АННОТАЦИЯ
диссертационной работы Сағынай Мөлдір
на тему: «Разработка моделей глубокого обучения для классификации деструктивного веб-контента на казахском языке» представленной на соискание степени доктора философии PhD по образовательной программе «8D06301 - Системы информационной безопасности»

Актуальность исследования. В современной цифровой среде автоматическое выявление деструктивного веб-контента стало значимым научным и практическим направлением обеспечения информационной безопасности и общественной стабильности. Распространяющиеся в интернете кибербуллинг, высказывания, выражающие ненависть, агрессивные призывы и противоправный контент причиняют вред отдельным людям и оказывают влияние на социальные отношения в обществе. Поэтому разработка интеллектуальных систем, способных своевременно выявлять и классифицировать такой контент, является одной из актуальных задач.

Актуальность исследования определяется тремя основными аспектами: законодательными требованиями, технологическими ограничениями и лингвистическими трудностями.

В законодательном аспекте противоправный характер действий, связанных с разжиганием социальной, национальной, этнической, расовой или религиозной розни, отражён в статье 174 Уголовного кодекса Республики Казахстан. Кроме того, Закон Республики Казахстан «О связи» содержит нормы, регулирующие распространение противоправной информации в сетях связи и интернет-пространстве. Следовательно, своевременное выявление деструктивного веб-контента является не только технической задачей, но и одним из необходимых условий обеспечения национальной информационной безопасности и правопорядка.

В технологическом аспекте большинство исследований ориентировано на английский, русский и другие языки с развитой ресурсной базой. Для выявления деструктивного веб-контента на казахском языке отсутствуют готовые классификаторы, достаточные по объёму аннотированные корпуса и системы, ориентированные на решение специализированных задач. Это обусловливает необходимость разработки новых методов и моделей автоматизированного анализа текстов.

В лингвистическом аспекте задача осложняется особенностями казахского языка, который относится к агглютинативным языкам и характеризуется сложной морфологической системой. Значение слова изменяется посредством присоединения аффиксов к корню, а смысловая окраска высказывания во многом зависит от контекста и семантических связей. Кроме того, деструктивное содержание часто выражается неявно: передаётся посредством подтекста, эмоциональной окраски и косвенных языковых конструкций. Поэтому модель анализа веб-контента на казахском языке должна учитывать наряду с контекстом морфологические и семантические особенности текста.

Таким образом, актуальность темы диссертационной работы обусловлена общественной опасностью деструктивного интернет-контента, необходимостью его своевременного выявления с учётом правовых требований и недостаточностью технологических решений для казахского языка. Указанные обстоятельства определяют научную и практическую значимость разработки усовершенствованной модели выявления деструктивного веб-контента на казахском языке.

Цель исследования – комплексное изучение и разработка моделей глубокого обучения для классификации деструктивного веб-контента на казахском языке.

Задачи исследования:

1. Сформировать набор данных деструктивных текстов на казахском языке, собранных с платформ социальных сетей.

2. Исследовать методы построения текстовых эмбеддингов, классического машинного обучения и глубокого обучения для классификации деструктивного веб-контента на казахском языке.

3. Разработать модуль морфологических признаков для казахского языка.

4. Разработать семантические признаки на основе онтологии для классификации деструктивного веб-контента на казахском языке.

5. Создать гибридную архитектуру для классификации деструктивного веб-контента по его типам и признакам с учётом особенностей казахского языка.

6. Провести испытания разработанной модели в экспериментальной среде.

Объект исследования: Деструктивный текстовый контент на казахском языке в социальных сетях и на веб-ресурсах.

Предмет исследования: Методы и модели объединения контекстных, морфологических и семантических признаков на основе онтологии для многоклассовой классификации деструктивного веб-контента на казахском языке.

Методы исследования: Методы машинного и глубокого обучения, классификации текстов, обработки естественного языка, анализа социальных сетей и онтологической классификации.

Научная новизна:

1. Создан аннотированный корпус деструктивного веб-контента на казахском языке с разметкой по нескольким классам для решения задачи его классификации.

2. Разработаны семантические признаки на основе онтологии, определяющие тип, явность выражения и целевую направленность деструктивного веб-контента на казахском языке и используемые для его классификации.

3. Разработана гибридная классификационная архитектура, объединяющая контекстные эмбеддинги, морфологические признаки казахского языка и семантические признаки на основе онтологии для классификации деструктивного веб-контента на казахском языке.

Основные положения, выносимые на защиту:

1. Для решения задачи многоклассовой классификации деструктивного веб-контента на казахском языке был разработан корпус деструктивного веб-контента и предложен метод интеграции семантических признаков на основе онтологии, характеризующих тип, явность выражения и целевую направленность контента, с эмбеддингами KazBERT и морфологическими признаками, при этом применение предложенного метода позволило повысить показатель Accuracy с 88,015% до 90,492%, а показатель Macro F1-score с 88,086% до 90,567% по сравнению с базовой моделью KazBERT.

2. Разработана гибридная архитектура для многоклассовой классификации деструктивного веб-контента на казахском языке, объединяющая эмбеддинги KazBERT, морфологические признаки, семантические признаки на основе онтологии и слой последовательностного анализа на основе BiLSTM. По сравнению с базовой моделью KazBERT разработанная архитектура позволила повысить показатель Accuracy с 88,015% до 91,045%, а показатель Macro F1-score – с 88,086% до 91,103%, обеспечив относительное снижение доли ошибочных классификаций, рассчитанной на основе Accuracy, на 25,28%.

Теоретическая значимость исследования: заключается в развитии теоретических и методологических основ классификации деструктивного веб-контента на казахском языке. В ходе исследования предложена система онтологически-ориентированных семантических признаков, описывающих тип, степень выраженности и целевую направленность деструктивного веб-контента, а также теоретически обоснована их роль в классификации текстов. Кроме того, сформированы научные основы гибридного подхода, объединяющего морфологические особенности казахского языка и контекстную семантику.

Предложенный подход вносит теоретический вклад в развитие научных исследований в области обработки естественного языка (NLP), автоматического обнаружения деструктивного контента и семантического анализа текстов на казахском языке. Результаты исследования могут служить теоретической основой для совершенствования моделей анализа контента на

казахском языке, а также для разработки интеллектуальных методов обработки текста, адаптируемых к другим тюркским языкам.

Практическая ценность. Практическая значимость определяется возможностью применения модели в системах автоматического выявления и классификации деструктивного веб-контента на казахском языке. Результаты исследования могут использоваться для мониторинга социальных сетей, веб-платформ и информационных систем, раннего выявления противоправного или деструктивного содержания, а также повышения эффективности модерации контента. Предложенные методы и разработанный набор данных создают основу для дальнейших прикладных исследований в области обработки естественного языка применительно к казахскому языку, информационной безопасности и искусственного интеллекта. По результатам исследования 27 марта 2026 года получено свидетельство о государственной регистрации № 69171 на компьютерную программу «Повышение эффективности моделей глубокого обучения на основе BERT для классификации деструктивного веб-контента на казахском языке с использованием метода Supervised Contrastive Learning (SupCon)».

Личный вклад соискателя. Все результаты диссертационной работы, выносимые на защиту, получены докторантом лично. К основным результатам относятся: система семантических признаков на основе онтологии, характеризующих тип, явность выражения и целевую направленность деструктивного веб-контента на казахском языке; гибридная классификационная архитектура, обеспечивающая выделение морфологических признаков казахского языка и их объединение с контекстными эмбеддингами KazBERT; усовершенствованная модель глубокого обучения для классификации деструктивного веб-контента, объединяющая эмбеддинги KazBERT, морфологические и онтологические семантические признаки, программное обеспечение для оценки эффективности предложенной модели, результаты экспериментальных исследований и их статистическая обработка.

Разработка алгоритмов, их программная реализация, обучение и тестирование моделей, оптимизация гиперпараметров, проведение экспериментальных исследований, сравнительный анализ полученных результатов, формулирование научных выводов и подготовка текста диссертации выполнены докторантом лично. Подготовка научных публикаций и оформление свидетельства о государственной регистрации компьютерной программы осуществлялись совместно с научными консультантами.

Связь темы диссертации с планами научно-исследовательских работ. Исследование выполнено в рамках проекта «Разработка моделей и методов выявления молодёжного экстремизма и обеспечения безопасности молодёжи в современном информационном пространстве», финансируемого Комитетом науки Министерства науки и высшего образования Республики Казахстан (грант AP19576868).

Уровень достоверности и результаты апробации. Результаты диссертационного исследования неоднократно докладывались и обсуждались на научных семинарах кафедры кибербезопасности и криптологии факультета информационных технологий и искусственного интеллекта Казахского национального университета имени аль-Фараби. Кроме того, результаты исследования были представлены на международных и отечественных научных конференциях и получили положительную оценку научного сообщества.

Вклад докторанта в подготовке каждой публикации. Результаты диссертационной работы опубликованы в 10 научных публикациях. Одна статья опубликована в издании, индексируемом в базе данных Scopus, три статьи – в материалах конференций, индексируемых в Scopus, две статьи – в журналах, рекомендованных Комитетом по контролю в сфере образования и науки Министерства образования и науки Республики Казахстан. Четыре статьи опубликованы в материалах международных конференций. Получено одно свидетельство о внесении сведений в Государственный реестр прав на объекты, охраняемые авторским правом Республики Казахстан.

В журналах, рекомендованных Комитетом по обеспечению качества в сфере науки и высшего образования Министерства науки и высшего образования Республики Казахстан.

1. Болатбек М.А., Сағынай М., Мусиралиева Ш.Ж. Применение методов машинного обучения для выявления деструктивного веб-контента на казахском языке // Известия НАН РК. Серия физико-математическая. – 2024. – № 4. – С. 99–111.

2. Сағынай М., Болатбек М. Морфологически и семантически усовершенствованная модель классификации деструктивного веб-контента на казахском языке // Вестник КазАТК. – 2026. – Т. 144. – № 3. – С. 151–166.

Научные статьи в журналах с индексацией в Scopus

1. Bolatbek M., Sagynay M., Mussiraliyeva S., Yeltay Z. Detection of offensive content in the Kazakh language using machine learning and deep learning approaches // PeerJ Computer Science. – 2025. – Vol. 11. – e3027. https://doi.org/10.7717/peerj-cs.3027 (индексируется в Web of Science: Q2; индексируется в Scopus; SJR за 2024 год – 0,719; процентиль – 80).

Научные статьи в материалах конференций с индексацией в Scopus

1. Milana Bolatbek, Moldir Sagynay, Shynar Mussiraliyeva, Yeltay Zhastay. Machine learning methods for identifying destructive texts in the Kazakh language // 2024 3rd International Conference on Algorithms, Data Mining, and Information Technology (ADMIT 2024). https://doi.org/10.1145/3701100.3701121

2. Milana Bolatbek, Moldir Sagynay, Shynar Mussiraliyeva, Kymbat Baisylbayeva, Zhastay Yeltay. Kazakh language dataset for hate speech detection on social media text // 2024 9th International Conference on Computational Intelligence and Applications (ICCIA 2024).

3. Moldir S., Bolatbek M., Mussiraliyeva S., Baisylbayeva K. Natural Language Processing and Machine Learning for Detecting Harmful Speech in Kazakh Social Media // 2025 11th International Conference on Computing and Artificial Intelligence (ICCAI). – Март 2025. – IEEE. – С. 347–351.

Статьи в материалах международных конференций

1. Сағынай М., Болатбек М. Разработка моделей глубокого обучения для классификации деструктивного веб-контента на казахском языке // Материалы международной конференции студентов и молодых учёных «Фараби әлемі».

2. Болатбек М., Сағынай М., Мусиралиева Ш. Проблема выявления вредоносного веб-контента на веб-ресурсах и её актуальность // Материалы IX Международной научно-практической конференции «Информатика и прикладная математика», посвящённой 90-летию Казахского национального университета имени аль-Фараби.

3. Сағынай М., Болатбек М. Создание моделей машинного обучения для выявления и классификации опасного веб-контента на казахском языке // «Фараби әлемі». – 2025. – С. 265.

4. Bolatbek M., Mussiraliyeva S., Sagynay M. Machine Learning-Based Hate Speech Detection in the Kazakh Language // Research in Language. – 2025. – Vol. 23. – P. 311–319.

Свидетельства о внесении сведений в Государственный реестр прав на охраняемые авторским правом объекты

1. Болатбек М.А., Сағынай М. «Повышение эффективности моделей глубокого обучения на основе BERT для классификации деструктивного веб-контента на казахском языке с использованием метода Supervised Contrastive Learning (SupCon)». Свидетельство № 69171 от 27 марта 2026 года.

Объём и структура работы. Диссертация состоит из введения, четырёх разделов, заключения, списка использованных источников и приложения. Общий объём диссертации составляет 98 страниц машинописного текста, включая 26 рисунков, 21 таблицу, список из 107 использованных источников и одно приложение.

Во введении обоснована актуальность темы диссертационного исследования, определены цель и задачи работы, объект и предмет исследования. Описаны методы исследования, научная новизна работы и практическая значимость её результатов. Сформулированы основные положения, выносимые на защиту, приведены сведения об

апробации результатов исследования и публикациях. Также указаны объём и структура диссертации.

В первом разделе диссертации рассмотрены теоретические и методические основы выявления деструктивного веб-контента на казахском языке. Описаны понятие деструктивного веб-контента, его основные виды и влияние на информационную безопасность, проанализированы языковые особенности веб-текстов на казахском языке. Выполнен сравнительный обзор подходов к обработке естественного языка, традиционных методов машинного обучения и методов глубокого обучения, применяемых для автоматического выявления деструктивных текстов. Определены научные проблемы, связанные с агглютинативной структурой казахского языка, контекстной зависимостью деструктивного значения и недостаточностью аннотированных данных. На основе проведённого анализа сформулирована исследовательская задача и обоснована необходимость объединения контекстных, морфологических и семантических признаков на основе онтологии.

Во втором разделе диссертации разработаны основы метода классификации деструктивного веб-контента на казахском языке. Описаны этапы сбора, фильтрации, предварительной обработки и экспертного аннотирования текстов с платформ Telegram, YouTube и VKontakte. В результате сформирован корпус из 12 059 текстов, распределённых по классам «Нейтральные тексты», «Буллинг», «Нацизм» и «Насильственный экстремизм», и проанализированы его статистические и языковые характеристики. Предложен способ формирования морфологических признаков, включающих категории части речи, времени, лица и наклонения. Разработана онтологическая модель знаний, описывающая деструктивный веб-контент по его типу, явности выражения и целевой направленности, и определён порядок извлечения семантических признаков на основе лексикона и графовой структуры. Описана методическая основа объединения этих признаков с контекстными представлениями, формируемыми KazBERT.

В третьем разделе диссертации разработаны модели глубокого обучения и гибридная архитектура для классификации деструктивного веб-контента на казахском языке. KazBERT выбран в качестве базовой контекстной модели; описаны способы его дополнения модулем морфологических признаков и онтологическим графовым модулем. Предложена математическая модель объединения контекстных, морфологических и семантических представлений на основе онтологии. Рассмотрены вариант классификации текстов с использованием представления [CLS] и альтернативные конфигурации, обрабатывающие последовательности токенов с помощью LSTM, BiGRU или BiLSTM. Описаны структура гибридной архитектуры, включающей KazBERT, MFA, онтологические семантические признаки и BiLSTM, порядок обработки признаков и способ вычисления вероятностей классов.

В четвёртом разделе диссертации эффективность предложенных моделей оценена посредством экспериментальных исследований. Описаны порядок разделения данных на обучающую, валидационную и тестовую выборки, экспериментальная среда и метрики оценки. Модели запускались с пятью различными значениями начального состояния генератора случайных чисел в соответствии с единым протоколом оценки; рассчитаны средние значения результатов и стандартные отклонения. Определён вклад морфологического, онтологического и последовательностного компонентов в качество классификации, а ошибочные предсказания проанализированы с использованием матриц ошибок. Различие между результатами предложенной модели и базовой модели KazBERT оценено с помощью теста Мак-Немара. Среди рассмотренных моделей гибридная архитектура на основе BiLSTM показала наилучшие результаты: Accuracy составила 91,045%, а Macro-F1 – 91,103%. Возможность практического применения модели продемонстрирована с помощью прототипа веб-интерфейса.

В заключении представлены основные результаты, полученные в диссертационной работе.