Аннотация к диссертационной работе Токтаровой Айгерім Бастарбеккызы на соискание степени доктора (PhD) по специальности 8D06115 – «Информационные системы» по теме «Обнаружение ненавистных речей с использованием методов обработки естественного языка и машинного обучения в онлайн-пользовательском контенте»

АННОТАЦИЯ

Актуальность исследования. В современном цифровом ландшафте алгоритмы машинного обучения позволяют автоматизировать и эффективно составлять базы данных языка с контекстом ненависти, вражды и скрытых угроз в социальных сетях. Язык вражды охватывает устные или письменные выражения, направленные на оскорбление или дискриминацию отдельных лиц или групп, затрагивающие такие темы, как религия, этническая принадлежность, пол и некоторые другие.

С ростом числа пользователей Интернета распространение вредоносных и оскорбительных комментариев в социальных сетях становится все более серьезной проблемой. Язык ненависти, включая угрозы и уничижительные заявления, существенно влияет на психологическое благополучие, что приводит к повышению рисков для жертв киберпреступлений в социальных сетях и других онлайн-платформах.

Язык ненависти проявляется различными способами, такими как оскорбительные замечания, распространение вредоносных слухов, публикация уничижительных изображений или видео и исключение из социальных сетей. Последствия воздействия такого контента могут быть серьезными и приводить к тревоге, депрессии, низкой самооценке и даже суицидальным мыслям. Онлайн-опрос, проведенный Facebook в 2022 году среди учащихся, посещающих летние лагеря, показал, что среди пользователей социальных сетей в Казахстане, 75% молодежи сталкивалась с той или иной формой кибербуллинга. Психологические последствия включают бессонницу, потерю аппетита, пренебрежение к уходу за собой и снижение успеваемости или социальной успеваемости. В крайних случаях постоянное воздействие онлайн-ненависти может приводить к трагическим последствиям. В 2024 году 165 подростков в Казахстане покончили жизнь самоубийством, а 378 несовершеннолетних пытались покончить жизнь самоубийством из-за оскорбительных или унизительных комментариев в Интернете.

Примерно 5% подростков подвергаются онлайн-травле два-три раза в месяц, а 12% сообщают как минимум об одном случае. Социальные и профессиональные последствия языка ненависти наносят ущерб репутации и создают трудности не только в построении дальнейших отношений, но и в трудоустройстве. Статистика показывает, что более трети жертв онлайн-травли получали личные сообщения, содержащие дискриминационные или

оскорбительные выражения, 24% сталкивались с унизительными сообщениями, а 31% сталкивались с негативными комментариями под своими фотографиями.

Просветительская деятельность в данном вопросе играет решающую роль в снижении онлайн-рисков. Вебинары, учебные пособия и информационно-образовательные мероприятия по повышению осведомленности как педагогов, так и родителей, могут помочь снизить распространенность языка вражды и связанного с ним вреда. Формирование культуры уважения и ответственности в онлайн-пространствах может способствовать созданию более безопасной цифровой среды для всех пользователей.

По состоянию на 2024 год уровень проникновения интернета в Казахстане составляет 92,3%, что составляет 18,9 млн. постоянных пользователей от общей численности населения в 20,07 млн. человек. Социальными сетями пользуются 14 млн. пользователей, или 71,5% населения. Учитывая постоянную доступность онлайн-информации и анонимность, предоставляемую цифровыми платформами, люди становятся все более уязвимыми для киберпреследований и атак в любое время.

Психологи подчеркивают, что эмоциональное насилие, включая онлайн-язык ненависти, может иметь более глубокое психологическое воздействие, чем физическое насилие. Поэтому выявление и смягчение вредного онлайн-контента является важнейшим приоритетом для исследований в области информационных технологий. В данной диссертации исследование касалось именно такой проблемы и ее решения путем реализации автоматизированного обнаружения подобного текста и слов с помощью машинного обучения.

Цель диссертационного исследования. Основной целью данного исследования является разработка модели глубокой нейронной сети, способной автоматически обнаруживать проявления ненависти в текстовых данных.

Задачи исследования:

Для достижения этой цели были поставлены следующие задачи:

1. Провести комплексный анализ алгоритмов машинного обучения для бинарной и многоклассовой классификации оскорбительных комментариев.

2. Собрать и предварительно обработать данные на казахском языке для обучения алгоритмов машинного обучения (ML) и глубокого обучения (DL).

3. Изучить основные архитектуры глубокого обучения и различные методы DL, включая:

а) сверточные нейронные сети (CNN);

б) глубокие нейронные сети (DNN);

в) рекуррентные нейронные сети (RNN);

г) сети с долговременной краткосрочной памятью (LSTM);

д) многослойные персептроны (MLP);

е) модели глубокого обучения на основе искусственного интеллекта.

4. Провести экспериментальные исследования, сравнить модели и оптимизировать гиперпараметры для повышения эффективности классификации.

Объект исследования – Социальные сети (ВКонтакте, Instagram, YouTube, Twitter) и новостные порталы (Nur.kz, Tengrinews).

Методы исследования: Алгоритмы машинного обучения и глубокого обучения для обнаружения языка вражды в текстовых данных.

Научная новизна исследования заключается в создании набора данных для казахского языка, предварительно обработки и дополнительного тестирования их для алгоритмов машинного и глубокого обучения. А также в разработке и обучении гибридной глубокой нейронной сети для определения слов языка вражды и ненависти.

Теоретическая значимость: изучены существующие исследования по обнаружению неоднозначных языковых выражений в текстовых данных и оценены инструменты обработки естественного языка (NLP) для автоматизированной классификации.

Практическая значимость: исследование имеет прикладной вклад в разработку и обучение глубоких нейронных сетей для обнаружения языка вражды. Полученные результаты подтверждают достижения в приложениях NLP для фильтрации и модерирования вредоносного контента в социальных сетях.

Основные положения, выносимые на защиту:

1. Сбор, предварительная обработка и ручная классификация набора данных на казахском языке для задач машинного и глубокого обучения.

2. Разработка и обучение глубокой нейронной сети, включающей механизм внимания для бинарной классификации языка вражды.

3. Сравнительный анализ алгоритмов машинного обучения и глубокого обучения для обнаружения языка вражды в текстах на казахском языке.

Основные результаты диссертационной работы:

1. Разработка набора данных на казахском языке:

Был собран, аннотирован и предварительно обработан полный набор данных на казахском языке, что создало фундаментальный ресурс для обучения и оценки алгоритмов машинного обучения и глубокого обучения. Для обеспечения согласованности и точности представления данных были реализованы ручная категоризация и лингвистическая нормализация.

2. Создание гибридной модели глубокой нейронной сети:

Была разработана и реализована уникальная гибридная архитектура глубокого обучения, которая интегрирует BERT с механизмом внимания. Модель была разработана для бинарной категоризации языка ненависти в онлайн-тексте и продемонстрировала мастерство в улавливании контекстуальных и семантических тонкостей оскорбительного языка.

3.Сравнительная оценка алгоритмов машинного обучения и глубокого обучения:

Были проведены комплексный анализ и сравнительный анализ нескольких алгоритмов, включая CNN, RNN, LSTM, MLP и BiLSTM, в сравнении с предлагаемой гибридной моделью. Исследование продемонстрировало улучшенное контекстуальное понимание и гибкость моделей на основе трансформатора при обработке сложных языковых шаблонов в многоязычных условиях.

4.Эффективные методы предварительной обработки для зашумленных онлайн-данных: исследование разработало устойчивые структуры предварительной обработки, подходящие для управления неформальным языком, переключением кодов, смайликами и специальными символами, обычно встречающимися в общении в социальных сетях. Это значительно повысило надежность модели в практических контекстах применения.

5. Эффективность предлагаемой модели оценивалась с использованием данных с казахстанских новостных сайтов Contur.kz и Serke.org. Эти платформы были выбраны за их отличительный акцент на социально-политических комментариях и редакционных материалах, часто отражающих устойчивое общественное мнение и потенциально деликатные дискуссии. Оценка выявила эффективность модели в тестовом режиме в точном обнаружении слов, связанных с ненавистью, в официальных, эмоционально заряженных новостных статьях.

Личный вклад докторанта: докторант провел обширный обзор литературы, проанализировал патенты, имеющие отношение к теме диссертации, выбрал соответствующие исследовательские методики и провел теоретические и экспериментальные исследования. Кандидат также сыграл центральную роль в практической реализации результатов исследования.

Подтверждение результатов исследования: основные выводы диссертации были представлены в следующих научных публикациях и на международных конференциях:

Апробация результатов исследования. Основные результаты диссертационной работы докладывались на семинарах и заседаниях кафедры «Компьютерная инженерия», в Международном казахско-турецком университете имени Ходжи Ахмеда Ясави, на международных конференциях, которые проходили в Астане, в Индии:

1. Расширенное заседание кафедры «Компьютерная инженерия», №6, 19.02.2025.

2. 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST), Астана

3. The11th International Conference on Cloud Computing, Data Science & Engineering (Confluence),2021, Индия

Статьи, опубликованные в журналах, индексируемых в базах Web of Science и Scopus:

1. Toktarova, A., Abushakhma, A., Adylbekova, E., Manapova, A., Kaldarova, B., Atayev, Y., ... & Aidarkhanova, A. (2023). Offensive language identification in low resource languages using bidirectional long-short-term memory network. International Journal of Advanced Computer Science and Applications, 14(6). DOI: http://dx.doi.org/10.14569/IJACSA.2023.0140687

2. Toktarova, A., Syrlybay, D., Myrzakhmetova, B., Anuarbekova, G., Rakhimbayeva, G., Zhylanbaeva, B., ... & Kerimbekov, M. (2023). Hate speech detection in social networks using machine learning and deep learning methods. International Journal of Advanced Computer Science and Applications, 14(5).

DOI: http://dx.doi.org/10.14569/IJACSA.2023.0140542

Статьи, опубликованные в научных журналах, рекомендованных Комитетом по контролю в сфере образования и науки МОН РК:

1. B. Toktarova, B.S.Omarov, G.N. Kazbekova, S.AMamikov, F.E. Temirbekova Collecting hate speech database on social network in Kazakh language by using machine learning/ Bulletin of the National Academy of Sciences of the Republic of Kazakhstan. Series of Physics and Mathematics, 1991-346X, No. 1, pp. 191-203, 2023, DOI:https://doi.org/10.32014/2023.2518-1726.177

2. A.B. Toktarova, Zh.Zh. Azhibekova, D.R. Sultan, M.A. Kerimbekov “Collecting hate speech in Kazakh language in online content using machine learning”/ Bulletin of Abai KazNPU, series “Physics and Mathematics Sciences”, vol. 81, No. 1, 2023

3. A.B. Toktarova, B.S. Omarov, Zh.Zh. Azhibekova, S. A. Mamikov “The importance of artificial intelligence in identifying offensive words in online content” Bulletin of Toraigyrov University. ISSN 2710-3420 Energy series. No. 1, p. 311-322, 2023

4. A.B. Toktarova, B.S. Omarov, Zh.Zh. Azhibekova, Automated classification of offensive words using the “emotional” opinions of network users / Bulletin of the National Academy of Sciences of the Republic of Kazakhstan, No. 2 (88), DOI:https://doi.org/10.47533/2023.1606-146X.9

5. A.B. Toktarova, B.S. Omarov, B.A. Kaldarova, Using bilstm in identifying offensive words from low-resource languages Bulletin of the National Academy of Sciences of the Republic of Kazakhstan. Physics and Mathematics Series, 1991-346X, No. 3, pp. 174-189, 2024, DOI: 10.32014/2024.2518-1726.299

6. A.B. Toktarova, B.S.Omarov, G.I. Beissenova, R.B. Abdrakhmanov Analysis of hate speech words in online content by using data mining Bulletin of the National Academy of Sciences of the Republic of Kazakhstan. Series: Physics and Mathematics, No. 2 (346), pp. 237-251, 2023, https://doi.org/10.32014/2023.2518-1726.196

7. A.B. Toktarova, B.S. Omarov, F.S. Temirbekova "Classification of Kazakh language obscene words and adaptation of machine learning methods to their detection" / Bulletin of KazNPU named after Abai, series "Physical and Mathematical Sciences", Vol. 82, No. 2 (2023)

8. A.Toktarova, Zh.Azhibekova, A.Aliyeva, N.Sarsenbieva "Bidirectional long short-term memory in hate speech detection problem on networks" / Bulletin of KazNPU named after Abai, series "Physical and Mathematical Sciences" Vol. 87, No. 3 (2024), DOI: 10.51889/29595894.2024.87.3.010

В трудах международных конференций:

1. Toktarova, A., Sultan, D., & Azhibekova, Z. (2024, May). Review of Machine Learning Models in Cyberbullying Detection Problem. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 233-238). IEEE.
2. Sultan, D., Suliman, A., Toktarova, A., Omarov, B., Mamikov, S., & Beissenova, G. (2021, January). Cyberbullying detection and prevention: Data mining in social media. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 338-342). IEEE.

Свидетельство о государственной регистрации прав на объект авторского права. База данных «Обнаружение ненавистных речей с использованием методов обработки естественного языка и машинного обучения в онлайн-пользовательском контенте». Запись в реестре РГП НИИС МЮ РК за №54674 от «14» февраля 2025 г.

Акт внедрения. Разработанная модель опробирована на информационных сайтах contur.kz и serke.org в тестовом режиме 31.01.2025 г.

Объем и структура диссертации.

Диссертационная работа состоит из введения, четерех глав, заключения. Работа выполнена печатным способом на 121 страницах, с применением компьютерных возможностей акцентирования внимания в виде иллюстраций, схем и таблиц. Список литературы состоит из наименований.

Автор выражает глубокую благодарность научному руководителю доктору философии (PhD), ассоциированному профессору Международного университета информационных технологий Омарову Батырхану Султановичу и зарубежному консультанту профессору Стамбульского технического университета Ешрефу Адалы (Турция, г. Стамбул) за неоценимую работу и консультации в ходе исследования.