https://doi.org/10.53360/2788-7995-2026-3(23)-3

BY 4.0
Check for updates

МРНТИ: 28.23.37

Р.Б. Абдрахманов[1], Ә.С. Төлеп[2], Д.Р. Сұлтан[3], А.С. Касымбеков[1], А.Б. Нуржанова[4]*
[1]Международный университет туризма и гостеприимства,
161200, Республика Казахстан, г. Туркестан, ул. Рабига Султан Бегим, 14А
[2]Международный казахско-турецкий университет им.Х.А.Ясави,
161200, Республика Казахстан, г. Туркестан, проспект Бекзата Саттарханова, 29
[3]Университет Нархоз,
050035, Республика Казахстан, г. Алматы, ул. Жандосова, 55
[4]Евразийский национальный университет имени Л.Н. Гумилева,
010008, Республика Казахстан, г. Астана, ул. Сатпаева, 2
*e-mail: nurzhanova_ab@enu.kz

ПОДХОДЫ ГЛУБОКОГО ОБУЧЕНИЯ К ОБНАРУЖЕНИЮ КИБЕРБУЛЛИНГА В МНОГОЯЗЫЧНОМ ОНЛАЙН-ТЕКСТОВОМ КОНТЕНТЕ

Аннотация: *В настоящем исследовании проведён системный сравнительный анализ эффективности классических алгоритмов машинного обучения и современных архитектур глубокого обучения для автоматического выявления кибербуллинга в многоязычных онлайн-текстах (русский и казахский языки). Особое внимание уделено сопоставлению методов на основе традиционных признаков (TF-IDF, n-граммы) и моделей, использующих контекстные представления текста. Для казахоязычного корпуса разработана и вручную аннотирована специализированная база данных, обеспечивающая высокую достоверность разметки в условиях низкой ресурсной обеспеченности языка и ограниченного объёма обучающих выборок.*

Экспериментальные результаты показали, что на русскоязычном корпусе наилучшие показатели продемонстрировала модель с механизмом внимания (Attention-based model), достигнув Accuracy 92,1%, F1-score 92,9% и ROC-AUC 0,95. На казахоязычном корпусе данная архитектура также показала лучшие результаты: Accuracy 91,5%, F1-score 92,9% и ROC-AUC 0,94. Рекуррентные модели LSTM и BiLSTM продемонстрировали стабильные показатели в диапазоне 90–91%, тогда как классические алгоритмы машинного обучения уступали по всем основным метрикам качества. Полученные результаты подтверждают эффективность глубокого обучения для выявления кибербуллинга в многоязычных текстах.

Работа подчёркивает важность интеграции мультимодальных и интерпретируемых подходов, расширения языковых ресурсов, а также применения transfer learning и предобученных языковых моделей для повышения устойчивости и точности классификации. Полученные результаты имеют практическую значимость для систем интеллектуальной модерации контента, обеспечения цифровой безопасности и раннего выявления рисков психологического состояния пользователей в онлайн-среде.

Ключевые слова: *кибербуллинг, многоязычный текст, глубокое обучение, рекуррентные нейронные сети, механизм внимания.*

Введение

Современная цифровая среда характеризуется быстрым ростом пользовательского контента в социальных медиа, что, наряду с расширением коммуникационных возможностей, способствует распространению кибербуллинга. Автоматическое выявление таких проявлений осложняется высокой лингвистической вариативностью (сленг, эмодзи, сарказм, код-свитчинг) и морфологической сложностью языков, особенно в условиях низкоресурсности казахского языка и ограниченности размеченных данных.

В этих условиях ключевую роль играют методы машинного и глубокого обучения. Классические алгоритмы обеспечивают базовую классификацию, однако ограничены в учёте контекста, тогда как современные нейросетевые архитектуры (RNN, CNN, Attention) эффективно выявляют сложные семантические зависимости в эмоционально окрашенных текстах.

©Р.Б. Абдрахманов, Ә.С. Төлеп, Д.Р. Сұлтан, А.С. Касымбеков, А.Б. Нуржанова, 2026

Целью работы является разработка и сравнительный анализ таких методов для выявления кибербуллинга в русскоязычных и казахоязычных текстах с использованием процедур предобработки и метрик оценки (accuracy, recall, F1-score, ROC-AUC). Исследование направлено на создание устойчивого и переносимого подхода для задач автоматической модерации и цифровой безопасности.

В ряде исследований подтверждается высокая эффективность методов глубокого обучения для обнаружения кибербуллинга в многоязычных текстах. Так, применение CNN и гибридных моделей CNN+LSTM позволило достигать точности до 0,82 на бенгальском и читтагонском языках, демонстрируя их применимость к морфологически сложным и низкоресурсным языкам [1].

В другом исследовании предложена мультимодальная модель HueID для языка урду, сочетающая Bi-LSTM для текстовых признаков и CNN для обработки эмодзи, что обеспечивает высокую точность классификации [2]. Отмечается также необходимость расширения многоязычной поддержки, поскольку большинство существующих решений ориентировано на английский язык [3].

Гибридные ансамблевые архитектуры (Bi-LSTM, Bi-GRU, CNN, LSTM) с использованием эмбеддингов GloVe демонстрируют высокую эффективность при обработке многоязычных данных, подтверждая потенциал комплексных моделей [4]. При этом дальнейшие исследования направлены на развитие многоязычных и мультимодальных подходов для повышения качества распознавания [5].

Несмотря на наличие работ, ориентированных на английские корпуса, вопросы многоязычной обработки остаются недостаточно изученными [6]. Аналогично, крупномасштабные системы на основе глубокого обучения не всегда учитывают специфику многоязычного контента [7].

Современные подходы фокусируются на интеграции мультимодальных данных и учёте морфологических особенностей языков [8]. Наряду с этим, гибридные методы машинного обучения (например, LinearSVC с TF-IDF) демонстрируют высокую точность (до 92,1%) и пригодность для реального времени [9].

В заключение, исследования с использованием моделей Bi-LSTM и BERT подтверждают эффективность глубокого обучения, однако подчеркивают необходимость дальнейшего развития многоязычных и мультимодальных решений для повышения качества выявления кибербуллинга [10].

Материалы и методы исследования

Сбор и подготовка данных: в рамках исследования разработан специализированный программный модуль (парсер), обеспечивающий автоматизированный сбор текстовых данных из социальных сетей и открытых медиаресурсов с учётом их тематической релевантности и лингвистического разнообразия. В результате сформированы два корпуса: русскоязычный, полученный преимущественно автоматизированным способом с последующей частичной верификацией, и казахоязычный, собранный и аннотированный вручную с привлечением экспертов-лингвистов, что позволило обеспечить высокую достоверность разметки в условиях низкоресурсности языка.

Практическая реализация исследования включала несколько последовательных этапов. На первом этапе был разработан программный парсер для автоматизированного сбора текстовых сообщений из открытых интернет-источников, включая социальные сети, комментарии пользователей и публичные обсуждения. При сборе данных учитывались тематическая релевантность сообщений, наличие агрессивной или оценочной лексики, а также языковая принадлежность текста. После первичного сбора выполнялась фильтрация данных: удалялись дубликаты, рекламные сообщения, нерелевантные фрагменты и тексты с недостаточным содержательным объёмом.

На втором этапе были сформированы русскоязычный и казахоязычный корпуса. Казахоязычный корпус был дополнительно проверен и аннотирован вручную, поскольку казахский язык относится к низкоресурсным языкам и имеет ограниченное количество доступных размеченных наборов данных. Разметка выполнялась по бинарной схеме: сообщения с признаками кибербуллинга и сообщения без признаков кибербуллинга. Для повышения достоверности использовалось межэкспертное согласование, что позволило снизить субъективность при определении спорных случаев.

На третьем этапе была проведена предобработка текстов. Она включала приведение текста к нижнему регистру, удаление лишних символов, ссылок, повторяющихся пробелов, токенизацию, удаление стоп-слов, лемматизацию и обработку эмодзи. Для классических моделей тексты были преобразованы в числовой вид с использованием Bag-of-Words и TF-IDF. Для нейросетевых моделей использовались последовательностные представления текста, позволяющие учитывать порядок слов и контекстные зависимости.

На четвёртом этапе корпуса были разделены на обучающую и тестовую выборки. На обучающей выборке проводилось обучение моделей, а на тестовой – оценка их качества. В качестве классических алгоритмов использовались Logistic Regression, Naive Bayes, SVM, KNN, Decision Tree и Random Forest. В качестве моделей глубокого обучения были реализованы LSTM, BiLSTM, CNN и модель с механизмом внимания. Реализация экспериментов выполнялась в среде Python с использованием библиотек Scikit-learn, TensorFlow/Keras, NumPy и Pandas.

Оценка качества моделей проводилась по метрикам Accuracy, Precision, Recall, F1-score и ROC-AUC. Дополнительно использовались матрицы ошибок и ROC-кривые, что позволило более подробно проанализировать характер ошибок классификации. Сравнение результатов показало, что нейросетевые модели лучше справляются с контекстно зависимыми сообщениями, а модель с механизмом внимания наиболее эффективно выделяет семантически значимые фрагменты текста.

Практическая значимость проведённой работы заключается в том, что предложенный подход может быть использован при разработке интеллектуальных систем автоматической модерации онлайн-контента. Такая система способна выявлять потенциально агрессивные сообщения, передавать их на дополнительную проверку модератору и тем самым снижать риски распространения кибербуллинга в цифровой среде.

Реализованные модели: в исследовании реализован и сопоставлен широкий спектр моделей, охватывающий как классические алгоритмы машинного обучения, так и современные нейросетевые архитектуры. В качестве базовых моделей использовались Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree и Random Forest. Для них применялись методы векторизации текста (Bag-of-Words, TF-IDF), обеспечивающие переход к числовому признаковому пространству с последующим обучением классификаторов.

Глубокие модели включали RNN, LSTM, BiLSTM, CNN, а также многослойный персептрон. Эти архитектуры позволяют учитывать последовательностную природу текста и извлекать скрытые семантические зависимости. Ключевое внимание уделено модели с механизмом внимания (Attention), которая реализует адаптивное взвешивание скрытых представлений входной последовательности. Формально механизм внимания вычисляется как нормализованная (например, с использованием softmax) функция сходства между скрытыми состояниями, что позволяет модели акцентировать наиболее информативные элементы текста. Это особенно эффективно при анализе кибербуллинга, где семантически значимые признаки могут быть распределены неравномерно по последовательности.

Метрики оценки: оценка эффективности моделей проводилась на основе совокупности стандартных метрик классификации: Accuracy, Precision, Recall, F1-score и ROC-AUC. Использование нескольких метрик обусловлено необходимостью комплексного анализа качества моделей, особенно в условиях возможного дисбаланса классов. Дополнительно применялись матрицы ошибок (confusion matrix) для детализированного анализа распределения ошибок классификации, а также ROC-кривые для оценки дискриминационной способности моделей при различных порогах принятия решений.

Сравнительный анализ проводился отдельно для русскоязычного и казахоязычного корпусов, что позволило выявить влияние морфологических, синтаксических и ресурсных особенностей языков на эффективность алгоритмов. Такой подход обеспечивает воспроизводимость результатов и позволяет обоснованно оценить применимость разработанных моделей в задачах многоязычной автоматической модерации контента.

Характеристика русскоязычного и казахоязычного корпусов. Для проведения исследования были сформированы два специализированных корпуса текстовых сообщений: русскоязычный и казахоязычный. Источниками данных послужили открытые интернет-

ресурсы, включая социальные сети, форумы, комментарии пользователей и публичные дискуссионные площадки. На этапе подготовки данных были удалены дубликаты, рекламные сообщения, спам и тексты, не относящиеся к тематике исследования.

Оба корпуса были аннотированы вручную по схеме бинарной классификации: сообщения с признаками кибербуллинга и сообщения без признаков агрессивного поведения. Для повышения качества разметки привлекались несколько экспертов, а окончательное решение принималось на основе процедуры межэкспертного согласования. Такой подход позволил минимизировать субъективность оценки и обеспечить высокую достоверность обучающих данных.

Русскоязычный корпус включает сообщения, содержащие различные формы агрессивного поведения, такие как оскорбления, унижение, угрозы и дискриминационные высказывания, а также нейтральные сообщения, использованные для формирования контрольного класса. Казахоязычный корпус формировался по аналогичному принципу и содержит тексты, отражающие особенности казахского языка, включая его агглютинативную структуру и разнообразие морфологических форм.

Созданный казахоязычный корпус представляет особую научную ценность, поскольку количество доступных размеченных наборов данных для казахского языка остается ограниченным. Использование двух корпусов позволило провести сравнительный анализ эффективности алгоритмов машинного и глубокого обучения в условиях различной языковой структуры и оценить их способность выявлять признаки кибербуллинга в многоязычной среде.

Результаты исследования

Русскоязычный корпус: экспериментальные результаты на русскоязычном корпусе демонстрируют выраженное преимущество моделей глубокого обучения над классическими алгоритмами машинного обучения (табл. 1).

Таблица 1 – Сравнение моделей (русский корпус)

Модель	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC
Naive Bayes	60.2	58.7	61.3	59.9	0.65
Logistic Regression	84.5	83.9	85.2	84.5	0.88
SVM	87.3	86.8	87.9	87.3	0.90
KNN	79.1	78.5	80.2	79.3	0.82
Decision Tree	81.4	80.9	82.1	81.5	0.84
Random Forest	88.2	87.6	88.9	88.2	0.91
LSTM	90.8	90.1	91.5	90.8	0.93
BiLSTM	91.2	90.7	91.8	91.2	0.94
CNN	89.7	89.0	90.3	89.6	0.92
Attention-based model	92.1	92.5	93.3	92.9	0.95

Лучшие результаты показала модель с механизмом внимания (Attention-based neural network): Accuracy – 92.1%, F1-score – 92.9%, ROC-AUC – 0.95, что свидетельствует о высокой обобщающей способности и устойчивости к вариативности текстов. Рекуррентные архитектуры (LSTM, BiLSTM) также продемонстрировали конкурентоспособные показатели (~90-91%), при этом BiLSTM показала небольшое улучшение за счёт учета двунаправленного контекста.

Классические методы машинного обучения оказались менее эффективными: Naïve Bayes показал наименьшую точность (~60%) из-за предположения условной независимости признаков, а ансамблевые и линейные модели (Random Forest, Logistic Regression, SVM) обеспечили умеренные результаты, уступая нейросетям в учёте сложных семантических и синтаксических зависимостей.

Казахоязычный корпус: результаты, полученные на казахоязычном корпусе, в целом согласуются с наблюдениями для русского языка, несмотря на более ограниченный объем данных и низкую ресурсность языка (табл. 2).

Лучшей моделью на казахоязычном корпусе стала архитектура с механизмом внимания: Accuracy – 91.5%, F1-score – 92.9%, ROC-AUC – 0.94, что подтверждает её высокую переносимость и эффективность для морфологически богатого языка. Рекуррентные модели (LSTM, BiLSTM) показали стабильные результаты, но их эффективность зависела от объема и качества обучающих данных, проявляясь большей вариативностью метрик при

ограниченных корпусах. Классические алгоритмы вновь уступили по всем показателям, что указывает на ограничения подходов с поверхностными признаками (TF-IDF, Bag-of-Words) для агглютинативного казахского языка.

Таблица 2 – Сравнение моделей (казахский корпус)

Модель	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC
Naïve Bayes	58.4	57.9	59.2	58.5	0.63
Logistic Regression	83.1	82.4	84.0	83.2	0.87
SVM	85.6	85.1	86.2	85.6	0.89
KNN	77.8	77.0	78.6	77.8	0.80
Decision Tree	80.2	79.6	81.0	80.3	0.83
Random Forest	86.7	86.0	87.5	86.7	0.90
LSTM	90.2	89.6	91.0	90.3	0.93
BiLSTM	90.9	90.3	91.6	90.9	0.93
CNN	89.1	88.5	89.9	89.2	0.91
Attention-based model	91.5	92.1	93.7	92.9	0.94

Сравнительный анализ и интерпретация результатов. Сравнительный анализ показал, что модели глубокого обучения превосходят классические алгоритмы благодаря способности извлекать высокоуровневые признаки и моделировать сложные нелинейные зависимости. Существенное повышение качества классификации достигается при использовании механизма внимания, обеспечивающего адаптивное выделение наиболее информативных элементов текста. Например, сообщение «Сен тағы да келдің бе? Әдеттегідей ештеңе түсінбейсің екен» не содержит прямых оскорбительных слов, однако в контексте общения может выражать унижение и насмешку над собеседником. Классические методы на основе TF-IDF и Bag-of-Words ориентируются преимущественно на частотные характеристики отдельных слов и часто не выявляют подобные случаи. В отличие от них модели BiLSTM и Attention способны учитывать контекст всей последовательности и корректно классифицировать такие сообщения как потенциальное проявление кибербуллинга.

Дополнительно установлено, что двунаправленные архитектуры (BiLSTM) повышают точность за счёт более полного учета контекста. Полученные результаты также подтверждают межъязыковую устойчивость моделей, демонстрируя их эффективность при обработке языков с различной морфологической структурой.

Обсуждение результатов и перспективы дальнейших исследований

Полученные результаты подтверждают высокую эффективность рекуррентных и гибридных нейросетевых архитектур в задачах текстовой классификации, особенно при анализе эмоционально и поведенчески окрашенного контента. Существенным научным вкладом является формирование и ручная аннотация казахоязычного корпуса, обладающего высокой ценностью в условиях ограниченности ресурсов для казахского языка и способного служить базой для дальнейших исследований в области NLP.

Вместе с тем выявлены ограничения, влияющие на интерпретацию результатов. К ним относятся зависимость глубоких моделей от объема размеченных данных, высокая вычислительная сложность их обучения, а также ограниченность анализа исключительно текстовой модальностью. Отсутствие учета мультимодальных и паралингвистических признаков (например, мимики, интонации) может снижать полноту выявления эмоционального состояния.

Таким образом, несмотря на достигнутые результаты, задача автоматического обнаружения кибербуллинга требует дальнейшего развития, включая расширение корпусов данных, интеграцию мультимодальных подходов и оптимизацию архитектур моделей.

Перспективные направления дальнейших исследований включают интеграцию трансформерных архитектур (например, BERT-подобных моделей), обеспечивающих более качественные контекстуализированные представления текста. Существенный потенциал также связан с разработкой мультимодальных систем, объединяющих текстовые, визуальные и аудиоданные, что позволяет более точно моделировать эмоциональное и психологическое состояние пользователей.

Важным направлением является применение transfer learning для переноса знаний с высокоресурсных языков на низкоресурсные, в частности казахский, что снижает зависимость от объема размеченных данных. Дополнительно, внедрение методов Explainable AI (XAI) способствует повышению интерпретируемости моделей и их применимости в социально значимых областях.

Практическую значимость представляет разработка систем раннего предупреждения для мониторинга психологического состояния пользователей, особенно в образовательной среде. При этом ключевым условием повышения эффективности моделей остается дальнейшее расширение и диверсификация казахоязычного корпуса, что формирует основу для перехода от экспериментальных решений к прикладным интеллектуальным системам.

Заключение

В настоящей работе выполнен сравнительный анализ классических алгоритмов машинного обучения и современных нейросетевых архитектур в задаче автоматического выявления кибербуллинга на русскоязычных и казахоязычных текстах. Это позволило оценить применимость и переносимость методов в условиях как высокоресурсного, так и низкоресурсного языка.

Экспериментальные результаты показывают, что модели глубокого обучения стабильно превосходят классические подходы, особенно при работе с морфологически сложными и контекстно насыщенными текстами. Наилучшие показатели достигнуты моделями с механизмом внимания, способными эффективно выделять значимые элементы последовательности, учитывать дальние контекстные зависимости и обеспечивать устойчивость к лингвистической вариативности.

Сформированный казахоязычный корпус представляет самостоятельную научную ценность и расширяет возможности исследований в области NLP для низкоресурсных языков. Практическая значимость созданного корпуса заключается в возможности его использования для обучения и тестирования интеллектуальных систем автоматической модерации контента, мониторинга цифровой безопасности и раннего выявления признаков агрессивного поведения в онлайн-среде. Кроме того, корпус может служить основой для дальнейших исследований в области обработки естественного языка и разработки специализированных языковых моделей для казахского языка.

Список литературы

1. Mahmud T.S.D.A. Exhaustive Study into Machine Learning and Deep Learning Methods for Multilingual Cyberbullying Detection in Bangla and Chittagonian Texts / T.S.D.A. Mahmud, M. Ptaszynski, F. Masui // Electronics. – 2024. https://doi.org/10.3390/electronics13091677.

2. Tahir S. HuEID: Hybrid Deep Learning for Cyberbullying Detection using Multi-Modal Urdu Text and Emojis / S. Tahir, A. Nawaz. – 2025. https://doi.org/10.1145/3769294.

3. Khafajeh H. Cyberbullying Detection in Social Networks Using Deep Learning / H. Khafajeh // The International Arab Journal of Information Technology. – 2024. – № 21(6). https://doi.org/10.34028/iajit/21/6/9.

4. Mahajan E. EnsMulHateCyb: Multilingual hate speech and cyberbully detection in online social media / E. Mahajan, H. Mahajan, S. Kumar // Expert Systems With Applications. – 2023. – № 236. – P. 121228. https://doi.org/10.1016/j.eswa.2023.121228.

5. Begum A. Cyberbullying Detection on Social Media using AI / A. Begum // International Journal For Science Technology And Engineering. – 2025. – № 13(4). – P. 7003-7010. https://doi.org/10.22214/ijraset.2025.70045.

6. Musyoka I. Multimodal Cyberbullying Detection Using Deep Learning Techniques: A Review / I. Musyoka, J. Wandeto, B. Kituku. – 2023. – P. 187-192. https://doi.org/10.1109/ict4da59526.2023.10302244.

7. Deep Learning for Multi-Labeled Cyberbully Detection / N. Islam et al // Enhancing Online Safety. – 2023. – P. 1-6. https://doi.org/10.1109/icdsns58469.2023.10245135.

8. Vijayakumar H.P.D.D. A study on deep learning algorithms for multimodal and multilingual cyberbullying detection / H.P.D.D. Vijayakumar // International Journal of Scientific Research. – 2021. – P. 50-54. https://doi.org/10.36106/IJAR/4620274.

9. Horate O. Cyber-Bullying and Harassment Detection using ML / O. Horate // International Journal for Science Technology and Engineering. – 2025. – № 13(4). – P. 5761-5766. https://doi.org/10.22214/ijraset.2025.69732.
10. AI-Powered Frameworks for the Detection and Prevention of Cyberbullying Across Social Media Ecosystems / M.A.M. Mondol et al // Techcomp Innovations. – 2025. – № 2(1). – P. 1-15. https://doi.org/10.70063/techcompinnovations.v2i1.59.

Благодарность

Данное исследование финансируется Комитетом науки Министерства науки и высшего образования Республики Казахстан (грант № AP23488900 – Автоматическое обнаружение кибербуллинга среди молодежи в социальных сетях с использованием искусственного интеллекта)».

Р.Б. Абдрахманов[1], Ә.С. Төлеп[2], Д.Р. Сұлтан[3], А.С. Касымбеков[1], А.Б. Нуржанова[4]*

[1]Халықаралық туризм және меймандостық университеті,
161200, Қазақстан Республикасы, Түркістан қ., Рабиға Сұлтан Бегім көшесі, 14А
[2]Қ.А.Ясауи атындағы Халықаралық қазақ-түрік университеті,
161200, Қазақстан Республикасы, Түркістан қ., Бекзат Саттарханов даңғылы, 29
[3]Нархоз университеті,
050035, Қазақстан Республикасы, Алматы қ., Жандосов көшесі, 55
[4]Л.Н.Гумилев атындағы ЕҰУ,
010008, Қазақстан Республикасы, Астана қ., Сәтбаев көшесі, 2
*e-mail: nurzhanova_ab@enu.kz

КӨПТІЛДІ ОНЛАЙН МӘТІНДЕРДЕ КИБЕРБУЛЛИНГТІ АНЫҚТАУДЫҢ ТЕРЕҢ ОҚЫТУ ТӘСІЛДЕРІ

Бұл зерттеуде көптілді онлайн мәтіндерде (орыс және қазақ тілдері) кибербуллингті автоматты түрде анықтау үшін классикалық машиналық оқыту алгоритмдері мен заманауи терең оқыту архитектураларының тиімділігіне жүйелі салыстырмалы талдау жүргізілді. Дәстүрлі белгілерге (TF-IDF, n-граммалар) негізделген әдістер мен мәтіннің контекстік көріністерін пайдаланатын модельдерді салыстыруға ерекше назар аударылды. Қазақ тіліндегі корпус үшін тілдік ресурстардың шектеулілігі және оқыту деректерінің аз көлемі жағдайында жоғары сапалы таңбалануды қамтамасыз ететін арнайы деректер қоры әзірленіп, қолмен аннотацияланды.

Эксперименттік нәтижелер орыс тіліндегі корпуста ең жоғары көрсеткіштерді назар механизміне негізделген модельдің (Attention-based model) көрсеткенін анықтады: Accuracy – 92,1%, F1-score – 92,9% және ROC-AUC – 0,95. Қазақ тіліндегі корпуста да аталған архитектура ең жақсы нәтижелерге қол жеткізді: Accuracy – 91,5%, F1-score – 92,9% және ROC-AUC – 0,94. LSTM және BiLSTM рекуррентті модельдері 90-91% аралығында тұрақты көрсеткіштер көрсетті, ал классикалық машиналық оқыту алгоритмдері негізгі сапа метрикалары бойынша төмен нәтижелер көрсетті. Алынған нәтижелер көптілді мәтіндердегі кибербуллингті анықтауда терең оқыту әдістерінің жоғары тиімділігін дәлелдейді.

Зерттеу мультимодальды және интерпретацияланатын тәсілдерді біріктірудің, тілдік ресурстарды кеңейтудің, сондай-ақ transfer learning және алдын ала үйретілген тілдік модельдерді қолданудың маңыздылығын көрсетеді. Алынған нәтижелер контентті интеллектуалды модерациялау жүйелерінде, цифрлық қауіпсіздікті қамтамасыз етуде және онлайн ортадағы пайдаланушылардың психологиялық тәуекелдерін ерте анықтауда практикалық маңызға ие.

Түйінді сөздер: *кибербуллинг, көптілді мәтін, терең оқыту, рекуррентті нейрондық желілер, назар механизмі.*

R.B. Abdrakhmanov[1], A.S. Tolep[2], D.R. Sultan[3], A.S. Kasymbekov[1], A.B. Nurzhanova[4]*

[1]International University of Tourism and Hospitality,
161200, Republic of Kazakhstan, Turkestan, 14A Rabiga Sultan Begim Street
[2]Khoja Akhmet Yassawi International Kazakh-Turkish University,
161200, Republic of Kazakhstan, Turkestan, 29 Bekzat Sattarkhanov Avenue
[3]Narxoz University, 050035, Republic of Kazakhstan, Almaty, 55 Zhandosov Street
[4]L.N. Gumlyov Eurasian National University, 010008, Republic of Kazakhstan, Astana, 2 Satpayev Street
*e-mail: nurzhanova_ab@enu.kz

DEEP LEARNING APPROACHES FOR CYBERBULLYING DETECTION IN MULTILINGUAL ONLINE TEXTS

This study presents a systematic comparative analysis of the effectiveness of classical machine learning algorithms and modern deep learning architectures for automatic cyberbullying detection in multilingual online texts (Russian and Kazakh). Particular attention is given to the comparison of traditional feature-based approaches (TF-IDF, n-grams) and models utilizing contextual text representations. For the Kazakh-language corpus, a specialized dataset was developed and manually annotated to ensure high-quality labeling under low-resource language conditions and limited training data availability.

Experimental results showed that the attention-based model achieved the best performance on the Russian-language corpus, reaching an Accuracy of 92.1%, an F1-score of 92.9%, and a ROC-AUC of 0.95. The same architecture also demonstrated the highest results on the Kazakh-language corpus, achieving an Accuracy of 91.5%, an F1-score of 92.9%, and a ROC-AUC of 0.94. Recurrent neural network models, including LSTM and BiLSTM, showed stable performance within the range of 90-91%, whereas classical machine learning algorithms yielded lower results across all major evaluation metrics. These findings confirm the effectiveness of deep learning approaches for cyberbullying detection in multilingual textual environments.

The study highlights the importance of integrating multimodal and interpretable approaches, expanding linguistic resources, and applying transfer learning along with pre-trained language models to improve robustness and classification accuracy. The obtained results have practical significance for intelligent content moderation systems, digital safety applications, and the early detection of psychological risk factors among users in online environments.

Key words: *cyberbullying, multilingual text, deep learning, recurrent neural networks, attention mechanism.*

Сведения об авторах

Рустам Абдрахманов – кандидат технических наук, доцент, Международный университет туризма и гостеприимства, Туркестан, Казахстан; e-mail: abdrakhmanov.rustam@iuth.edu.kz. ORCID: https://orcid.org/0000-0002-5508-389X.

Абимухан Тулеп – кандидат технических наук, доцент, МКТУ им. Х.А. Ясави, Туркестан, Казахстан; e-mail: abdimuhan.tolep@ayu.edu.kz. ORCID: https://orcid.org/0000-0002-0089-2446.

Данияр Султан – PhD, доцент факультета цифровых технологий Университета Нархоз, Алматы, Казахстан; e-mail:daniyar.sultan@narxoz.kz. ORCID:https://orcid.org/0000-0002-1611-1923.

Аббазбек Касымбеков – кандидат технических наук, доцент, Международный университет туризма и гостеприимства, Туркестан, Казахстан; e-mail: abbazbek@mail.ru. ORCID: https://orcid.org/0000-0001-8022-5637.

Айжан Нуржанова* – докторант, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан; e-mail: nurzhanova_ab@enu.kz. ORCID: https://orcid.org/0009-0006-9871-9823.

Авторлар туралы мәліметтер

Рустам Абдрахманов – техника ғылымдарының кандидаты, Халықаралық туризм және меймандостық университетінің доценті, Түркістан қ., Қазақстан; e-mail: abdrakhmanov.rustam@iuth.edu.kz. ORCID: https://orcid.org/0000-0002-5508-389X.

Әбдімухан Төлеп – техника ғылымдарының кандидаты, Қ.А. Ясауи атындағы Халықаралық қазақ-түрік университеті доценті, Түркістан қ., Қазақстан; e-mail: abdimuhan.tolep@ayu.edu.kz. ORCID: https://orcid.org/0000-0002-0089-2446.

Данияр Султан – Нархоз университетінің Цифрлық технологиялар мектебінің доценті, Алматы, Қазақстан; e-mail: daniyar.sultan@narxoz.kz. ORCID: https://orcid.org/0000-0002-1611-1923.

Аббазбек Касымбеков – техника ғылымдарының кандидаты, Халықаралық туризм және меймандостық университетінің доценті, Түркістан қ., Қазақстан; e-mail: abbazbek@mail.ru. ORCID: https://orcid.org/0000-0001-8022-5637.

Айжан Нуржанова* – докторант, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана қ., Қазақстан; e-mail: nurzhanova_ab@enu.kz. ORCID: https://orcid.org/0009-0006-9871-9823.

Information about the authors

Rustam Abdrakhmanov – Candidate of Technical Sciences, Associate Professor of the Humanities school, International University of Tourism and Hospitality, Turkistan, Kazakhstan; e-mail: abdrakhmanov.rustam@iuth.edu.kz. ORCID: https://orcid.org/0000-0002-5508-389X.

Abdimukhan Tolep – Candidate of Technical Sciences, Associate Professor, Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan; e-mail: abdimuhan.tolep@ayu.edu.kz. ORCID: https://orcid.org/0000-0002-0089-2446.

Daniyar Sultan – PhD, Ac. Associate Professor of the School of Digital Technologies, Narxoz University, Almaty, Kazakhstan; e-mail: daniyar.sultan@narxoz.kz. ORCID:https://orcid.org/0000-0002-1611-1923.

Abbazzbek Kasymbekov – Candidate of Technical Sciences, Associate Professor of the Humanities school, International University of Tourism and Hospitality; e-mail: abbazbek@mail.ru. ORCID: https://orcid.org/0000-0001-8022-5637.