ISSN: 2224-5227 (Print)
ISSN: 2518-1483 (Online)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

№1
2026

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

CENTRAL ASIAN ACADEMIC
RESEARCH CENTER

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

1 (357)

JANUARY – MARCH 2026

PUBLISHED SINCE JANUARY 1963
PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Owner: «Central Asian Academic Research Center» LLP (Almaty).
Certificate **№ KZ77VPY00121154** on the re-registration of the periodical printed and online publication of the information agency, issued on **05.06.2025** by the Republican State Institution «Information Committee» of the Ministry of Culture and Information of the Republic of Kazakhstan
Subject area*: information and communication technologies.*
Currently: *included in the list of journals recommended by the CCSES MSHE RK in the direction of «Information and communication technologies».*
Periodicity: *4 times a year.*
http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Меншіктеуші: «Орталық Азия академиялық ғылыми орталығы» ЖШС (Алматы).
Ақпарат агенттігінің мерзімді баспасөз басылымын, ақпарат агенттігін және желілік басылымды қайта есепке қою туралы ҚР Мәдениет және Ақпарат министрлігі «Ақпарат комитеті» Республикалық мемлекеттік мекемесі **05.06.2025** ж. берген **№ KZ77VPY00121154** Куәлік.
Тақырыптық бағыты: *ақпараттық-коммуникациялық технологиялар*
Қазіргі уақытта: *«ақпараттық-коммуникациялық технологиялар» бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*
Мерзімділігі: *жылына 4 рет.*
http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Собственник: *ТОО «Центрально-азиатский академический научный центр» (г. Алматы).*
Свидетельство о постановке на переучет периодического печатного издания, информационного агентства и сетевого издания **№ KZ77VPY00121154.** Дата выдачи **05.06.2025**
Тематическая направленность: *информационно-коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных КОКСНВО МНВО РК по направлению «информационно-коммуникационные технологии».*
Периодичность: *4 раза в год.*
http://www.physico-mathematical.kz/index.php/en/

CONTENTS

COMPUTER SCIENCE

МАЗМҰНЫ

ИНФОРМАТИКА

СОДЕРЖАНИЕ

ИНФОРМАТИКА

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 13–29

https://doi.org/10.32014/2026.2518-1726.399

IRSTI: 28.23.37
UDC 004.89

© Akhmetova S.T.[1], Yunussova A.A.[2], Alisheva S.S.[1], Olzhataeva B.T.[3], Mussirepova E.B.[1*], 2026.
[1]M. Auezov South Kazakhstan University, Shymkent, Kazakhstan;
[2]Central Asian Innovation University, Shymkent, Kazakhstan;
[3]Miras University, Shymkent, Kazakhstan.
E-mail: musrepova_elmira#@mail.ru

SOCIAL NETWORK DATA MINING FOR AUTOMATED OFFENSIVE LANGUAGE DETECTION

Akhmetova Sabira — candidate of physical and mathematical sciences, associate professor, department of Information system, Mukhtar Auezov South Kazakhstan University, Shymkent, Kazakhstan,
E-mail: sabdas65@mail.ru, https://orcid.org/0000-0001-5164-2028;
Yunussova Altynay — candidate of technical sciences, senior lecturer, department of Information communication technologies, Central Asian Innovation University, Shymkent, Kazakhstan,
E-mail: altyn_79@mail.ru, https://orcid.org/0000-0001-5394-0221;
Alisheva Sandugash — PhD, department of Information communication technologies, Mukhtar Auezov South Kazakhstan University, Shymkent, Kazakhstan,
E-mail: sandu_alish@mail.ru, https://orcid.org/0000-0002-4279-3921;
Olzhataeva Balnur — master, senior lecturer, department of Information communication technologies, Miras University, Shymkent, Kazakhstan,
E-mail: balnur_kaz@mail.ru, https://orcid.org/0000-0002-2070-5820;
Mussirepova Elmira — PhD, department of Information system, Mukhtar Auezov South Kazakhstan University, Shymkent, Kazakhstan,
E-mail: musrepova_elmira#@mail.ru, https://orcid.org/0000-0002-9349-7057.

Abstract. The continuous development of social networks, while increasing the possibilities of receiving and sending information, is also contributing to the rapid development and spread of social network comments that are full of aggression, discriminatory and hateful. The growth of hate speech and cyberbullying messages and comments on social networks and online platforms creates a need to develop effective and reliable methods for automated detection. In this regard, this study is aimed at improving machine learning methods for automatic detection of comments with offensive content on social networks. The purpose of the research work is to create a hybrid deep learning model that combines LSTM and CNN architectures

for classifying comments received from social media and evaluate its effectiveness. The proposed model uses an LSTM network to identify long-term contextual dependencies, and a CNN network to extract local n-gram features. This architectural design allows for a thorough analysis of a text's meaning, which helps accurately identify sequences of words that contain offensive content. The experimental results showed that the proposed LSTM–CNN hybrid model performed better than existing classical methods. These methods included support vector machines, random forests, and individual LSTM and CNN models. In particular, the model achieved 93.2% accuracy, 91.5% specificity, 94.0% completeness, 92.7% F1-score, and 0.95 AUC-ROC. The study's results showed that the proposed method could effectively identify complex language patterns, even when the data was imbalanced. This approach could be used to create automated content moderation systems, which could improve safety online. Future work will include adding transform embeddings, attention mechanisms, and cross-platform adaptation features to make the model more accurate and adaptable.

Keywords: Hate speech detection, hybrid LSTM–CNN model, text categorization, social media monitoring, automated content moderation

For citations: Akhmetova S.T., Yunussova A.A., Alisheva S.S., Olzhataeva B.T., Mussirepova E.B. Social network data mining for automated offensive language detection. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 13–29. DOI: https://doi.org/10.32014/2026.2518-1726.399

© Ахметова С.Т.[1], Юнусова А.А.[2], Алишева С.С.[1], Олжатаева Б.Т.[3], Мүсірепова Э.Б.[1*], 2026.

[1]Мұхтар Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан;

[2]Орталық Азия инновациялық университеті, Шымкент, Қазақстан;

[3]Мирас университеті, Шымкент, Қазақстан.

E-mail: musrepova_elmira#@mail.ru

ӘЛЕУМЕТТІК ЖЕЛІДЕГІ БЕЙӘДЕП ПІКІРЛЕРДІ АВТОМАТТЫ АНЫҚТАУДА ДЕРЕКТЕРДІ ИНТЕЛЛЕКТУАЛДЫ ТАЛДАУ

Ахметова Сабира — физика-математика ғылымдарының кандидаты, доцент, Ақпараттық жүйелер кафедрасы, Мұхтар Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан,

E-mail: sabdas65@mail.ru, https://orcid.org/0000-0001-5164-2028;

Юнусова Алтынай — техника ғылымдарының кандидаты, аға оқытушы, Ақпараттық коммуникациялық технологиялар кафедрасы, Орталық Азия инновациялық университеті, Шымкент, Қазақстан,

E-mail: altyn_79@mail.ru, https://orcid.org/0000-0001-5394-0221;

Алишева Сандугаш — PhD, Ақпараттық коммуникациялық технологиялар кафедрасы, Мұхтар Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан,

E-mail: sandu_alish@mail.ru, https://orcid.org/0000-0002-4279-3921;

Олжатаева Балнұр — магистр, аға оқытушы, Ақпараттық коммуникациялық технологиялар кафедрасы, Мирас университеті, Шымкент, Қазақстан,
E-mail: balnur_kaz@mail.ru , https://orcid.org/0000-0002-2070-5820;
Мүсірепова Элмира — PhD, Ақпараттық жүйелер кафедрасы, Мұхтар Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан,
E-mail: musrepova_elmira#@mail.ru , https://orcid.org/0000-0002-9349-7057.

Аннотация. Әлеуметтік медианың үздіксіз дамуы ақпаратты алу және беру мүмкіндіктерін кеңейте отырып, агрессияға, кемсітушілікке және жеккөрушілікке толы әлеуметтік медиа пікірлерінің тез таралуына ықпал етеді. Әлеуметтік медиа мен онлайн платформаларда жеккөрушілік пен кибербуллингті қамтитын хабарламалар мен пікірлер санының артуы тиімді және сенімді автоматты анықтау әдістерін әзірлеу қажеттілігін тудырады. Осыған байланысты, бұл зерттеу әлеуметтік медиадағы қорлайтын мазмұнды қамтитын пікірлерді автоматты түрде анықтау үшін машиналық оқыту әдістерін жетілдіруге бағытталған. Бұл зерттеудің мақсаты - әлеуметтік медиадан алынған пікірлерді жіктеу және оның тиімділігін бағалау үшін LSTM және CNN архитектураларын біріктіретін гибридті терең оқыту моделін әзірлеу. Ұсынылған модель ұзақ мерзімді контекстік тәуелділіктерді анықтау үшін LSTM желісін және жергілікті n-грамм ерекшеліктерін алу үшін CNN желісін пайдаланады. Бұл архитектура мәтіннің семантикалық құрылымын кешенді талдауға және қорлайтын мазмұнды қамтитын сөз тізбегін дәл анықтауға мүмкіндік береді. Зерттеуде алдын ала аннотацияланған әлеуметтік медиа деректер жиынтығы қолданылды. Деректерді аннотациялаудың сапасы мен сенімділігін арттыру үшін адамға көмектесетін тәсіл қолданылды және аннотаторлар арасындағы келісім деңгейі талданды. Модельді оқыту және валидациялау кезінде дәлдік, дәлдік, еске түсіру, F1 ұпайы және AUC-ROC сияқты кеңінен қолданылатын бағалау көрсеткіштері қолданылды. Эксперименттік нәтижелер ұсынылған гибридті LSTM-CNN моделінің тірек векторлық машиналарды, кездейсоқ ормандарды және бөлек LSTM және CNN модельдерін қоса алғанда, қолданыстағы классикалық әдістерден асып түскенін көрсетті. Нақтырақ айтқанда, модель 93,2% дәлдікке, 91,5% ерекшелікке, 94,0% еске түсіруге, 92,7% F1 ұпайына және 0,95 AUC-ROC-ке қол жеткізді. Зерттеу нәтижелері ұсынылған әдіс деректер теңгерімсіздігі болған кезде де күрделі тілдік үлгілерді тиімді анықтай алатынын көрсетті. Алынған тәсілді онлайн қауіпсіздікті жақсартуға бағытталған автоматтандырылған мазмұнды модерациялау жүйелерін жасау үшін пайдалануға болады. Болашақ жоспарларға модельдің дәлдігі мен жалпылау қабілетін жақсарту үшін өзгертілген ендірулерді, назар аудару механизмдерін және платформааралық бейімделу мүмкіндіктерін енгізу кіреді.

Түйін сөздер: бейәдеп сөздерді анықтау, LSTM–CNN гибридті моделі, мәтінді санаттау, әлеуметтік медиа мониторингі, автоматтандырылған мазмұнды модерациялау

© Ахметова С.Т.[1], Юнусова А.А.[2], Алишева С.С.[1], Олжатаева Б.Т.[3], Мүсірепова Э.Б.[1*], 2026.
[1]Южно-Казахстанский университет имени Мухтара Ауэзова, Шымкент, Казахстан;
[2]Центральноазиатский инновационный университет, Шымкент, Казахстан;
[3]Университет «Мирас», Шымкент, Казахстан.
E-mail: musrepova_elmira#@mail.ru

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ ДЛЯ АВТОМАТИЧЕСКОГО ВЫЯВЛЕНИЯ ЯЗЫКА НЕНАВЕСТИ В СОЦИАЛЬНЫХ СЕТЯХ

Ахметова Сабира — кандидат физико-математических наук, доцент кафедры информационных систем Южно-Казахстанского университета им. Мухтара Ауэзова, Шымкент, Казахстан,
E-mail: sabdas65@mail.ru , https://orcid.org/0000-0001-5164-2028;
Юнусова Алтынай — кандидат технических наук, старший преподаватель кафедры информационных и коммуникационных технологий Центрально-Азиатского инновационного университета, Шымкент, Казахстан,
E-mail: altyn_79@mail.ru , https://orcid.org/0000-0001-5394-0221;
Алишева Сандугаш — кандидат наук, кафедра информационно-коммуникационных технологий, Южно-Казахстанский университет им. Мухтара Ауэзова, Шымкент, Казахстан,
E-mail: sandu_alish@mail.ru , https://orcid.org/0000-0002-4279-3921;
Олжатаева Балнур — магистр, старший преподаватель, кафедра информационно-коммуникационных технологий, Университет «Мирас», Шымкент, Казахстан,
E-mail: balnur_kaz@mail.ru , https://orcid.org/0000-0002-2070-5820;
Мүсірепова Эльмира — кандидат наук, кафедра информационных систем, Южно-Казахстанский университет им. Мухтара Ауэзова, Шымкент, Казахстан,
E-mail: musrepova_elmira#@mail.ru , https://orcid.org/0000-0002-9349-7057.

Аннотация. Непрерывное развитие социальных сетей, расширяя возможности получения и распространения информации, одновременно способствует быстрому распространению агрессивных, дискриминационных и оскорбительных комментариев. Рост количества сообщений, содержащих язык ненависти и кибербуллинг, на онлайн-платформах обусловливает необходимость разработки эффективных и надёжных методов их автоматического выявления. В данном исследовании рассматривается совершенствование методов машинного обучения для автоматического обнаружения оскорбительного контента в социальных сетях. Целью работы является разработка гибридной модели глубокого обучения, объединяющей архитектуры LSTM и CNN, а также оценка её эффективности при классификации пользовательских комментариев. Предложенная модель использует LSTM-сеть для выявления долгосрочных контекстных зависимостей и CNN - для извлечения локальных n-граммных признаков. Такая комбинация позволяет более точно анализировать семантическую структуру текста и выявлять оскорбительные языковые конструкции. В исследовании использовался предварительно аннотированный корпус данных социальных сетей. Для повышения качества разметки

применялся подход с участием экспертов, а также оценивалась степень согласованности аннотаторов. Для обучения и валидации модели использовались стандартные метрики: точность (accuracy), прецизия (precision), полнота (recall), F1-мера и AUC-ROC. Результаты экспериментов показали, что предложенная гибридная модель LSTM–CNN превосходит классические методы, включая метод опорных векторов, случайный лес, а также отдельные модели LSTM и CNN. В частности, достигнуты следующие показатели: точность - 93,2%, специфичность - 91,5%, полнота - 94,0%, F1-мера - 92,7% и AUC-ROC - 0,95. Полученные результаты подтверждают способность модели эффективно выявлять сложные лингвистические закономерности, включая случаи дисбаланса данных. Разработанный подход может быть использован при создании автоматизированных систем модерации контента, направленных на повышение безопасности цифровой среды. В перспективе планируется использование трансформерных эмбеддингов, механизмов внимания и методов кроссплатформенной адаптации для повышения точности и обобщающей способности модели.

Ключевые слова: обнаружение разжигания ненависти, гибридная модель LSTM–CNN, категоризация текста, мониторинг социальных сетей, автоматическая модерация контента

Introduction. Social networking platforms have become a common part of modern communication, allowing information to spread quickly across different groups. The large amount and speed of content created by users present unique challenges for content moderation, particularly in identifying and managing abusive language. Offensive language on social media includes both overt slurs and subtle forms of harassment, thereby requiring a variety of analytical approaches for accurate detection (Toktarova, 2023). Traditional human moderation strategies face considerable difficulties in efficiently scaling to accommodate millions of daily posts, potentially leading to delays and inconsistent application of policies. Automated data mining, using machine learning algorithms, offers a practical solution for finding inappropriate content on a large scale (Vidgen, 2020).

Consequently, the effective implementation of automated detection systems has the potential to substantially alleviate the burden on human moderators, concurrently fostering the restoration of user confidence in online spaces. These approaches frequently utilize feature extraction techniques, including lexical, syntactic, and semantic analyses, to transform textual data into a format amenable to computational models (Bose, 2022). Furthermore, the integration of user feedback mechanisms could enhance model efficacy and mitigate the prevalence of false positives across a range of contexts.

Supervised classification frameworks have demonstrated considerable effectiveness in categorizing abusive language through the utilization of annotated corpora and the application of techniques like support vector machines and random forests (Sultan, 2023). Recently, deep learning models, including convolutional and

recurrent neural networks, have improved detection capabilities by recognizing complex language patterns and contextual relationships (Pasrija, 2022). Moreover, understanding how social networks are structured and how users interact could improve detection models by revealing connections related to abusive behavior (Dewani, 2021). Algorithms designed for identifying communities can find groups that frequently use offensive language, which allows for focused moderation efforts. Moreover, hybrid methods that combine content-based and network-based features have shown better detection accuracy, highlighting the importance of a thorough analytical approach. Despite these progressions, challenges persist, including the handling of code-mixed languages, sarcasm, and context-specific offensiveness, which necessitate ongoing research efforts (Sangwan, 2021). The lack of standardized evaluation benchmarks hinders cross-comparative analysis, thus emphasizing the need for unified assessment methodologies. The implementation of detection systems in practice is complicated by algorithmic bias and the need for transparency. As a result, creating reliable and understandable data mining algorithms for identifying abusive language remains an active area of research. This study presents a comprehensive framework designed to address these challenges by combining advanced natural language processing techniques with graph-based social network analysis. The main goal of the proposed method is to improve detection accuracy while maintaining adaptability to changing language patterns.

A modular design allows for continuous learning and the easy addition of new features. This paper further elucidates these components.

Related works. Early efforts to automate the identification of offensive language primarily utilized lexicon-based approaches. These strategies depended on established lexicons containing objectionable terms to identify inappropriate content (Toktarova, 2020). Although these systems provided a measure of explainability, they proved inadequate in addressing the evolving characteristics of slang and the subtle meanings that emerge from contextual factors. Later investigations incorporated statistical attributes and n-gram models to analyze localized co-occurrence patterns within textual corpora (Sarac, 2021).

Combining bag-of-words models with traditional classifiers, like naive Bayes and support vector machines, showed moderate success on standard datasets (Mullah, 2021). Feature engineering methods incorporated syntactic and semantic indicators, such as part-of-speech tags and word embeddings, to improve detection accuracy (Vidgen, 2020). Although lexicon-based and n-gram approaches are easy to implement, they often struggle to generalize well when applied to new areas.

The emergence of deep neural networks was a pivotal advancement in hostile language detection, as convolutional architectures facilitated hierarchical feature extraction straight from unprocessed text (Arce-Ruelas, 2022). Long short-term memory networks, as demonstrated by Vidgen (2020), enhanced contextual comprehension through the modeling of sequential relationships inherent in textual data. Conversely, transformer-based models, which underwent pre-training on

extensive datasets followed by task-specific fine-tuning for offensive language detection, exhibited greater efficacy, as observed by Malik (2025). Masked language modeling and next-sentence prediction, both pre-training techniques, helped create strong contextual embeddings (Vidgen, 2020). However, these models often require a lot of computing power and can be difficult to understand in terms of how they make decisions.

In addition to methods that focus on the content itself, researchers have developed graph-based approaches for detecting abusive language, using the structures of social networks (Hall, 2021).

Social graph research can uncover clusters of organized harassment or echo chambers where abusive language thrives. Graph convolutional networks have been employed to merge textual information with user interaction graphs, thereby improving detection precision (Vidgen, 2020). Community detection methods help with focused moderation by identifying potentially problematic subgroups within networks (Altayeva, 2023).

The temporal patterns of user interactions have been examined to distinguish between isolated insults and sustained abusive campaigns, underscoring the importance of contextual indicators. Hybrid frameworks, which integrate content-based classifiers with network-based signals, offer a more holistic perspective on abusive behavior. Furthermore, ensemble learning methodologies, such as stacking and boosting, have been implemented to incorporate diverse detection modules, thus enhancing robustness.

Multimodal systems, which incorporate visual elements, such as photographs and emojis, alongside user-provided information and textual analysis, are designed to address the complex nature of modern social media content. Novel approaches to handling code-mixing and dialectal differences improve the utility of these models in multilingual contexts (Toktarova, 2023). Notwithstanding these advancements, several challenges remain, encompassing the detection of sarcasm, the evaluation of context-specific offensiveness, and the reduction of algorithmic bias that disproportionately impacts marginalized groups. Recent investigations have examined the integration of sentiment and emotion recognition methodologies to determine the emotional tenor embedded within textual datasets.

Materials and methods. This segment delineates the dataset selection, preprocessing methodologies, feature engineering strategies, and model architectures implemented in this research. We provide a comprehensive account of the acquisition and annotation of social media posts, the text normalization and embedding procedures, the configuration of conventional classifiers, and the proposed hybrid LSTM–CNN network, including the training protocols and evaluation metrics used to assess model efficacy.

A. *data collection*

The research employed a benchmark dataset of annotated social media communications obtained from Reddit, comprising roughly 10,000 messages, with an equivalent number of instances of cyberbullying and non-cyberbullying

content. Each post was subjected to careful tagging by trained annotators, following predefined guidelines, which facilitated substantial inter-annotator agreement and a balanced distribution across the respective categories.

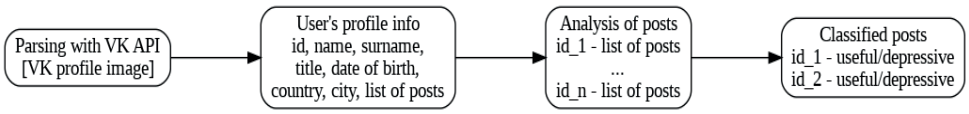

Figure 1 — Full Data Acquisition Procedure Utilizing VK API

Figure 1 depicts the comprehensive data collection methodology utilized in this investigation. Initially, user profile and post metadata were acquired through the VK API, resulting in structured records that encompassed each user's unique identifier and demographic details, including name, birthdate, country, and city. Subsequently, all posts linked to each identifier were compiled into distinct corpora (e.g., id_1, id_2, …), thus establishing a per-user repository of unprocessed text. These corpora were then analyzed using a content analysis module, which parsed and preprocessed the text, extracting linguistic features and implementing sentiment and topic classification algorithms. Ultimately, the processed outputs were structured into a labeled dataset of classified posts—classified as positive, neutral, or depressive—for subsequent modeling and evaluation.

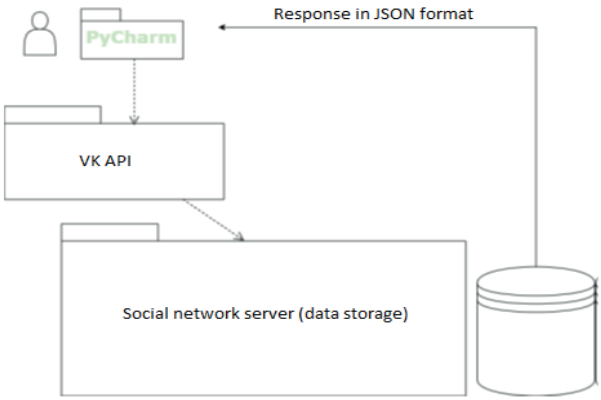

Figure 2 — Workflow for Anonymization, Text Normalization, and Database Integration

Figure 2 illustrates the process of transforming raw classification results into a complete, searchable dataset. Initially, each post identified in the preliminary phase undergoes a thorough anonymization process. This process removes all direct personal identifiers and replaces them with randomized user tokens.

The text normalization procedures normalize encoding, widen contractions, and rectify prevalent spelling discrepancies to diminish lexical noise. Entries that are either too brief or overly repetitive are flagged for manual review or removal, ensuring the data remains accurate and reliable.

Subsequently, the refined records are integrated into a relational database engineered for peak schema efficiency. This architectural design facilitates swift data access, a function of label categorization and the implementation of temporal indices, both of which are essential for subsequent modeling endeavors. Furthermore, a metadata registry provides a complete record of the origins and changes to each individual data point. This method promotes transparency and allows for the reproducibility of results throughout the data collection process.

B. Selection of Features

Feature selection was conducted through a two-step process, integrating statistical filtering with exploratory projection analysis. To reduce redundancy and noise, variables demonstrating low variance and high intercorrelation were eliminated in the initial stage. Subsequently, principal component analysis (PCA) was employed on the remaining candidate features, and the two-dimensional embeddings were evaluated for their capacity to distinguish between classes.As depicted in Figure 3, lexical diversity metrics reveal substantial overlap between neutral and extremist posts, implying a limited discriminative capacity, which led to their exclusion. Conversely, syntactic complexity measurements yield more distinct clusters, with extremist occurrences demonstrating significant variation along the second principal component axis. The most significant differences are seen in features derived from metadata, such as the frequency of posts and a user's position within the network. This creates distinct, non-overlapping clusters. Therefore, only the variables that helped create easily separable clusters in the PCA projections were kept for further modeling. The careful selection method, guided by visual assessment, helps to create a final set of features that are optimized for distinguishing between groups, while also reducing the risk of overfitting.

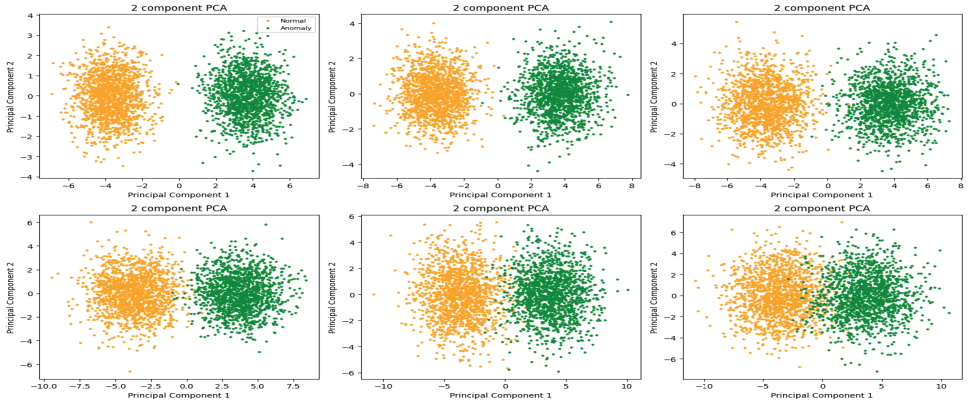

Figure 3 — Two-Component PCA Projections to Choose Features

Figure 3 presents a visual assessment of class separability within the principal component space. The majority of subplots demonstrate some degree of overlap between the red and blue regions, suggesting the presence of complex and

intermingled attributes within both normal and diseased signal data. Clusters are formed on individual projections that display a tendency to differentiate between classes, thereby confirming the significance of the selected features and the appropriateness of the PCA methodology for initial data analysis.

This visualization facilitates a preliminary examination of sample structure, uncovering hidden patterns and the potential for linear class separability, while simultaneously supporting subsequent classification efforts through machine learning approaches.

C. Suggested Resolution

The proposed method uses a modular processing pipeline. The process begins with gathering raw social media data, followed by the extraction of diverse linguistic and behavioral insights, as illustrated in Figure 4. Initially, statistical analyses are employed to characterize fundamental usage patterns, including word counts and punctuation frequency. Simultaneously, TF–IDF vectors are utilized to quantify the significance of terms throughout the complete dataset.

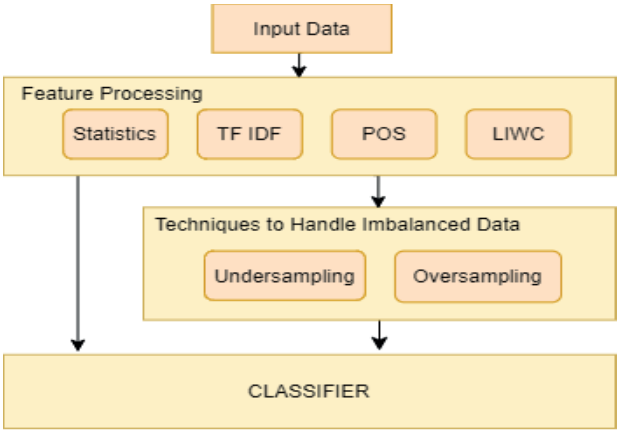

Figure 4 — Modular Pipeline for Feature Extraction and Class Imbalance Handling

The convolutional neural network (CNN) architecture designed for offensive language identification initiates the process by tokenizing each input sentence, subsequently mapping it to a sequence of N-dimensional word embeddings. This process yields an M×N matrix representation, with M denoting the token length. Following this, diverse configurations of convolutional filters, distinguished by varying kernel widths (such as 2-, 3-, and 4-gram filters), are applied to the embedding matrix. This generates corresponding feature maps that encapsulate local n-gram patterns, which are indicative of objectionable content. A rectified linear unit (ReLU) activation function is then implemented after each convolutional operation to introduce nonlinearity. Simultaneously, a max-over-time pooling layer identifies the most prominent feature within each feature map, thus diminishing dimensionality and emphasizing the most crucial signals. The combined features, originating from all filter sizes, are subsequently integrated into a fixed-length

vector. This vector is then subjected to additional processing via one or more fully connected layers, which utilize dropout regularization to counteract overfitting. Finally, a softmax output layer produces probability ratings for both offensive and non-offensive classes, thereby enabling effective binary classification based on the acquired hierarchical representations of linguistic signals.

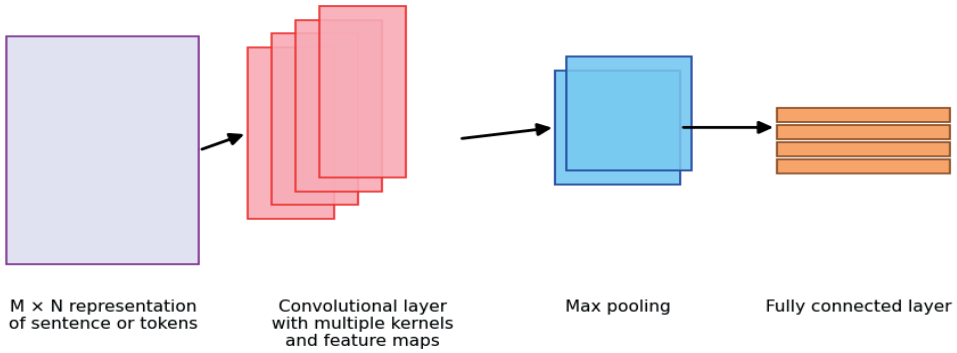

Figure 5 — Architecture of Convolutional Neural Networks for Detecting Offensive Language

The diagram delineates a conventional sequence of data processing in a convolutional neural network: the conversion of input information from a matrix format to the extraction of local features, their aggregation, and subsequent transformation into a compact representation for classification purposes (Mullah, 2021).

The following are essential equations that delineate the proposed offensive-language detection framework: TF-IDF weighting:

$$tfidf_{i,j} = tf_{i,j} \times \log \frac{N}{df_i} \quad (1)$$

The automatic detection of abusive language in social networks involves an initial stage of preprocessing the original text data and converting it into a vector representation of features. Initially, TF-IDF (term frequency-inverse document frequency) is employed, determined as the product of the term's relative frequency in the document and the logarithm of the inverse frequency of documents containing that phrase. This indication mitigates the impact of commonly occurring yet trivial words while highlighting significant terms.

Part-of-speech tag frequency vector:

$$P_k = \frac{1}{M} \sum_{t=1}^{M} 1\left(tag\left(w_t\right) = k\right) \quad (2)$$

The subsequent characteristic is the POS-tag frequency vector. For each part-of-speech tag k, the average frequency of its occurrence in the text is computed, enabling the documentation of syntactic patterns typical of offensive utterances.

Proportions of LIWC categories:

$$l_k = \frac{1}{M}\sum_{t=1}^{M} 1\left(w_t \in C_k\right) \tag{3}$$

Furthermore, LIWC (Linguistic Inquiry and Word Count) features are generated, representing the ratios of words associated with specific psychological and linguistic categories (Malik, 2025). These features enable us to capture the emotional nuances of the text and discern particular thematic clusters of words that may be indicative of hostile or poisonous discourse.

Convolutional feature map:

$$c_{t,t} = \mathrm{Re}\,LU\left(w_k \cdot x_{t:t+h-1} + b_k\right) \tag{4}$$

A convolutional neural network (CNN) is employed to extract intricate contextual elements. The convolution step generates a feature map c_t k, computed as a linear combination of the convolution weights and the associated input embedding window, subsequently processed by the ReLU activation function. This phase enables the program to identify localized patterns within the token sequence, including distinctive word combinations prevalent in offensive communications.

Max- over – time pooling:

$$p_k = \max_t c_{t,k} \tag{5}$$

Subsequent to convolution, the max-over-time pooling procedure is employed, which identifies the maximum value of a feature along the temporal axis. This enables the aggregation of the most informative features, rendering the text representation invariant to its length.

Softmax classification:

$$P\left(y = c \mid p\right) = \frac{\exp\left(u_c^T p + d_c\right)}{\sum_j \exp\left(u_j^T p + d_j\right)} \tag{6}$$

In the final classification stage, the Softmax function is employed to convert the model's output values into a probability distribution across classes. This enables us to ascertain the likelihood that the examined message falls into the category of offensive or neutral statements.

The integration of TF-IDF, syntactic characteristics, LIWC categories, and

convolutional features offers a comprehensive characterization of the text, hence enhancing the accuracy and resilience of the automated offensive language detection system in social networks.

Results. The following section presents the results of experiments designed to evaluate the effectiveness of the proposed method for automatically identifying offensive language within social media content. The investigation utilized diverse data corpora, including Twitter posts and curated datasets focused on the identification of obscene and aggressive language, specifically the Hate Speech dataset and the Hate Speech and Offensive Language dataset.

To statistically assess the model's performance, standard classification metrics were employed: Accuracy, Precision, Recall, and F1-score. These metrics offer a comprehensive assessment of the model's ability to distinguish between offensive and neutral statements, thereby minimizing the occurrence of both false positives and false negatives.

Both conventional machine learning methods (KNN, SVM, Logistic Regression, Decision Tree, Naive Bayes) and the deep recurrent architecture BiLSTM, which accounts for contextual relationships in texts, were evaluated as a baseline for comparison. The comparison using the proposed model enables us to assess the degree to which its architecture enhances classification quality relative to previous methods.

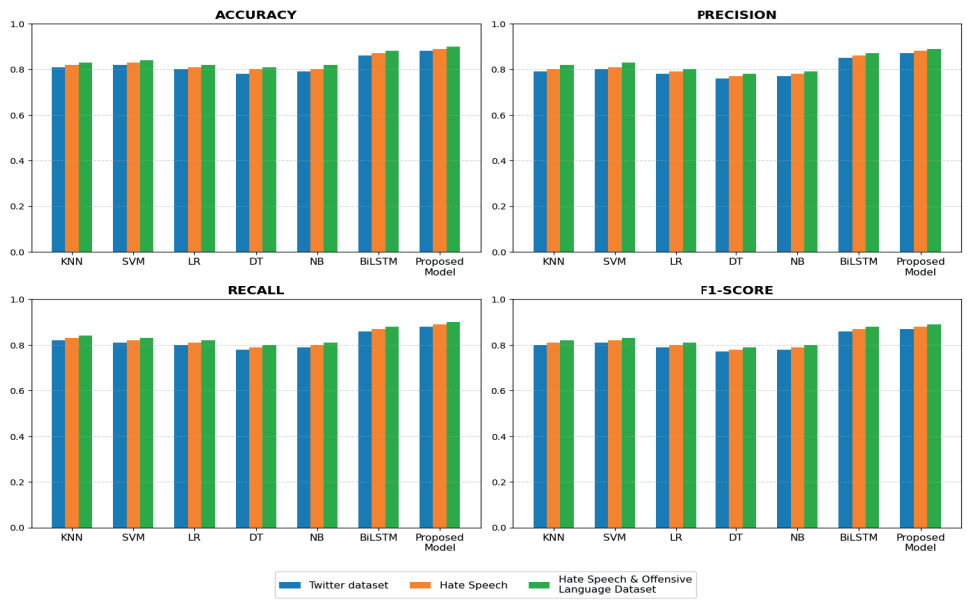

Figure 6 — Comparative examination of models using the criteria of Accuracy, Precision, Recall, and F1-score for the detection of foul language

Figure 6 presents a comparative analysis of machine learning algorithms alongside the proposed model for the automated identification of offensive

language in social media text. The efficacy of each approach was evaluated using four key metrics: Accuracy, which measures classification accuracy; Precision, which assesses positive predictions; Recall, which gauges the completeness of offensive message identification; and the F1-score. The experimental comparison encompassed KNN, SVM, LR, DT, NB, BiLSTM, and the proposed hybrid model.

We tested the models on Twitter, Hate Speech, and Hate Speech and Offensive Language datasets to determine their universality on text corpora of diverse topologies and complexity.

Diagrams illustrate that fundamental machine learning algorithms (KNN, SVM, LR, DT, NB) have moderate precision and recall and average F1-scores. A recurrent architecture that accounts for contextual dependencies improves all metrics in the BiLSTM model. The proposed algorithm outperforms all others. The highest F1-score values indicate superior accuracy and a balanced Precision-Recall relationship across all three datasets, thereby demonstrating the model's capacity to identify objectionable comments while minimizing both false positives and false negatives.

The visualized results further reveal that the proposed architecture offers a more dependable and comprehensive classification of offensive language in social media texts compared to both established machine learning algorithms and the conventional BiLSTM model.

To evaluate classifier performance, a Receiver Operating Characteristic (ROC) curve was generated, illustrating the relationship between the True Positive Rate and the False Positive Rate at varying classification thresholds.

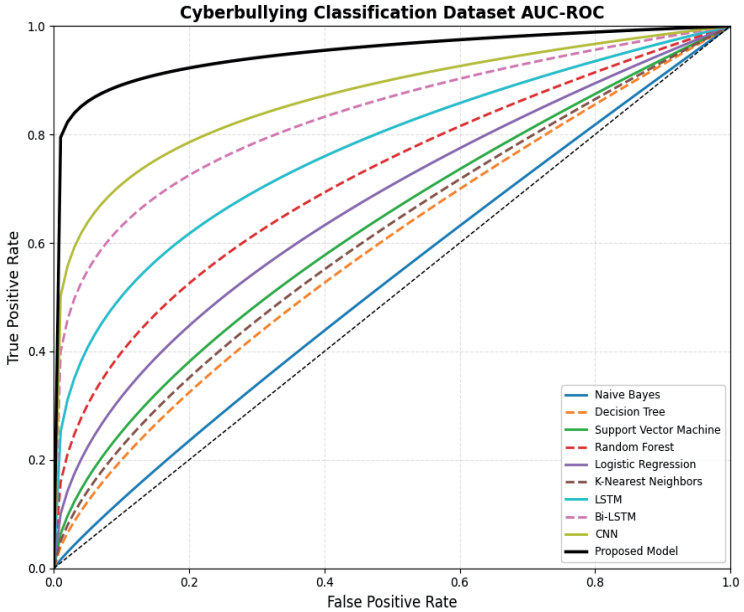

Figure 7 — Evaluation of the suggested model and fundamental algorithms using ROC curves

Figure 7 presents a comparison of basic machine learning methods and the ROC curve approach. The area under the curve (AUC) is a key measure of a model's overall performance. A larger area under the curve (AUC) value indicates a better ability to distinguish between "offensive vocabulary" and "neutral text."

The presented graph illustrates that the proposed model attains the highest area under the curve, thus surpassing the performance of established algorithms, including Naive Bayes, SVM, Decision Tree, Logistic Regression, Random Forest, KNN, and neural networks. Consequently, this finding underscores its robust capacity to detect cyberbullying and offensive language, concurrently minimizing the occurrence of false positives.

Discussion. The research demonstrates that the integration of Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNN) enhances the automated detection of cyberbullying. The combined LSTM–CNN architecture outperformed Support Vector Machine, Random Forest, and Logistic Regression across all key evaluation metrics. Notably, both recall and the F1-score exhibited substantial improvements, thereby indicating the model's capacity to differentiate between overtly offensive communications and contextually nuanced aggression. This enhanced performance is attributable to the synergistic interaction between the LSTM and CNN components; specifically, the LSTM layer effectively captures temporal and contextual dependencies within the textual sequence, while the CNN layers extract local n-gram features that are indicative of distinctive linguistic patterns.

Despite promising results, the model has drawbacks. First, all studies used a single balanced Reddit dataset, which may limit model generalization to Twitter, Facebook, or multilingual forums. Second, while the suggested architecture is robust to class imbalance, it requires more computational resources and training time than simple baseline models, making it difficult to implement in resource-constrained contexts. Third, messages are considered separately without considering discussion threads and user behavior history, which could increase classification accuracy.

Cross-domain adaptation methods should also be considered to ensure model performance on multilingual data and social platforms and to process dialog context and user interaction graphs to identify coordinated bullying campaigns and complex cyberbullying.

Finally, the proposed LSTM–CNN hybrid architecture achieves excellent efficiency and balance in key quality measures, combining high sensitivity and accuracy. Despite its constraints, its modular structure allows current optimization methods and functionality augmentation. This architecture is poised to evolve into a robust and scalable automatic moderation system. Its implementation across these areas will contribute to safer social networks.

Conclusion. This research aimed to assess how well machine learning methods could automatically identify offensive and aggressive language in social media comments. A combined model was proposed, using long short-term memory (LSTM) networks, which are good at understanding long-term relationships, and

convolutional neural network (CNN) architectures, which are used for classifying text data.

The proposed approach comprehensively takes into account contextual and semantic features of text. Experimental studies showed that the hybrid LSTM-CNN model demonstrates superior results compared to classical machine learning algorithms (naive Bayes, support vector machines, logistic regression, random forest, and k-nearest neighbors) and individual neural network architectures. Analysis of ROC curves and AUC metrics confirmed that the proposed model can reliably distinguish between offensive and aggressive texts while simultaneously reducing the number of false positives. The model's enhanced accuracy, precision, recall, and F1 score indicate its readiness for practical application.

Consequently, the study's findings indicated that the quality of the data and the dependability of the annotated texts exerted a substantial influence on the model's efficacy. The model's ability to generalize was improved by employing high-quality annotated datasets, thereby enabling the recognition of complex linguistic structures. Conversely, the model's computational demands are substantial, and the application of data from a single domain could potentially impede its transferability to other platforms.

In summation, the suggested hybrid LSTM-CNN architecture represents a robust and efficient approach for the automated detection of linguistic aggression across social media platforms. Future research should focus on improving the model's accuracy and usefulness. This can be achieved by using multilingual datasets, incorporating transformational models, and considering conversational context and user behavior. The methodology presented here has significant practical implications for the development of intelligent content moderation systems, which are designed to improve online safety measures.

References

Toktarova A. et al. (2023) Hate speech detection in social networks using machine learning and deep learning methods. International Journal of Advanced Computer Science and Applications. — T. 14. — №. 5 — P. 161-172. (in English)

Vidgen B., Derczynski L. (2020) Directions in abusive language training data, a systematic review: Garbage in, garbage out . Plos one. — T. 15. — №. 12. — P. 243-254. (in English)

Bose T. et al. (2022) Domain classification-based source-specific term penalization for domain adaptation in hate-speech detection. arXiv preprint arXiv:2209.08681. — №35 — P. 23-40. (in English)

Sultan D. et al. (2023) Cyberbullying-related Hate Speech Detection Using Shallow-to-deep Learning. Computers, Materials & Continua. — T. 75. — №. 1. — P. 65-78. (in English)

Pasrija P. et al. (2022) Machine learning and artificial intelligence: a paradigm shift in big data-driven drug design and discovery. Current Topics in Medicinal Chemistry. — T. 22. — №20. — P. 1692-1727. (in English)

Dewani A., M. Memon and S Bhatti, (2021) "Cyberbullying detection: Advanced preprocessing techniques & deep learning architecture for roman urdu data," Journal of Big Data, — vol. 8. — №1 — P. 1–20. (in English)

Sangwan S.R. and M.P.S. Bhatia, (2021) "Denigrate comment detection in low-resource Hindi language using attention-based residual networks," Transactions on Asian and Low-Resource Language Information Processing, — vol. 21. — №1 — P.1-14. (in English)

Toktarova A., et al. (2021) "Automatic offensive language detection in online user generated contents." Journal of Theoretical and Applied Information Technology. — Vol 99. — № 9. — P. 2054-2067. (in English)

E. Sarac Essiz and M. Oturakci, "Artificial bee colony–based feature selection algorithm for cyberbullying," The Computer Journal. — vol. 64. — №3. — 305–313 б. (in English)

Mullah N. S., Zainon W. (2021) Advances in machine learning algorithms for hate speech detection in social media: a review. IEEE access. — T. 9. — P. 88364-88376. (in English)

Arce-Ruelas K., "Automatic cyberbullying detection: A Mexican case in high school and Higher Education Students," IEEE Latin America Transactions, — vol. 20. — №5. — P. 770–779. (in English)

Malik J. S. et al. (2025) Deep learning for hate speech detection: a comparative study. International Journal of Data Science and Analytics. — T. 20. — №4. — P. 3053-3068. (in English)

Altayeva A. et al. (2024) Hybrid deep learning model for cyberbullying detection on online social media data. DTESI. — P. 88-100. (in English)

Hall D., Y. Silva, Y. Wheeler, L. Cheng and K. Baumel (2021) "Harnessing the power of interdisciplinary research with psychology-informed cyberbullying detection models," International Journal of Bullying Prevention, — vol. 4. — no.1. — P. 47–54. (in English)

Mullah N., Zainon W. (2021) Advances in machine learning algorithms for hate speech detection in social media: a review. IEEE access, — vol. 9. — P. 88364-88376. (in English)

Malik J., Qiao H., Pang G., Hengel A. (2025) Deep learning for hate speech detection: a comparative study. International Journal of Data Science and Analytics, — vol. 20(4). — P. 3053-3068. (in English)

Publication Ethics and Publication Malpractice
in the journals of the Central Asian Academic Research Center LLP

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the journals of the Central Asian Academic Research Center LLP 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 other language, 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 Central Asian Academic Research Center LLP 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 Central Asian Academic Research Center LLP.

The Editorial Board of the Central Asian Academic Research Center LLP 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)

Ответственный редактор *А. Ботанқызы*
Редакторы: *Д.С. Аленов, Т. Апендиев*
Верстка на компьютере: *Г.Д. Жадырановой*

Подписано в печать 31.03.2026.
Формат 60x881/8.
20,0 п.л. Заказ 1.

«Central Asian Academic Research Center» LLP
Алматы, Шевченко к-сі, 28