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

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

№4
2025

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

CENTRAL ASIAN ACADEMIC
RESEARCH CENTER

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

4 (356)

OCTOBER – DECEMBER 2025

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

МАЗМҰНЫ

СОДЕРЖАНИЕ

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 4. Namber 356 (2025). 234–258

https://doi.org/10.32014/2025.2518-1726.394

УДК 004.855.5

© L. Rzayeva[1], A. Ryzhova[1], M. Zhaparkhanova[1], A. Myrzatay[1], Zh. Kozhakhmet[1], 2025.
[1]Astana IT University, Astana, Kazakhstan;
[2]RSE on PCV "Digital Government Support Center", Astana, Kazakhstan.
E-mail: l.rzayeva@astanait.edu.kz

A NEW LSTM-BASED WEB APPLICATION FOR AUTOMATED PASSWORD STRENGTH EVALUATION

Rzayeva Leila — PhD, Associate Professor, Director of the "CyberTech" Research and innovation Center, Astana IT University, Astana, Kazakhstan,
E-mail: l.rzayeva@astanait.edu.kz, https://orcid.org/0000-0002-3382-4685;
Ryzhova Alissa — Junior Researcher at the "CyberTech" Research Center, Astana IT University, Astana, Kazakhstan,
E-mail: alissaryzh@gmail.com, https://orcid.org/0009-0006-3781-1773;
Zhaparkhanova Merei — Junior Researcher at the "CyberTech" Research and innovation, Kazakhstan,
E-mail: zhaparkhanovamerei.16@gmail.com**,** zhaparkhanovamerei.16@gmail.com, https://orcid.org/0009-0002-7177-7805;
Myrzatay Ali — PhD, Junior Researcher at the "CyberTech" Research and innovation Center, Astana IT University LLP, Astana, Kazakhstan,
E-mail: mirzataitegiali@gmail.com; ORCID ID: https://orcid.org/0000-0002-5339-2437;
Kozhakhmet Zhaksylyk — Junior Researcher at the "CyberTech" Research and innovation Center, Astana IT University LLP, Astana, Kazakhstan,
E-mail: zh.kozhakhmet@astanait.edu.kz; ORCID ID: https://orcid.org/0009-0002-5449-3317.

Abstract. According to the latest Verizon DBIR report, poor handling of credentials, including password reuse and the human factor in their creation remains one of the key attack vectors. A study conducted by the authors showed that most users change passwords only when they are lost, and 35% consider mandatory regular changes inconvenient. This underscores the need for technical solutions capable of clearly demonstrating system vulnerabilities and raising security awareness. Within the study, the human factor in creating usernames and passwords is considered a vulnerability. Identifying the patterns and rules people use significantly reduces the number of combinations an attacker must try to gain access. The proposed method is based on a character-level LSTM model that detects recurring structures and generates generalized masks reflecting characteristic patterns. Training was performed on public datasets containing 31,000 compromised passwords. The model achieved over 90% accuracy on the test set without signs of overfitting.

The approach combines the analysis of individual user habits in generating logins and passwords with automatic keyword extraction from open sources using a keyword extraction algorithm. The method is integrated into a web application that allows local fine-tuning of the model, running it via ONNX, and performing all computations on the device, ensuring data confidentiality and compliance with information security requirements.

Keywords: cybersecurity; password security; machine learning; neural networks; social engineering; digital forensic; cryptography; behavioural analysis

© Л. Рзаева[1], А. Рыжова[1], М. Жапарханова[1], А. Мырзатай[1,2], Ж. Кожахмет[1], 2025.
[1]Astana IT University, Астана, Қазақстан;
[21]Цифрлық үкіметті қолдау орталығы" ШЖҚ РМК, Астана, Қазақстан.
E-mail: l.rzayeva@astanait.edu.kz

ҚҰПИЯСӨЗДІҢ БЕРІКТІГІН АВТОМАТТЫ БАҒАЛАУҒА АРНАЛҒАН LSTM НЕГІЗІНДЕГІ ЖАҢА ВЕБ-ҚОСЫМША

Рзаева Лейла — PhD, қауымдастырылған профессор, «CyberTech» ҒЗО директоры, Astana IT University, Астана, Қазақстан,
E-mail: l.rzayeva@astanait.edu.kz, https://orcid.org/0000-0002- 3382-4685;
Рыжова Алиса — кіші ғылыми қызметкер, «CyberTech» ҒЗО, Astana IT University, Астана, Қазақстан,
E-mail: alissaryzh@gmail.com, https://orcid.org/0009-0006-3781-1773;
Жапарханова Мерей — кіші ғылыми қызметкер, «CyberTech» ҒЗО, Astana IT University, Астана, Қазақстан,
E-mail:zhaparkhanovamerei.16@gmail.comб https://orcid.org/0009-0002-7177-7805;
Мырзатай Али — PhD, аға ғылыми қызметкер, «CyberTech» ҒЗО, Astana IT University, Астана, Қазақстан,
E-mail:mirzataitegiali@gmail.comб https://orcid.org/0000-0002-5339-2437;
Кожахмет Жаксылык — ғылыми қызметкер, «CyberTech» ҒЗО, Astana IT University, Астана, Қазақстан,
E-mail:zh.kozhakhmet@astanait.edu.kzб https://orcid.org/0009-0002-5449-3317.

Аннотация. Бұл мақалада көптілді мәтіндерді өңдеу мен семантикалық қауіптілік талдауына арналған цифрлық криминалистика жүйесінің заманауи архитектурасы ұсынылады. Негізгі мәселе – құрылымдалмаған хабарламалардан маңызды ақпаратты автоматты түрде бөліп алып, оны нақты қауіп категорияларына жатқызу қажеттілігі. Бұл мәселе құқық қорғау органдарының заманауи тергеу процестерінде жиі кездеседі және оны шешу жедел әрекет етуді, масштабталатын жүйелер мен көптілділікке төзімділікті талап етеді. Ұсынылған жүйе Python/FastAPI негізінде құрылған серверлік архитектураны, Qdrant векторлық дерекқорын, алдын ала үйретілген jina-embeddings-v3 үлгісін және Llama-Guard-3-1B қауіп классификациялау моделін пайдаланады. Жүйе мәтінді векторизациялау, семантикалық іздеу, тілдерді автоматты анықтау және көптілді аударма, сонымен қатар

қауіпті контентті санатқа бөлу сынды бірнеше кезеңнен тұрады. Негізгі гипотеза – семантикалық векторлық модельдер мен категориялық гибридті классификацияны біріктіру арқылы қауіпсіздік деңгейін жоғары дәлдікпен анықтауға болады. Эксперименттік нәтижелер жүйенің нақты уақытта өңдеу қабілетін көрсетті (7448 хабарлама ~3 секундта индекстеледі) және Llama-Guard моделінің 75.19% дәлдікпен хабарламаларды қауіптілік деңгейі мен типі бойынша жіктей алатынын растады. Орташа жауап уақыты аудармасыз сұраныстарда ~0.45 сек және аударманы қажет ететін жағдайларда ~5.23 сек құрайды. Жүйенің визуалды модулі пайдаланушыға хабарламаларды топтар бойынша, қолданбалар, алушылар, геолокациялар және уақыттық үлгілер бойынша талдауға мүмкіндік береді. Бұл функционал криминалистикалық сараптаманың тиімділігін арттырып, тергеудің негізгі кезеңдерін автоматтандыруға жағдай жасайды. Интерфейс жүйені ыңғайлы басқаруға, деректерді жеңілдетуге және шешім қабылдауды жылдамдатуға көмектеседі. Ұсынылған шешім құқық қорғау, ұлттық қауіпсіздік, ақпараттық қауіпсіздік және киберқылмысқа қарсы күрес салаларында практикалық қолдануға бейімделген және болашақта аудио/видео деректермен кеңейтілуі мүмкін.

Түйін сөздер: табиғи тілдерді өңдеу (NLP), мәтінді векторизациялау, семантикалық іздеу, көптілді талдау, автоматты машиналық аударма, криминалистика, қатерлерді жіктеу

© Л. Рзаева[1], А. Рыжова[1], М. Жапарханова[1], А. Мырзатай[1,2], Ж. Кожахмет[1], 2025.

[1] НИЦ “CyberTech” Astana IT University, Астана, Казахстан;

[2] РГП на ПХВ Центр поддержки цифрового правительства, Астана, Казахстан.

E-mail: l.rzayeva@astanait.edu.kz

НОВОЕ ВЕБ-ПРИЛОЖЕНИЕ НА ОСНОВЕ LSTM ДЛЯ АВТОМАТИЗИРОВАННОЙ ОЦЕНКИ НАДЕЖНОСТИ ПАРОЛЕЙ

Рзаева Лейла — PhD, Ассоциированный профессор, Директор НИЦ «CyberTech», Astana IT University, Астана, Казахстан,
E-mail: l.rzayeva@astanait.edu.kz, https://orcid.org/0000-0002-3382-4685;
Рыжова Алиса — младший научный сотрудник НИЦ «CyberTech», Astana IT University, Астана, Казахстан,
E-mail: alissaryzh@gmail.com, https://orcid.org/0009-0006-3781-1773;
Жапарханова Мерей — младший научный сотрудник НИЦ «CyberTech», Astana IT University, Астана, Казахстан,
E-mail: zhaparkhanovamerei.16@gmail.com, https://orcid.org/0009-0002-7177-7805;
Мырзатай Али — PhD, старший научный сотрудник НИЦ «CyberTech», Astana IT University, Астана, Казахстан,
E-mail: mirzataitegiali@gmail.com, https://orcid.org/0000-0002-5339-2437;
Кожахмет Жаксылык — научный сотрудник НИЦ «CyberTech», Astana IT University, Астана, Казахстан,
E-mail: zh.kozhakhmet@astanait.edu.kz, https://orcid.org/0009-0002-5449-3317.

Аннотация. Согласно последнему отчёту Verizon DBIR, неграмотное обращение с учётными данными, включая повторное использование паролей и влияние человеческого фактора при их создании, остаётся одним из ключевых векторов атак. Исследование, проведённое авторами, показало, что большинство пользователей меняют пароли лишь при их утрате, а 35 % считают обязательную регулярную смену неудобной. Это показывает, насколько необходимы решения, которые способны наглядно демонстрировать уязвимости и повышать осведомленность в вопросах безопасности паролей. В рамках данного исследования человеческий фактор при создании паролей рассматривается как уязвимость. Определение типичных шаблонов и правил значительно сокращает количество комбинаций и время, которые злоумышленнику необходимо для получения доступа. Предложенный метод основан на LSTM-модели, выявляющей повторяющиеся структуры и маски, отражающие характерные паттерны на уровне отдельных символов. Для обучения использовались датасеты в открытом доступе, включающие 31 000 скомпрометированных паролей. Модель показала более 90% точности на тестовой выборке без переобучения. Подход объединяет анализ индивидуальных привычек и шаблонов пользователя при генерации паролей с автоматическим извлечением ключевых слов из открытых источников с помощью алгоритма keyword extraction. В настоящее время метод интегрирован в веб-приложение, которое позволяет предварительно обучать модель, запускать ее через ONNX и выполнять все вычисления на конечном устройстве пользователя. Это обеспечивает конфиденциальность данных и соответствие требованиям информационной безопасности.

Ключевые слова: кибербезопасность, безопасность паролей, машинное обучение, нейронные сети, социальная инженерия, цифровая криминалистика, криптография, поведенческий анализ

Введение. В эпоху цифровых технологий большинство людей неосознанно раскрывают личную информацию, (имена, даты рождения, клички домашних животных, личные хобби), в социальных сетях, которые часто служат основой для их учетных данных. Эти элементы, особенно в сочетании с распространенными шаблонами или «масками», могут быть легко использованы злоумышленниками с помощью социальной инженерии.

Хотя существует множество инструментов для оценки надежности паролей на основе традиционных политик, таких как длина и разнообразие символов (Bergeron & Dearden, 2024; Atzori et al., 2024; He et al., 2021) они не учитывают индивидуальную логику, лежащую в основе формирования паролей пользователями, поскольку для этого потребовалось бы отслеживать историю использования логинов и паролей, личные данные и привычки формирования пароля.

Несмотря на широкое применение дополнительных методов аутентификации, например таких как биометрия и двухфакторная верификация, односложные буквенные пароли по-прежнему остаются

основным методом доступа. Представленный метод, который оценивает потенциальную уязвимость пароля путем анализа открытых данных и экспорта из менеджеров паролей, показывает, как индивидуальные привычки и присутствие в Интернете влияют на его предсказуемость. В результате они представляют собой серьезную угрозу безопасности: в 2024 году было скомпрометировано более 5,5 миллиарда учетных записей по всему миру, что почти в восемь раз больше, чем в 2023 году (Surfshark, 2020). Данные опроса выявили, что пользователи часто используют привычные методы: 42% смешивают значимые слова и цифры, 34% следуют базовым требованиям платформы, а 32% повторно используют части предыдущих паролей (Urrico, 2024). Существующие решения, как правило, не учитывают индивидуальные правила формирования паролей конкретного пользователя. Некоторые допускают такие пароли, как «имяфамилия1!», или вовсе не анализируют часто повторяющиеся шаблоны в базах данных утечек.

Это исследование, являясь частью большого исследовательского проекта (Rzayeva et al., 2025; Idrissova et al., 2025) фокусируется на широко распространенной тенденции пользователей игнорировать предложения системы по созданию паролей, при этом почти 90% предпочитают создавать собственные комбинации на основе знакомых и предсказуемых шаблонов (Buckman, 2025). Эти шаблоны, как правило, следуют одному из нескольких общих правил: слова из словаря, числовые последовательности, (такие как номера телефонов или даты рождения), или символьные маски — причем символы редко используются отдельно. Описанный подход фокусируется на личных привычках пользователя, учитывает характерные паттерны и предлагает удобный формат в виде расширения для Google Chrome с возможностью анализа в режиме реального времени.

Научный вклад данного исследования содержит: анализ существующих методов взлома паролей и выявление слабых мест в паролях, созданных пользователями; анализ правил формирования паролей на основе данных из утечек; проектирование и обучение LSTM (Long Short-Term Memory) для выявления структурных шаблонов; разработку инструмента, предоставляющего индивидуальную обратную связь с применением методов социальной инженерии.

В отличие от существующих работ, данное исследование сосредоточено на взломах, вызванных человеческим фактором, а не на технических уязвимостях или методах перебора. Если злоумышленник получает доступ к предыдущим комбинациям логина и пароля, (например, через локальные устройства или базы утечек), и сочетает это с открытыми данными из социальных сетей, он может вывести правила генерации паролей пользователя, что значительно ускоряет получение несанкционированного доступа к его учетным записям. Новизна исследования заключается в предложенном методе проверки паролей, который учитывает человеческий фактор и зависимость от социальной инженерии, игнорируемые традиционными средствами проверки. Идентифицируя общие паттерны, мы создаем обобщённые маски паролей,

выделяющие структурные слабые места, потенциально позволяющие в будущем взломать другие учетные записи.

Материалы и методы:

1.2 **Новый метод обнаружения повторяющихся правил при создании логина/пароля**

Анализ поведения пользователей показывает, что при отказе от использования автоматически сгенерированных системой паролей они демонстрируют склонность к повторению одних и тех же логических схем и правил их формирования, внося минимальные изменения. Такая повторяемость может быть связана с намерением упростить процесс запоминания, снижая общую стойкость паролей к подбору.

Новый метод, логическое объяснение которого представлено на рисунке 1, вводится путем анализа прозрачности и предсказуемости логинов и паролей.

Важно отметить, что метод можно использовать только при наличии хотя бы части истории логинов и паролей пользователя. Именно на основе этой истории создается новая научная и прикладная работа. В отличие от существующих методов, этот подход предполагает, что каждый пользователь имеет свои собственные устойчивые правила формирования паролей.

Метод объединяет два типа анализа. Первый - поиск повторяющихся структур и паттернов в последовательности логинов и паролей пользователя. Второй - выделение его личного «словаря» на основе активности в социальных сетях. Используя результаты этих двух шагов, система в режиме реального времени оценивает, насколько новый пароль выбивается из привычных шаблонов, и формирует персональные рекомендации по усилению его надёжности.

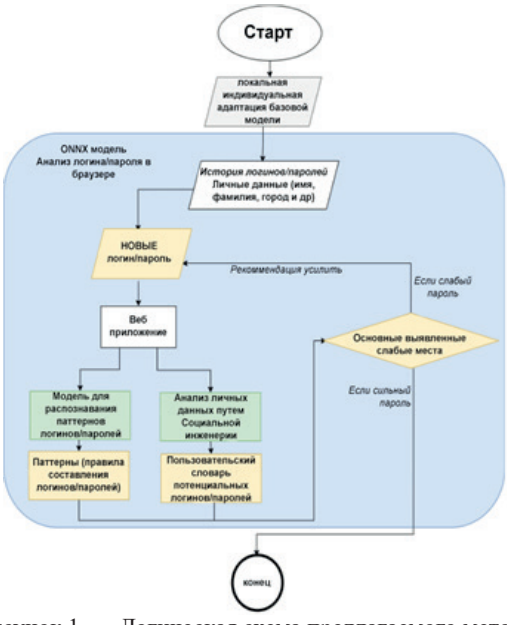

Рисунок 1 — Логическая схема предлагаемого метода

Сбор эмпирических данных, их техническая обработка, обучение моделей и проверка пользовательских сценариев — все это были шаги, которые использовались для достижения целей исследования. Исходные данные были созданы с использованием открытых наборов скомпрометированных паролей (RockYou, LinkedIn, GitHub и другие). После удаления дублей, служебных символов и неправильных записей был создан очищенный массив более 31 000 анонимизированных строк, который был готов к дальнейшему анализу.

Обучающая выборка дополнительно пополнялась открытыми датасетами Kaggle (Password dataset 2024) и GitHub (Kkrypt0nn., n.d.-b). На подмножестве этих данных обучалась модель Decision Tree Classifier; перед запуском обучения каждому паролю вручную присваивался тип. В ходе обучения алгоритм последовательно отбирает наиболее информативные признаки и формирует древовидную структуру (рис. 2), в которой внутренние узлы соответствуют логическим условиям, разделяющим выборку на более однородные группы по целевому признаку.

После завершения обучения модель была протестирована на отложенной выборке, что позволило оценить ее способность обобщать закономерности на новые, ранее не встречавшиеся данные. Для анализа результатов были получены прогнозы модели, которые затем сравнивались с фактическими метками классов, что позволило количественно оценить точность классификации. Однако впоследствии это решение было изменено в пользу более гибких и выразительных моделей, способных учитывать последовательный характер паролей и выявлять сложные скрытые зависимости в структуре символов.

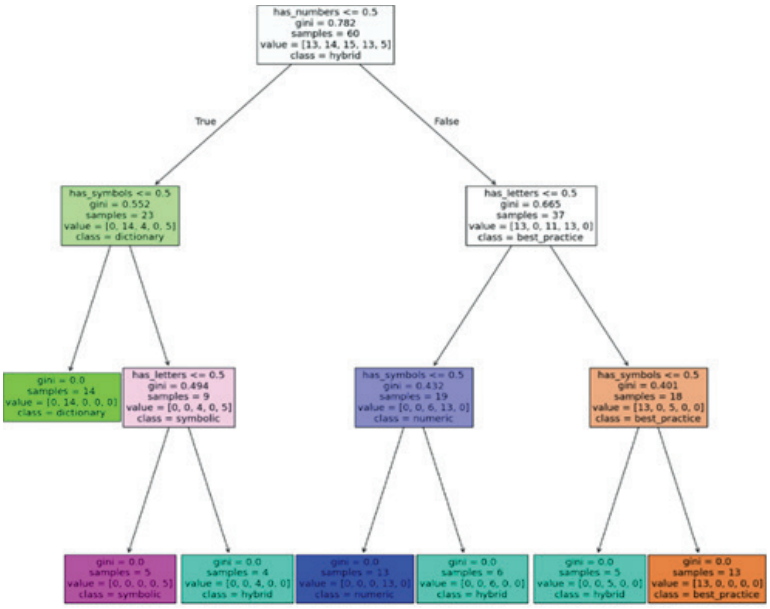

Рисунок 2 – Результаты Decision Tree

Классификатор на основе дерева решений обучается сопоставлению:

$$f(x) \rightarrow \{0, 1, \ldots, K-1\} \quad (1)$$

где x = [x_1, x_2, x_3] бинарные features показывающее присутствие букв (x_1), цифр (x_2) и символов (x_3) в пароле, и K это количество возможных типов пароля.

В каждом узле модель выбирает признак и пороговое значение для разделения данных таким образом, чтобы минимизировать нечистоту Джини:

$$G = 1 - \Sigma\, p_i^2 \quad (2)$$

где p_i – это доля класса i в узле. Дерево рекурсивно разбивает пространство признаков и присваивает наиболее часто встречающуюся метку класса каждому листу. В результате получается интерпретируемая модель классификации типов паролей на основе лексических свойств.

Дерево решений, традиционный класс алгоритмов, используемый в задачах классификации, использовался в качестве одной из основных моделей исследования. На начальном этапе они использовались для разделения паролей по простым признакам, таким как наличие букв, цифр, особых символов и их комбинаций. Это помогло задокументировать стандартные схемы паролей и повторяющихся комбинаций, которые впоследствии использовались как основа для более тщательного анализа и создания обобщенных масок. При этом модель дерева решений служила как стартовым фильтром, так и источником эмпирических гипотез. Это привело к переходу к более гибким последовательностным моделям, таким как LSTM.

В дополнение к этому были проведены исследования метода контекстно-свободной грамматики на основе вероятности (PCFG), который был направлен на определение устойчивых структур в паролях пользователей. В этом случае пароль представляется последовательностью логических фрагментов, состоящих из подпоследовательностей букв (L), цифр (D) и специальных символов (S). Вероятностная грамматика строится по множеству этих разбиений, учитывая тип и длину каждого сегмента.

Для каждого пароля была сгенерирована структура, например, L5-D4 для пароля Merei1611. С точки зрения формальности, PCFG модель разбивает пароль p по частям s_1, s_2, …, s_n, где каждая часть $s_i = (\tau_i, v_i)$, с $\tau_i \in$ {L, D, S} представляет тип символов (Буквы, цифры, специальные символы), и v_i относится к части пароля.

Структурная часть пароля представлена так:

$$\sigma(p) = \tau_1|v_1| - \tau_2|v_2| - \ldots - \tau_n|v_n| \quad (3)$$

Например пароль "Merei1611" соответствует шаблону L5–D4.

Вероятность шаблона представлена как:

$$P(\sigma) = C(\sigma) / N$$

где C(σ) это частота шаблона σ в датасете, и N это общее количество возможных шаблонов.

Условная вероятность сегментов:

$$P(v_i \mid \tau_i) = C_i(v_i) / \Sigma C_i \quad (4)$$

где $C_i(v_i)$ подсчет под частей слова v_i среди сегментов типа τ_i, и ΣC_i это общее количество всех сегментов (Фигура 6)

Предполагая независимость между сегментами, общая вероятность пароля составляет:

$$P(p) = P(\sigma(p)) \times \Pi P(v_i \mid \tau_i) \quad (5)$$

Затем надёжность пароля рассчитывается как нормализованный отрицательный логарифм по основанию 10 этой вероятности:

$$Score(p) = -(1 / |p|) \times log_{10} P(p) \quad (6)$$

Чем выше балл, тем более редкий и надежный пароль. Модель оценивает уникальность и предсказуемость паролей путем анализа их структурных и лексических паттернов. После обучения на специально подобранном наборе данных, в котором были зафиксированы частоты встречаемости этих структур, а также частоты встречаемости определенных подстрок в рамках категорий. На этапе оценки паролей модель использует логарифмическую метрику, которая отражает вероятность структуры и ее элементов с поправкой на длину.

Модель обучалась итеративно, наблюдая за прогрессом по среднему баллу и точности распознавания структур, которые были известны. Чтобы продемонстрировать это, были созданы два графика. На рис. 4 показано изменение среднего логарифмического балла на обучающих и валидационных выборках по мере увеличения обучающей группы, а на рис. 3 показана точность сопоставления структур на обеих выборках. Это позволяет наблюдать за тем, насколько хорошо модель адаптируется к новым данным, и обнаруживать паттерны, которые часто встречаются в паролях.

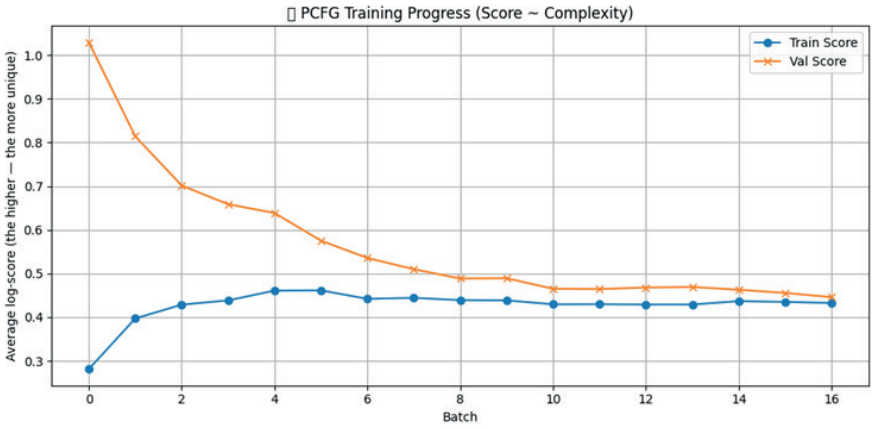

Рисунок 3 – PCFG Training Progress

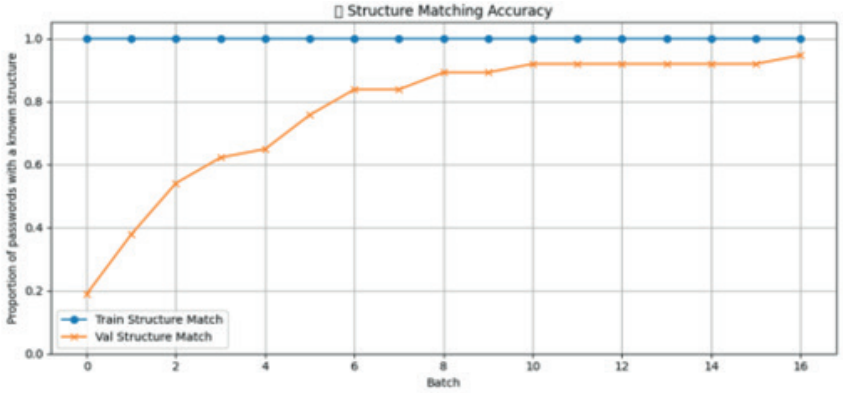

Рисунок 4 –Structure Matching Accuracy

Модели PCFG и деревья решений опираются на заранее определенные правила и характеристики для анализа паролей, например PCFG представляет пароли в виде фрагментов и определяет вероятность появления этих паттернов. Модель была выбрана потому что является негласным стандартом в анализе структуры паролей. Этот метод был применен для изучения распределения структурных паттернов и оценки вероятности использования определенных форматов паролей.

Однако оба подхода, недостаточно гибкие, так как не учитывают влияние человеческого фактора на составление учетных данных. Деревья решений используют жесткие логические разделения, а PCFG плохо справляется с нестандартными шаблонами, индивидуальными привычками или эмоционально заряженными комбинациями.

Эти ограничения привели к переходу на другую архитектуру. Основой анализа паролей, описанного в этой статье, является LSTM. Без ручной настройки правил эта модель способна обучаться на сырых данных и

определять сложные персонализированные шаблоны. В области анализа паролей способность LSTM обрабатывать последовательные данные имеет решающее значение, поскольку сложность пароля может быть изменена даже незначительными изменениями символов. Кроме того, было показано, что модели LSTM эффективны для задач, таких как генерация текста и предсказание следующего элемента.

3.3. Техническое выполнение предложенного метода

Архитектура разработанной модели состоит из трех логических уровней. Первый этап включает входных данных посимвольно изменяя в плотные векторные представления фиксированной размерности, что позволяет проводить эффективное обучение по сравнению с классическим однозначным кодированием, поскольку дает модели возможность улавливать синтаксические и структурные особенности символов, а не только их расположение в алфавите.

На втором этапе применяется двухслойная LSTM модель, которая анализирует контекстуальные отношения во всей строке пароля. Таким образом, модель может распознавать не только комбинации цифр и букв, но и положения особых символов, специфичных для способностей пользователей. Механизм регуляризации помогает предотвратить переобучение на ограниченных наборах данных, а использование двух слоев облегчает обработку долгосрочных и краткосрочных зависимостей между символами.

На последнем этапе используется полностью подключенный слой для преобразования выходных состояний LSTM в вероятностное распределение по всем символам, доступным в словаре. Модель может предсказать не только следующий символ, но и оценить вероятность того, что определенные паттерны будут присутствовать в пароле.

Основные цели моделирования структуры паролей и уровень абстракции отличают модель LSTM, предложенную в этом исследовании, от подхода на основе нейронных сетей (Melicher et al., 2016). Использование рекуррентной нейронной сети на уровне символов позволяет оценить вероятность перехода между отдельными символами. Это позволяет провести детальное моделирование угадываемости паролей на основе зависимостей от места. Их архитектура адаптирована к крупномасштабному моделированию атак, уделив особое внимание вероятности последовательностей символов в паролях.

В отличие от этого, в исследовании уделяется больше внимания поведенческому анализу при создании паролей с целью выявления семантически значимых правил и закономерностей, отражающих когнитивные привычки человека и социальную активность. Для построения обобщенных структурных масок предполагается использование предлагаемой архитектуры LSTM модели для обнаружения повторяющихся подстрок и типичных комбинаций, таких как имена и годы рождения.

Таким образом, наш метод предоставляет дополнительную информацию

о человеческом факторе, который способствует созданию предсказуемых и уязвимых паролей. Это отличается от методов, которые используются для моделирования вероятности очевидности паролей (Melicher et al., 2016).

Алгоритм оптимизации Адам, который обеспечивает стабильную и быструю конвергенцию в глубоких нейронных сетях, использовался для проведения обучения. Кросс-энтропия, которая является стандартной для задач многоклассовой классификации, использовалась в качестве функции потерь. Для обеспечения воспроизводимости и стабильности результатов во всех модулях случайности было установлено фиксированное значение семени. Это позволяет проводить эксперименты снова и снова в тех же условиях и получать те же результаты. Даже при обработке больших объемов строковых данных процесс выполнялся на локальной рабочей станции с графическим ускорителем, что обеспечивало достаточное время обучения. Набор данных для обучения состоял из тысяч паролей, которые были получены из общедоступных утечек, которые были предварительно очищены, чтобы убрать дублики и неинформативные символы. Входные последовательности обычно состояли от 8 до 24 символов, что свидетельствует о том, насколько часто пользователи используют пароли.

В математическом плане LSTM представляет собой набор формул, которые обрабатывают последовательность, такую как текст, шаг за шагом. С помощью специальных «ворот», управляемых значениями от 0 до 1, он определяет, что запомнить, а что забыть. Все это основано на стандартной математике, включая умножение, матричные операции и нелинейные функции, такие как сигмоид и гиперболический тангенс.

В соответствии с документацией PyTorch torch.nn.LSTM, следующие формулы описывают операции LSTM для первого слоя модели PasswordLSTM с embedding_dim = 128, hidden_dim = 256 и num_layers = 2.

Основная часть архитектуры LSTM, входные ворота(it), отвечают за регулирование включения информации в состояние ячейки. Он контролирует степень влияния входных данных на обновление ячейки памяти на текущем временном шаге.

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \quad (7)$$

где

$W_{ii} \in R^{256\times128}, W_{hi} \in R^{256\times256}, b_{ii}, b_{hi} \in R^{256}, x_t \in R^{128}, h_{t-1} \in R^{256}$

$x_t \in R^{128}$ – входной вектор по временному периоду t

$h_{t-1} \in R^{256}$ – скрытое состояние прошлого временного отрезка

$W_{ii} \in R^{256} \times 128$ – матрица весов для входа

$W_{hi} \in R^{256} \times 256$ – матрица весов для скрытого состояния

$b_{ii}, b_{hi} \in R^{256}$ – векторы погрешности

$\sigma(\cdot)$ – сигмоидная функция активации

Ворота забывания (f_t) — часть LSTM, решающая, какую информацию удалить из памяти. На вход подаются текущие данные и предыдущее состояние; на выходе — числа от 0 до 1, где 0 означает «полностью забыть», а 1 — «оставить без изменений».

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \quad (8)$$

где $x_t \in R^{128}$ – входной вектор на шаге t,

$h_{t-1} \in R^{256}$ – скрытое состояние с предыдущего шага

$W_{if} \in R^{256\times128}$ – матрица весов для входа x_t

$W_{hf} \in R^{256\times256}$ – матрица весов для скрытого состояния h_{t-1}

$b_{if}, b_{hf} \in R^{256}$ – векторы смещений,

$\sigma(\cdot)$ – сигмоидальная функция активации.

Клеточные ворота (кандидатное состояние ячейки) ($\tilde{C}_t$) — часть LSTM, формирующая новое содержимое памяти, т.е. предлагающая, что можно добавить. Значения вычисляются на основе входа и предыдущего состояния, но напрямую не добавляются — это делает входной клапан.

$$\tilde{C}_t = tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \quad (9)$$

где $x_t \in R^{128}$ – входной вектор на шаге t

$h_{t-1} \in R^{256}$ – скрытое состояние с предыдущего шага,

$W_{ig} \in R^{256\times128}$ – матрица весов по входу для кандидата,

$W_{hg} \in R^{256\times256}$ – матрица весов по скрытому состоянию для кандидата,

$b_{ig}, b_{hg} \in R^{256}$ – векторы смещений,

$tanh(\cdot)$ – гиперболический тангенс.

Выходные ворота (o_t) — часть LSTM, определяющая, какую часть текущего состояния памяти показать «наружу», т.е. что станет новым скрытым состоянием ht. Используется текущее состояние памяти Ct и его фильтрация сигмоидой (значения от 0 до 1).

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \quad (10)$$

где $x_t \in R^{128}$ – входной вектор на шаге t

$h_{t-1} \in R^{256}$ – скрытое состояние с предыдущего шага,

$W_{io} \in R^{256\times128}$ – матрица весов по входу для выходных ворот,
$W_{io} \in R^{256\times256}$ – матрица весов по скрытому состоянию для выходных ворот,
$b_{io}, b_{ho} \in R^{256}$ – векторы смещений,
$\sigma(\cdot)$ – сигмоидальная функция активации.

Состояние ячейки(C_t) процесс обновления внутреннего состояния памяти C_t в LSTM. Оно сочетает старую память C_{t-1}, умноженную на ворота забывания, и новое предложение $\tilde{C}_t$ умноженное на входные ворота.

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t\, b_{ho}) \tag{11}$$

где $C_{t-1} \in R^{256}$ – предыдущее состояние ячейки,
$f_t \in R^{256}$ – выход ворот забывания,
$i_t \in R^{256}$ – выход входных ворот,
$C_t \in R^{256}$ – кандидатное состояние ячейки,
$\odot$ – обозначает покомпонентное умножение.

Скрытое состояние (h_t) — выход LSTM на текущем шаге, то, что «выдаёт» сеть. Вычисляется на основе обновлённой памяти C_t (после $tanh$) и выходных ворот o_t, решающих, какую её часть показать.

$$h_t = o_t \odot tanh(C_t) \tag{12}$$

где $C_t \in R^{256}$ – текущее состояние ячейки,
$o_t \in R^{256}$ – выход выходных ворот,
$h_t \in R^{256}$ – результирующее скрытое состояние,
$tanh(\cdot)$ – гиперболический тангенс,
$\odot$ –обозначает покомпонентное умножение.

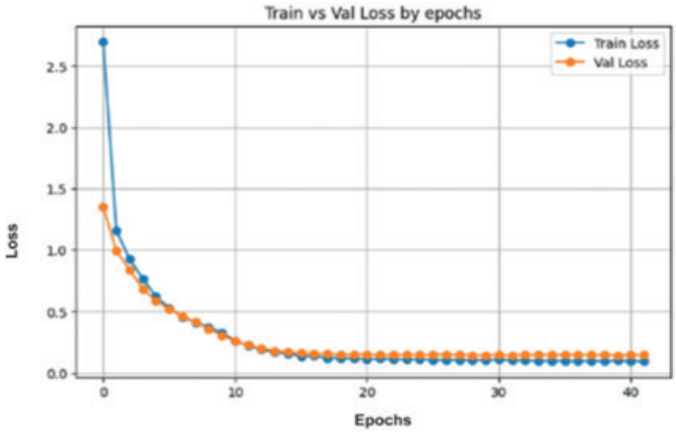

Рисунок 5 –Train vs Validation Loss

Динамика функции потерь при обучении модели Password LSTM на протяжении 43 эпох продемонстрирована на рисунке 8. На начальном этапе кривые обучающей и валидационной потерь демонстрируют резкое снижение, особенно в первые 10 эпох. Модель захватывает базовые структурные зависимости в последовательностях паролей и эффективно уменьшает ошибку предсказания на ранней стадии обучения, что и иллюстрирует быстрый спад.

После ~10-й эпохи скорость снижения потерь замедляется, и примерно к 15-й эпохе обе кривые стабилизируются на значении 0,15. Это плато свидетельствует о достижении сходимости, после чего улучшаются лишь маргинальные величины. Близкое совпадение двух кривых на протяжении всего процесса обучения отражает высокую обобщающую способность и отсутствие переобучения, что особенно важно при работе с посимвольными последовательностями.

В целом динамика обучающей функции потерь подтверждает, что модель была корректно настроена и обучалась на репрезентативном датасете. Согласованное поведение обеих кривых по эпохам демонстрирует устойчивость архитектуры LSTM и этапов препроцессинга, включающих токенизацию символов, нормализацию последовательностей и обогащение признаков. Полученные результаты подтверждают надежность модели в задачах распознавания структур паролей и её готовность к практическому внедрению.

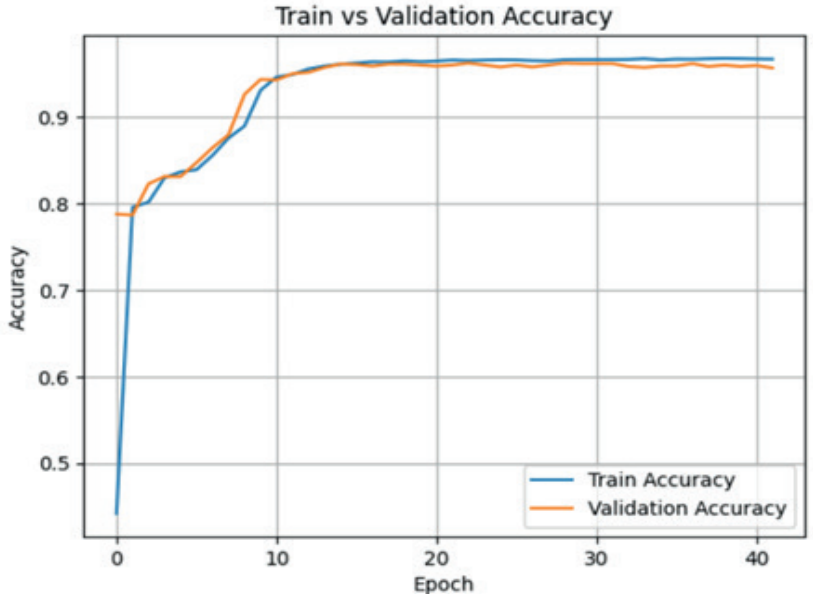

Рисунок 6 –Точность на обучающем и валидационном наборах

Результаты классификации модели Password LSTM как на обучающем, так и на валидационном наборах показаны в рисунке 6. На первых десяти поколениях точность резко увеличивается. Модель переходит от случайного угадывания к стабильно высоким значениям. Несмотря на ограниченный алфавит и сравнительно короткие входы, модель демонстрирует этот рост, быстро извлекая важные паттерны и семантические зависимости из последовательностей паролей.

С пятнадцатого века различия на кривых остаются минимальными, а точность обеих выборок улучшилась до 95%. Это поведение особенно важно для посимвольных моделей последовательностей, потому что оно показывает сохранение качества предсказаний на данных, не использовавшихся при обучении, и отсутствие явных признаков переобучения.

В более поздние времена правильная настройка конвейера предварительной обработки и выбранные механизмы регулирования связаны с высокими значениями точности. Исключение между слоями LSTM и нормализация длины последовательностей обеспечивает устойчивость предсказаний при работе с разнородными паролями. Эти возможности позволяют модели использоваться для задач оценки устойчивости паролей в Интернете и анализа пользовательских шаблонов в средствах безопасности браузеров.

Результаты обучения, показанные на Рисунках 5 и 6, подтверждают надёжность и устойчивость предложенной LSTM-модели. Модель эффективно обнаруживает скрытые закономерности в данных о паролях без переобучения, что подтверждают быстрые сходимости функции потерь и стабильная точность свыше 95% на обучающем и валидационном наборах. Благодаря близкому совпадению кривых модель подходит для задач анализа и прогнозирования структуры паролей. В целом процесс обучения оказался успешным, что привело к созданию надежной и высокопроизводительной архитектуры.

Был проведен двухэтапный анализ, чтобы оценить влияние общедоступных персональных данных на структуру пользовательских паролей. На первом этапе использовался Serper API для сбора текстовых сниппетов с именами, городом, датой рождения и адресом. Ключевые токены и пароли из корпуса были объединены после нормализации текста. Следующим шагом было использование метрик RapidFuzz для оценки точных и частичных совпадений имени/логина, а также семантической близости к OSINT-токенам. Основные характеристики пароля, такие как наличие цифр и символов, были объединены с результатами и представлены в виде корреляционной матрицы.

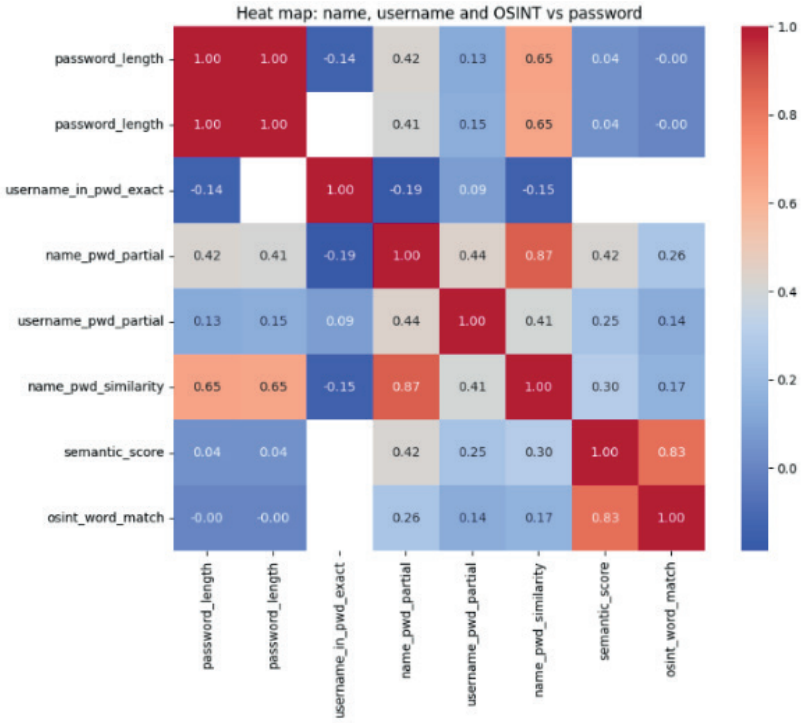

Рисунок 7 –Тепловая карта корреляций между паролями и персональными данными

Проведенный анализ (Рис. 7) показал, что длина пароля положительно коррелирует с его сходством с именем (r = 0,65), тогда как частичные совпадения демонстрируют сильную корреляцию как с полной похожестью (r = 0,87), так и с семантической оценкой (r = 0,42). При этом, если в пароле сохраняются персональные токены, добавление цифр или подстановочных символов лишь незначительно усложняет его предсказуемость. Отсюда следует, что имена и связанные с ними элементы остаются основным источником уязвимости, а усложнение структуры без удаления этих элементов не гарантирует надежную защиту.

После этапа обучения модель была экспортирована в формат ONNX, что позволило интегрировать её в два прикладных инструмента: веб-приложение на Vue 3 и расширение браузера Google Chrome, благодаря этому всё вычисление выполняется локально: без отправки чувствительных данных на удалённые серверы. Исключая серверное хранение или анализ, такая архитектура обеспечивает максимальную конфиденциальность данных: учётные сведения пользователей не собираются, не сохраняются и не обрабатываются на стороне сервера. Интерфейсы работают в режиме реального времени, поэтому пользователь получает мгновенную обратную связь в виде отображения разложения строки на структурные элементы, присвоение уровня безопасности (WEAK, MED, STRONG) и оценку

вероятности на основе выходного распределения модели. Расширение использует Indexed DB для хранения повторяющихся паттернов. Это позволяет отслеживать стандартные шаблоны ввода и адаптировать советы к поведению пользователя. Таким образом, цель реализации модели состоит не только в достижении технической точности, но и в том, насколько она полезна для повседневной кибергигиены.

Опираясь на вероятностное сканирование модели LSTM, система встроена в комплексную платформу анализа паролей. Приложение веб-браузера и расширение Chrome предоставляют обратную связь в режиме реального времени. Веб-приложение загружает экспортированную в ONNX модель и словарь «символ→индекс» при открытии страницы через onnxruntime-web (WASM), что позволяет клиенту выполнять все вычисления инферентности. Асинхронная процедура вызывает функцию предсказания силы для определения двух важных метрик: средней отрицательной лог-вероятности (avg(–log P)) и нормализованного процента «силы» (0–100). Это происходит после того, как пользователь вводит пароль в привязанное поле и нажимает «Проверить». Реактивные переменные Vue.js мгновенно обновляются при помощи этих значений: красная — WEAK (менее 40 %), желтая — MED (40–70 %), зелёная — STRONG (более 70 %), а также выводится среднее значение для продвинутых пользователей (рис. 8).

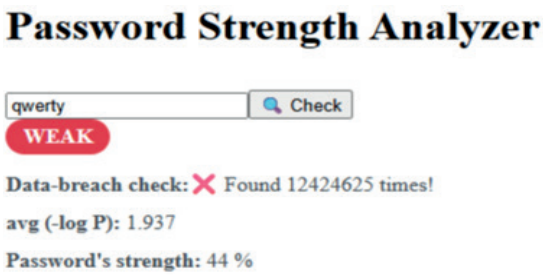

Рисунок 8. Веб-приложение: анализ пароля

После вычисления оценки система сопоставляет пароль с предварительно рассчитанным списком уязвимых подстрок, извлеченных в ходе обучения, помечая совпадения предупреждающими иконками для выделения структурных слабых мест. Но в API Have I Been Pwned отправляется только первые пять символов SHA-1-хеша пароля, таким образом предотвращая возможные атаки повторного использования данных, а возвращенный список суффиксов проходит локальную фильтрацию для отображения количества утечек без вскрытия чувствительных данных. Ниже основной панели динамический компонент Password Chart (Рис. 12) отображает семь наиболее часто повторяющихся паролей, а текстовые списки «Маски, упрощающие угадывание паролей» и «Рекомендуется избегать» (Рис. 13) направляют пользователей вдали от как общих, так и персонализированных паттернов.

Рисунок 9. Веб-приложение: график наиболее повторяемых паролей

Рисунок 10. Веб-приложение: паттерны, которые рекомендуется избегать

Расширение для Chrome воспроизводит эту логику на каждой странице с полями <input type="password">. Определённое в manifest.json, оно включает всплывающий интерфейс (popup UI), фоновый воркер и контент-скрипт, реализованные на Vue 3, TypeScript и Vite. Во время онбординга пользователь при желании вводит своё имя и город; это инициирует поиск в DuckDuckGo, из которого извлекаются значимые ключевые слова для персонализированных проверок паттернов и сохраняются локально в chrome. storage и IndexedDB. Контент-скрипт добавляет рядом с каждым полем пароля живой бейдж и слушает нажатия клавиш; при каждом вводе выполняются те же ONNX-инференс, сопоставление паттернов и проверка утечек через унифицированную функцию rate (), обновляющие бейдж и встроенную диагностику за миллисекунды. Поддержка «горячей» перезагрузки через сообщения chrome.runtime держит паттерны актуальными без перезагрузки страниц. Все вычисления — исполнение модели, парсинг ключевых слов, проверки паттернов — происходят на устройстве пользователя; из браузера

уходят только неидентифицирующие хеш-префиксы или поисковые запросы, что полностью сохраняет приватность при обеспечении быстрой и информативной обратной связи (Рис. 11).

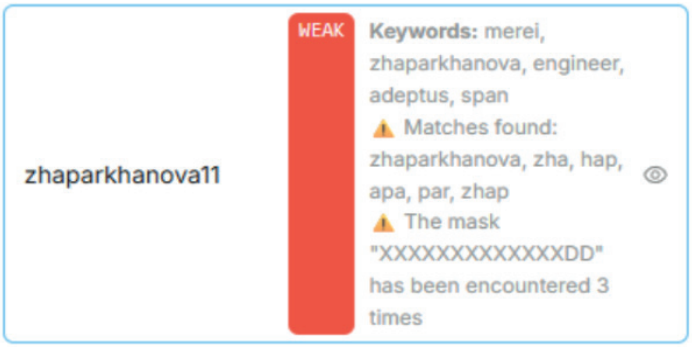

Рисунок 11. Расширение Chrome

Для оценки устойчивости пользовательских паролей к перебору была применена LSTM-модель, обученная на персонализированном датасете. В ходе обучения сеть автоматически выявляла часто повторяющиеся последовательности символов, характерные для конкретного пользователя, и формировала набор «масок» и подстрок, которые статистически повышают предсказуемость паролей данного пользователя на основе его правила формирования паролей и набора наиболее частых слов. Такой персонализированный подход дополнен локальным инференсом (ONNX) и допускает дообучение на конечном устройстве клиента, что исключает вывоз сырых данных и соответствует требованиям корпоративной ИБ.

Сходимость подтверждена динамикой функции потерь и точности: на Рисунке 5 показано, что обучающие и валидационные потери быстро падают в первые 10 эпох и стабилизируются ниже ~0,15, признаков переобучения не наблюдается; Рисунок 6 фиксирует достижение валидационной точности 96,7 % при близких значениях на обучении. Для калибровки «знакомости» структуры использовалась средняя отрицательная лог-вероятность (Avg Log-Score): на валидации получено 0,224 (меньше — значит более предсказуемо). В связке «accuracy + Avg Log-Score» модель не только классифицирует «сильные/слабые» пароли, но и количественно оценивает вероятность структурных и семантических шаблонов.

Согласно исследованию, персональные токены основательно влияют на степень предсказуемости паролей. Например, анализ сформированного моделью набора характерных данных (подстрок) демонстрируют доминацию числовых и именных последовательностей. Это в свою очередь отображает предпочтения пользователей при формировании их паролей, как показано на рисунке 12.

```json
```

{
 "merei": 20,
 "erei04": 1,
 "merei04": 1,
 "zha": 49,
 "apa": 96
}```

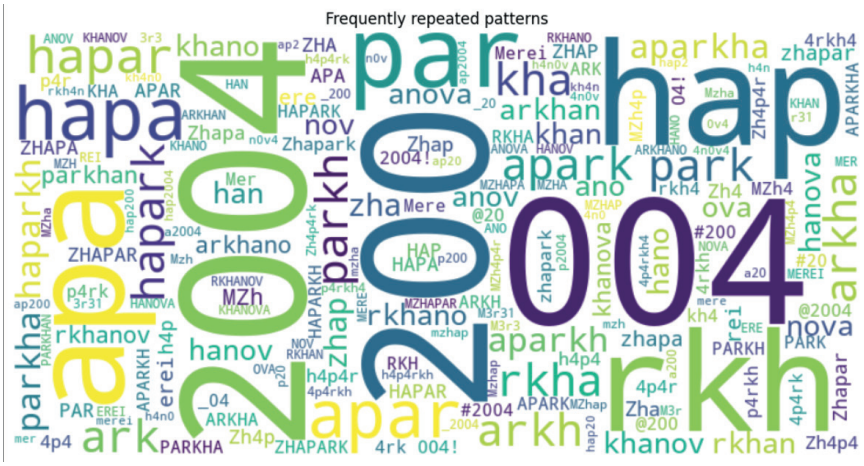

Рисунок 12 – Облако слов по часто повторяющимся паттернам

На рисунке 12 визуализировано какие последовательности преобладают среди числовых последовательностей и токенов, например «2004» и «zhap», «khan», «Merei», что объясняет повышенную предсказуемость даже формально «сложных» паролей с заглавными буквами и спецсимволами. Вместе с этим, проведенный кореляционный анализ OSINT-ключевых слов, извлечённых через Serper API, продемонстрировал выраженную положительную связь между паролями и личными именами ($r \approx 0{,}65$; см. Рисунок 7); где частичные совпадения строк и обнаруженная семантическая близость также увеличивают риск угадывания. Все эти наблюдения поддерживают стратегию персонализированного контроля предсказуемости при формировании паролей, где через контексные подсказки и блокировок определенных персонально клишированных «масок» достигается более высокая непредсказуемость при формирования паролей до их сохранения в системе.

Ниже, в таблице 1 мы провели сравнение полученных результатов с базовыми методами, где модель LSTM показала более высокую точность предсказания лучше оценивала вероятность структуры паролей а также точнее отражает поведенческие особенности испытуемых.

Таблица 1. Сравнение результатов моделей

Модель	Точность	Avg Log-Score	Адаптируемость	Ручное проектирование признаков	Моделирование поведения человека
Decision Tree	81.4%	0.421	Низкая	Требуется	Отсутствует
PCFG	88.2%	0.318	Средняя	Требуется	Частично
LSTM	96.7%	0.224	Высокая	Не требуется	Реализовано

Результаты проведенных опытов подтверждает статистику (All About Cookies, 2023) что большинство пользователей, около 84%, ставят небезопасные пароли опираясь на легко предсказуемые паттерны, такие как: любимые числа (24%), имя питомцев (23%), дата рождения (19%), имена членов семьи или партнеров (16%), любимые произведения (8%), памятные даты (8%) или вообще повторное использование того же пароля в той же системе (17%), если это не запрещено.

Сопоставление с онлайн-проверками также демонстрирует ограниченность эвристик. PasswordMonster (PasswordMonster. (n.d.)) имитирует реальные атаки (словари, подстановки а→@, е→3, последовательности 12345/qwerty) и оценивает время взлома, выступая одновременно диагностическим и образовательным инструментом. Напротив, Kaspersky Password Checker (Kaspersky Lab, 2025) и Security.org (Security.org, 2025) полагаются на простые правила (длина, классы символов), игнорируя структурные/поведенческие паттерны. UIC Password Checker (University of Illinois Chicago. (n.d.)) — правило-ориентированная система с фиксированными критериями. Подобные методы склонны ошибаться и не отражают фактическую стойкость, что подчёркивает потребность в интеллектуальных адаптивных подходах.

Для полноты приведено сравнение классов методов, включая PassGAN в таблице 2(Hitaj, B., et al., 2019), генерирующий «человекоподобные» пароли. Хотя PassGAN силён в массовой генерации кандидатов, ему не хватает интерпретируемости и возможностей обратной связи в реальном времени, необходимых для персонализированных рекомендаций.

Таблица 2. Сравнение различных подходов

Критии	Decision Tree	PCFG	PassGAN	LSTM
Тип модели	Дискретная логика на деревьях	Статистическая модель на грамматике	Генеративно-состязательная сеть	Рекуррентная нейросеть
Обработка последовательностей	Не поддерживается	Ограничена заданной грамматикой	Полная поддержка (генеративная выборка)	Полная поддержка (контекстное моделир.)
Требуется ручное проектирование признаков	Да	Да	Нет	Нет

Гибкость для новых паттернов	Низкая	Средняя	Высокая	Высокая
Учёт человеческого фактора	Практически не учитывается	Частичный учёт через вероятности	Косвенно, по распределению данных	Присутствует, моделирует реальное поведение
Устойчивость к нетипичным данным	Низкая	Средняя	Высокая	Высокая
Интерпретируемость	Высокая	Средняя	Очень низкая (чёрный ящик-генератор)	Низкая (чёрный ящик)
Представление пароля	Дерево логических правил	Структура + частоты сегментов	Латентное векторное кодирование	Эмбеддинги и память
Основное ограничение	Требует ручного отбора признаков	Не выявляет вложенные паттерны	Недостаток интерпретируемости и персональной обратной связи	Требует больших данных и времени обучения

Для задач цифровой криминалистики в настоящее время представлена модульная система семантического анализа сообщений, которая комбинирует векторный поиск, классификацию и визуальную аналитику (FastAPI, Qdrant, Llama-Guard-3-1B, React). Платформа сосредоточена на семантической интерпретации и категоризации, отличаясь от традиционных решений. Время построения индекса на корпусе из 7448 сообщений составило 3 с; многоязычные запросы с динамической маршрутизацией на перевод обрабатываются в среднем за 5,23 с, а запросы без перевода — 0,45 с. Классификатор Llama-Guard-3-1B достиг точности 75,19 % при пакетной обработке; типичные ошибки, включая смешение близких категорий и ложные негативы на трех «тонких» кейсах, показывают, что доменная адаптация и балансировка классов являются необходимыми. Платформа облегчает поиск узлов социальных графов, временных аномалий и критичных категориальных паттернов, визуализируя структурные и поведенческие характеристики коммуникаций (по приложениям, получателям, геолокациям, динамике и статусам).

Заключение

Предложенное решение является отечественным альтернативным решением проверки качества паролей пользователей и при интеграции в корпоративную сеть может являться продуктивным инстру

В совокупности результаты показывают: (i) в парольном сценарии LSTM-подход переводит контроль качества пароля от формальных правил к персонализированному on-device анализу, снижая предсказуемость и повышая реальную стойкость; (ii) инфраструктура семантической аналитики

пригодна для экспресс-разведки больших многоязычных массивов текста в криминалистике при приемлемой латентности. Перспективные направления включают расширение персональных словарей (в т.ч. семантическая близость), автоматическую генерацию «анти-масок», интеграцию с менеджерами паролей и корпоративными SSO; мультимодальный OSINT (LinkedIn, X/Twitter, Instagram) с агентами CV/NLP и связкой сущностей для более точного профилирования; а также лонгитюдное исследование, оценивающее, от каких небезопасных привычек (персональные данные, популярные паттерны, минимальные правки старых паролей) пользователи отказываются под воздействием контекстной обратной связи. Дополнительно для криминалистического модуля планируются расширение обучающих корпусов, hard-negative mining, добавление аудио-/видеоаналитики и интеграция с платформами OSINT/Threat Intelligence.

Таким образом, предложенная система сочетает персонализированную оценку стойкости паролей в реальном времени и масштабируемую семантическую аналитику текстов, закрывая разрыв между «безопасностью по правилам» и безопасностью на практике.

Благодарности

Данное исследование проведено при финансовой поддержке Комитета науки Министерства науки и высшего образования Республики Казахстан в рамках договора №388/ПЦФ-24-26 от 01.10.2024 по научному проекту BR24993232 «Разработка инновационных технологий проведения цифровых криминалистических исследований с применением интеллектуальных программно-аппаратных комплексов».

Этическое одобрение

Данное исследование не затрагивает участие людей или животных.

References

Bergeron A., & Dearden T.E. (2024) Secret sharing in online communities: A comparative analysis of offender and non-offender password creation strategies. Journal of Economic Criminology, 6, 100110. *https://doi.org/10.1016/j.jeconc.2024.100110* (in English)

Atzori M., Calò E., Caruccio L., Cirillo S., Polese G., & Solimando G. (2024) Evaluating password strength based on information spread on social networks: A combined approach relying on data reconstruction and generative models. Online Social Networks and Media, 42, 100278. *https://doi.org/10.1016/j.osnem.2024.100278* (in English)

He D., Yu H., Zhou B., Zhu S., Zhang M., Chan S., & Guizani M. (2021) How does social behavior affect your password? IEEE Network, 35(5). — P. 284–289. https://doi.org/10.1109/mnet.101.2000762 (in English)

Buckman B. (n.d.-b) 36 Must-Know Password Statistics for 2025. Huntress. https://www.huntress.com/blog/password-statistics? (in English)

Data breach statistics in 2024 - Surfshark (2020, December 9) Surfshark. *https://surfshark.com/research/study/data-breach-recap-2024* (in English)

Urrico R. (2024, February 13) Reports show rising ransomware attacks and bad password habits threaten financial accounts, among others. Finopotamus. https://www.finopotamus.com/post/reports-show-rising-ransomware-attacks-and-bad-password-habits-threaten-financial-accounts-among-ot (in English)

Password dataset (2024, April 25) Kaggle. https://www.kaggle.com/datasets/soylevbeytullah/password-datas 7 (in English)

Kkrypt0nn. (n.d.-b) wordlists/wordlists/passwords at main kkrypt0nn/wordlists. GitHub. https://github.com/kkrypt0nn/wordlists/tree/main/wordlists/passwords 8 (in English)

Melicher W., Ur B., Segreti S.M., Komanduri S., Shay R., Bauer L., Christin N., & Cranor L.F. (2016) Fast, lean, and accurate: Modeling password guessability using neural networks. In 25th USENIX Security Symposium. — P. 175–191. USENIX Association. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/melicher (in English)

All About Cookies. (2023, October 12). 84% of people use unsafe passwords: Password behavior survey. https://allaboutcookies.org/password-users-behavior-survey (in English)

Pass word Monster (n.d.) Password Strength Meter. https://www.passwordmonster.com/ (in English)

Kaspersky Lab. (2025) Password Checker & Secure Random Password Generator. https://password.kaspersky.com/ (in English)

Security.org. (2025) How Secure Is My Password?. Password Strength Checker. https://www.security.org/how-secure-is-my-password (in English)

University of Illinois Chicago. (n.d.) Password Strength Test. https://www.uic.edu/apps/strong-password/(in English)

Hitaj B., Gasti P., Ateniese G., & Perez-Cabo A. (2019) PassGAN: A deep learning approach for password guessing. In A. Biggio & F. Roli (Eds.), Advances in information security. — Vol. 79. — P. 1–20. Springer. https://doi.org/10.1007/978-3-030-29959-5_1 (in English)

Rzayeva L., Pogolovkin D., & Myrzatay A. (2025) DEVELOPMENT OF A MODULAR NLP-BASED CORRESPONDENCE ANALYSIS SERVICE FOR DIGITAL FORENSICS. News of the National Academy of Sciences of the Republic of Kazakhstan. Physico-Mathematical Series, (2). — P. 212–233. https://doi.org/10.32014/2025.2518-1726.354 (in English)

Idrissova M., Kim S., Amirgaliyev B., Yedilkhan D., & Rzayeva L. (2025) DIGITAL FOOTPRINTS: CLUSTERING BROWSER HISTORY FOR USER PROFILING USING MACHINE LEARNING. Journal of Problems in Computer Science and Information Technologies, 3(2). — P. 16-28. https://doi.org/10.3390/info16080655 (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 22.12.2025.
Формат 60x881/8. Бумага офсетная.
Печать –ризограф. 20,0 п.л. Заказ 4.

«Central Asian Academic Research Center» LLP
Алматы, Қонаев к-сі, 142