Инженерный вестник Дона, №3 (2026)
ivdon.ru/ru/magazine/archive/n3y2026/10850

Разработка и верификация системы защиты информации от деструктивного контента на основе трансформерных моделей

М.С. Анурьева[1,2], А.А. Скворцов[1,2], Д.В. Лопатин[1], Н.Л. Королева[1]
[1] Тамбовский государственный университет имени Г.Р. Державина
[2] Московский государственный университет технологий и управления имени К.Г. Разумовского (Первый казачий университет)

Аннотация: Рассматривается задача автоматизированного выявления деструктивных речевых воздействий в пользовательском контенте цифровых платформ как элемент системы обеспечения информационной безопасности. Предложен метод контекстно-семантической идентификации агрессивных и дискриминационных высказываний на основе модели русскоязычной версии двунаправленного трансформера для представления текста, адаптированной на специализированном размеченном корпусе русскоязычных сообщений. Описаны процедура формирования данных, схема обучения бинарного классификатора и вероятностная интерпретация результатов. Экспериментальная оценка подтверждает эффективность и устойчивость метода к лексической вариативности и контекстно-зависимым формам речевой агрессии, а также возможность его интеграции в автоматизированные системы мониторинга и защиты информационного пространства.
Ключевые слова: информационная безопасность, деструктивный контент, речевая агрессия, автоматическая модерация, контекстно-семантический анализ, трансформерная модель, бинарная классификация, машинное обучение, обработка естественного языка, система мониторинга, интеллектуальная фильтрация.

Введение

Современная цифровая коммуникационная среда является не только пространством социального взаимодействия, но и объектом целенаправленных информационных воздействий, оказывающих дестабилизирующее влияние на личность, социальные группы и устойчивость информационных процессов. Существенную долю таких воздействий составляют деструктивные речевые формы: оскорбления, угрозы, кибербуллинг, высказывания на основе языка ненависти и дискриминации, характеризующиеся нарушением нормативных моделей коммуникации и направленностью на конкретных адресатов [1]. В условиях массовости и анонимности сетевого общения данные формы приобретают системный характер и рассматриваются как лингвистически опосредованные

информационные угрозы, требующие формализованных и технологически воспроизводимых средств противодействия.

Актуальность автоматизированного выявления деструктивного контента определяется, в том числе, нормативно-правовыми факторами. Федеральные законы от 27.07.2006 № 149-ФЗ и от 01.07.2021 № 236-ФЗ возлагают на операторов цифровых платформ обязанность по выявлению и ограничению распространения противоправной и социально опасной информации, что переводит задачи модерации в контур обеспечения информационной безопасности. В данном контексте системы интеллектуальной фильтрации и классификации пользовательских сообщений следует рассматривать как элементы средств защиты информации, функционирующие в режиме, близком к реальному времени.

Исторически первичным механизмом противодействия являлась ручная экспертиза, обладающая высокой интерпретативной точностью, но не масштабируемая в условиях экспоненциального роста объемов сетевых сообщений и подверженная влиянию человеческого фактора [2]. Это обусловило переход к автоматизированным методам и формированию алгоритмически воспроизводимых процедур выявления речевых атак.

Эволюция исследований в данной области демонстрирует переход от правиловых и словарных фильтров к статистическим и нейросетевым моделям, учитывающим распределенные семантические представления и контекст высказывания [3]. Показано, что методы глубокого обучения превосходят поверхностные алгоритмы по устойчивости к лексической вариативности и завуалированным формам агрессии, что делает их более перспективными для интеллектуальных систем защиты информации [4]. Вместе с тем, ограничения поверхностных и последовательных нейросетевых моделей при обработке контекстно-зависимых и имплицитных форм речевой агрессии обусловили необходимость перехода к более выразительным

архитектурам, способным учитывать глобальный семантический контекст высказывания.

Ключевым этапом стало появление трансформерных архитектур и модели двунаправленного трансформера для представления текста (Bidirectional Encoder Representations from Transformers – BERT), обеспечивающих двунаправленное контекстное кодирование и интерпретацию имплицитных речевых воздействий [5, 6]. Адаптация данной архитектуры к русскому языку в виде модели русскоязычной версии BERT (Russian BERT – RuBERT) и методы ее тонкой настройки показали высокую эффективность в задачах контекстно-зависимой классификации токсичного и агрессивного контента [7, 8]. При этом достоверность таких систем существенно зависит от качества и типологической репрезентативности обучающих корпусов, а также от когнитивных и прагматических факторов аннотирования [9].

В работах авторского коллектива показана эффективность интеллектуальных методов классификации текстов и обоснована целесообразность применения трансформерных архитектур для семантической фильтрации в условиях лингвистической неопределенности [10]. В смежных исследованиях подтверждена применимость методов машинного обучения для автоматизированного мониторинга и поддержки принятия решений в сложных технических системах, в том числе в контурах информационной безопасности [11, 12].

Таким образом, сформирована теоретическая и технологическая база для построения контекстно-чувствительных средств выявления деструктивных речевых воздействий, однако сохраняется методологическая проблема их адаптации и интеграции в состав прикладных систем защиты информации, функционирующих в условиях реальных цифровых коммуникационных платформ и нормативных ограничений. Целью

настоящей работы является разработка и экспериментальное обоснование метода автоматического выявления и классификации враждебных и дискриминационных речевых конструкций на основе современных моделей представления и анализа текстовой информации.

Постановка задачи

Ключевая проблема автоматизированного выявления деструктивного контента заключается в несоизмеримости его лингвистической формы и прагматического содержания. Тексты, содержащие речевую агрессию, дискриминацию или манипуляцию, нередко не имеют явных маркеров, а их деструктивная интенция выявляется на основе анализа семантики, контекста и социокультурных импликаций. Это делает недостаточными подходы, опирающиеся на ключевые слова и поверхностные статистические закономерности, и требует моделей, способных к контекстно-ориентированному смысловому анализу.

В данном контексте трансформерные архитектуры, в частности RuBERT, являются адекватным инструментарием, поскольку формируют контекстуализированные векторные представления текста и учитывают глобальные зависимости между его элементами. Вместе с тем их применение для фильтрации деструктивного контента в русскоязычном цифровом пространстве требует решения ряда методических задач.

Основная сложность связана с адаптацией предобученной модели к узкому, но семантически сложному домену: необходимы формирование специализированного корпуса, разработка схемы тонкой настройки, ориентированной на выявление прагматически вредоносных высказываний, и верификация результата не только по стандартным метрикам, но и по устойчивости к попыткам обхода и по соответствию требованиям пропускной способности систем автоматизированного мониторинга.

Материалы и методика исследования

В качестве эмпирической базы исследования сформирован специализированный корпус русскоязычных текстовых сообщений, предназначенный для обучения и валидации методов автоматизированного выявления деструктивных речевых воздействий в пользовательском контенте. Совокупный объем корпуса после процедур очистки, нормализации и устранения дубликатов составил 120 000 текстовых единиц.

Корпус сформирован на основе интеграции данных из двух взаимодополняющих источников: открытых аннотированных коллекций русскоязычных комментариев (в том числе корпусов *«Toxic Russian Comments»* и *«Russian Language Toxic Comments»*), приведенных к единому формату и согласованной схеме классификации, а также пользовательских комментариев социальной сети «ВКонтакте», полученных посредством автоматизированного сбора через программный интерфейс платформы (метод *wall.getComments*).

Каждое сообщение представлено в виде структурированной записи, содержащей текст и числовую метку класса. На этапе подготовки данных выполнялась лингвистическая нормализация, включавшая удаление эмодзи, гиперссылок, пунктуационных символов и избыточных пробелов с использованием регулярных выражений и специализированных библиотек обработки текста. Классовая разметка носит бинарный характер и отражает принадлежность сообщения к нейтральному контенту либо к контенту с признаками деструктивного речевого воздействия; корректность и согласованность меток контролировались в ходе валидации корпуса.

В основе предлагаемого подхода лежит *контекстно-семантическая модель автоматизированной идентификации деструктивных речевых воздействий*, ориентированная на выявление в пользовательских сообщениях признаков агрессии, оскорблений, угроз и иных форм речевой

дестабилизации, не сводимых к набору поверхностных лексических индикаторов. Метод опирается на представление текста в виде распределенных контекстно-зависимых векторных описаний и последующую бинарную классификацию с использованием глубинной трансформерной модели.

В качестве базового аппроксиматора используется предобученная языковая модель RuBERT, реализующая архитектуру двунаправленного трансформера и способная формировать контекстуальные представления, учитывающие как локальные синтаксические зависимости, так и дальние семантические связи, что обеспечивает корректную интерпретацию имплицитных и завуалированных форм речевой агрессии.

Адаптация модели к прикладной задаче осуществлялась в режиме тонкой настройки на специализированном размеченном корпусе. Трансформерная сеть дополнялась классификационной надстройкой с одним выходным нейроном, формирующим логит, интерпретируемый как апостериорная оценка принадлежности текста к классу деструктивных высказываний. Обучение формализовано как минимизация взвешенной бинарной кросс-энтропии с учетом дисбаланса классов и оптимизацией методом *AdamW* с регуляризацией весов.

Программный комплекс реализован в виде набора взаимосвязанных модулей, объединенных в единую архитектуру, обеспечивающую полный цикл автоматизированной контекстно-семантической модерации текстового контента. Взаимодействие системы с внешними субъектами и общая логика обработки представлены на контекстной диаграмме в нотации «контекст–контейнеры–компоненты–код» (C4 model – C4) уровня 1 (Рис. 1).

Рис. 1. – Контекстная диаграмма (C4 уровень 1) системы интеллектуальной модерации русскоязычного текстового контента

Внутренняя декомпозиция на функциональные контейнеры и потоки данных – на контейнерной диаграмме (C4 уровень 2, Рис. 2).

В соответствии с архитектурной моделью выделяются два логических контура: контур подготовки данных и обучения модели и контур эксплуатации, реализующий онлайн-классификацию сообщений. В первом контуре осуществляется сбор комментариев через программный интерфейс приложения (Application Programming Interface – API) социальной сети «ВКонтакте», их очистка и нормализация, формирование размеченного корпуса и тонкая настройка трансформерной модели RuBERT с сохранением весов и конфигурации для последующего использования.

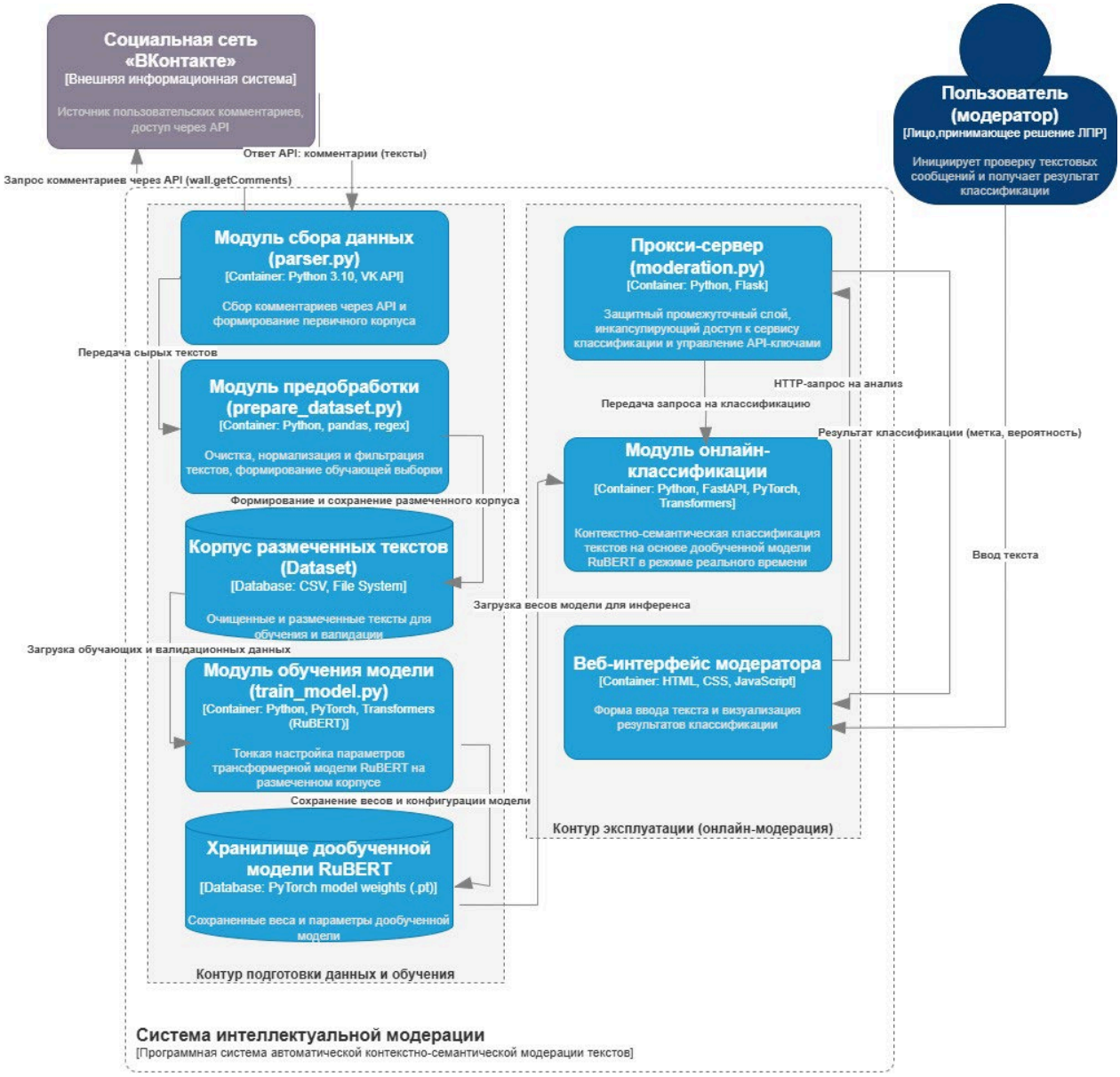

Рис. 2. – Контейнерная диаграмма (C4 уровень 2) архитектуры системы контекстно-семантической идентификации деструктивных сообщений на основе модели RuBERT

Контур эксплуатации обеспечивает обработку текстов в режиме, близком к реальному времени. Ввод сообщений и визуализация результатов выполняются через веб-интерфейс модератора, а взаимодействие с сервисом классификации осуществляется через прокси-сервер, реализующий функции защитного промежуточного слоя и управления доступом. Контекстно-семантическая классификация реализована в виде сервиса, построенного в

соответствии с архитектурным стилем «передача состояния представления» (Representational State Transfer – REST), который загружает дообученную модель, выполняет токенизацию входных данных, инференс и формирует выходные оценки в виде метки класса и апостериорной вероятности.

Представленная архитектура обеспечивает разделение вычислительных и сервисных функций, изоляцию критически важных компонентов и возможность масштабирования, что позволяет рассматривать разработанный программный комплекс как основу для интеграции в контуры автоматизированных систем мониторинга и защиты информационного пространства цифровых платформ.

Экспериментальная оценка и анализ результатов

Оценка метода выполнена на обучающей, валидационной и тестовой выборках (80% / 10% / 10%) с фиксированным параметром воспроизводимости. В ходе обучения контролировалась функция потерь, итоговая оценка проводилась на тестовой выборке по точности классификации и F1-мере. Результаты базового эксперимента представлены в таблице № 1.

Таблица № 1

Результаты обучения и оценки модели RuBERT

Параметр	Валидационная выборка	Обучающая выборка	Тестовая выборка
Функция потерь	0.3515	0.2735	0.3109
Точность классификации	–	–	0.8962
F1-мера	–	–	0.8946
Время выполнения, сек	431.44	24396.91	456.68
Образцов в секунду	55.29	11.73	52.24

Классификационные метрики (точность и F1-мера) вычислялись на тестовой выборке как независимой контрольной подвыборке, предназначенной для оценки обобщающей способности модели. На

обучающей и валидационной выборках в ходе обучения анализировалась динамика функции потерь, используемая для контроля сходимости и отсутствия переобучения.

Сопоставимые значения функции потерь на всех выборках (0.2735, 0.3515, 0.3109) свидетельствуют об отсутствии переобучения. Значения точности классификации (0.8962) и F1-меры (0.8946) на тестовых данных подтверждают высокую обобщающую способность адаптированной модели, достигающую целевых порогов эффективности.

Для определения конкурентных преимуществ предложенного подхода был проведен сравнительный анализ с рядом альтернативных методов. Результаты, сведенные в таблицу 2, демонстрируют сравнительную эффективность различных архитектур.

Таблица № 2

Сравнительные результаты классификации деструктивного контента

Модель / Метод	Точность	F1-мера	Точность (деструктивный класс)	Полнота (деструктивный класс)	Площадь под кривой операционной характеристики приемника
Логистическая регрессия (взвешивание «частота термина–обратная частота документа») (N-граммы 1-3)	0.831	0.705	0.668	0.695	0.754
Сверточная нейронная сеть (Convolutional Neural Network – CNN)	0.848	0.795	0.773	0.784	0.831
Двунаправленная сеть с долгой краткосрочной памятью (Bidirectional Long Short-Term Memory – BiLSTM)	0.859	0.822	0.805	0.815	0.861
RuBERT (предобученная, без дообучения)	0.857	0.758	0.789	0.747	0.848
RuBERT (дообученная модель, предложенный метод)	0.896	0.895	0.912	0.881	0.943

Анализ полученных результатов позволяет сделать несколько принципиальных выводов. Во-первых, сопоставление строк 4 и 5 демонстрирует решающую роль доменной адаптации: тонкая настройка на целевом корпусе существенно повышает качество классификации – прирост F1-меры составляет 0,137, а точности распознавания деструктивного класса – 0,123. Тем самым эмпирически подтверждается, что даже сильная предобученная языковая модель без специализированного дообучения не обеспечивает требуемого уровня эффективности в прикладной задаче модерации.

Во-вторых, предложенный подход (строка 5) устойчиво превосходит классические нейросетевые архитектуры (CNN, BiLSTM) по всем ключевым показателям, что свидетельствует о более адекватном моделировании контекстных зависимостей трансформерной архитектурой. Для задач семантической фильтрации это критично, поскольку деструктивная интенция часто выражается не отдельными словами, а их контекстной конфигурацией и прагматической направленностью высказывания.

Наконец, с точки зрения практической эксплуатации наибольшую значимость имеет достижение максимального значения точности распознавания деструктивного класса (0,912), поскольку именно этот показатель напрямую связан с минимизацией ложных срабатываний в контуре модерации и, следовательно, с снижением избыточных блокировок и нагрузки на ручную проверку. Дополнительно, подтвержденная производительность порядка 52 сообщений в секунду при работе на центральном процессоре указывает на принципиальную пригодность метода для применения в режимах, близких к реальному времени, в составе автоматизированных средств мониторинга.

Для содержательной интерпретации количественных метрик и оценки семантической адекватности работы модели был проведен экспертный

анализ ее решений на репрезентативной выборке высказываний. Целью анализа была проверка способности модели корректно идентифицировать не только явные, но и завуалированные формы деструктивного контента. Результаты, представленные в таблице 3, подтверждают высокую контекстную чувствительность дообученной модели.

Таблица № 3

Примеры классификации текстов системой

Комментарий	Предсказанный статус
только приехали в Тамбов, ваша помощь была бы очень кстати...	Безопасен (0)
Ты полный идиот, если в это веришь	Токсичен (1)
ох, ну и уш1лепок	Токсичен (1)
Получаем 5000 рублей за подписку https://t.me/...	Токсичен (1)
Я хочу устроить терр**	Токсичен (1)

Как видно из таблицы, модель успешно распознает разнородные проявления деструктивности: от прямых оскорблений до сообщений с намеренными орфографическими искажениями, спама и завуалированных угроз. Корректная классификация нейтрального высказывания (первая строка) свидетельствует о низкой склонности модели к гипердиагностике.

Для систематизации ограничений метода был выполнен анализ ошибок классификации на тестовой выборке. Основные типы выявленных проблем категоризированы в Таблице 4.

Представленная категоризация ошибок показывает, что основные ошибки метода связаны с прагматикой и имплицитными смыслами. Ложноположительные срабатывания возникают при конструктивной критике и сарказме из-за негативной или инвертированной лексики, тогда как ложноотрицательные ошибки характерны для скрытых угроз, намеков и культурно-специфичных аллюзий, слабо представленных в обучающем корпусе.

Таблица № 4

Категоризация основных типов ошибок классификации

Категория ошибки	Пример	Причина
Ложноположительные (конструктивная критика или резкий тон)	*«Ваш подход неэффективен и ведет к потерям»* (нейтральный).	Лексика с негативной коннотацией («неэффективен», «потери») интерпретируется как агрессия без учета прагматики профессиональной дискуссии.
Ложноотрицательные (имплицитные угрозы или намеки)	*«Мы с тобой еще поговорим наедине»* (токсичный).	Отсутствие явных маркеров угрозы; для классификации требуется учет внелингвистического контекста и интенции.
Ложноположительные (сарказм и ирония)	*«Ну ты, конечно, большой специалист!»* (нейтральный).	Модель реагирует на поверхностно-позитивную лексику, но не распознает инвертированный ироничный смысл.
Ложноотрицательные (культурно-специфичные инсинуации)	Использование узкосубкультурных аллюзий или мемов с агрессивным подтекстом.	Отсутствие соответствующих паттернов в обучающем корпусе.

В целом результаты подтверждают высокую эффективность дообученной RuBERT-модели и ее преимущество над альтернативными подходами, а также пригодность для применения в режиме, близком к реальному времени.

Заключение

В работе разработан и экспериментально верифицирован метод контекстно-семантической идентификации деструктивного речевого контента на основе трансформерной модели RuBERT, ориентированный на применение в системах обеспечения информационной безопасности цифровых платформ. Показано, что использование контекстно-зависимых представлений позволяет повысить устойчивость классификации к лексической вариативности и завуалированным формам речевой агрессии. Формализована задача автоматической модерации как задача вероятностной бинарной классификации и реализована программная архитектура, обеспечивающая интеграцию модели в контуры автоматизированного

мониторинга через защищенный интерфейс. Полученные результаты подтверждают эффективность предложенного подхода и его применимость в качестве компонента интеллектуальных средств обнаружения лингвистических информационных угроз.

Литература

1. Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N., Kumar R. Predicting the Type and Target of Offensive Posts in Social Media. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. pp. 1415–1425.

2. Roberts S.T. Behind the Screen: Content Moderation in the Shadows of Social Media. New Haven: Yale University Press. 2019. 352 p.

3. Fortuna P., Nunes S. A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys. 2018. T. 51. № 4. p. 85.

4. Toktarova A., Syrlybay D., Myrzakhmetova B., Anuarbekova G., Rakhimbayeva G., Zhylanbaeva B., Suieuova N., Kerimbekov M. Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods. International Journal of Advanced Computer Science and Applications. 2023. T. 14. № 5. pp. 396–403.

5. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is All You Need. Advances in Neural Information Processing Systems. 2017. T. 30. pp. 5998–6008.

6. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. pp. 4171–4186.

7. Kuratov Y., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. Computational Linguistics and

Intellectual Technologies: Papers from the Annual International Conference “Dialogue”. 2019. pp. 333–347.

8. Sun C., Qiu X., Xu Y., Huang X. How to Fine-Tune BERT for Text Classification? Chinese Computational Linguistics. 2019. pp. 194–206.

9. Vidgen B., Derczynski L. Directions in Abusive Language Training Data: A Systematic Review. Computational Linguistics. 2021. Т. 47. № 4. pp. 787–821.

10. Скворцов А.А., Анурьева М.С., Солодовников А.Н. Интеллектуальная система классификации текстов в условиях лингвистической неопределенности. Программная инженерия. 2025. Т. 16. № 11. С. 583–593.

11. Скворцов А.А., Анурьева М.С., Солодовников А.Н. Применение алгоритмов машинного обучения для прогнозирования отказов и адаптивного управления производственными системами. Инженерный вестник Дона. 2025. № 5. URL: ivdon.ru/ru/magazine/archive/n5y2025/10059.

12. Скворцов А.А., Анурьева М.С., Солодовников А.Н. Интеллектуальная система поддержки принятия решений для автоматизированного мониторинга пожаров в технических системах. Моделирование систем и процессов. 2025. Т. 18. № 1. С. 85–96.

References

1. Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N., Kumar R. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Pp. 1415–1425.

2. Roberts S.T. Behind the Screen: Content Moderation in the Shadows of Social Media. New Haven: Yale University Press, 2019. 352 p.

3. Fortuna P., Nunes S. ACM Computing Surveys. 2018. V. 51. No. 4. P. 85.

4. Toktarova A., Syrlybay D., Myrzakhmetova B., Anuarbekova G., Rakhimbayeva G., Zhylanbaeva B., Suieuova N., Kerimbekov M. International Journal of Advanced Computer Science and Applications. 2023. V. 14. No. 5. Pp. 396–403.

5. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Advances in Neural Information Processing Systems. 2017. V. 30. Pp. 5998–6008.

6. Devlin J., Chang M.-W., Lee K., Toutanova K. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Pp. 4171–4186.

7. Kuratov Y., Arkhipov M. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue”. 2019. Pp. 333–347.

8. Sun C., Qiu X., Xu Y., Huang X. Chinese Computational Linguistics. 2019. Pp. 194–206.

9. Vidgen B., Derczynski L. Computational Linguistics. 2021. V. 47. No. 4. Pp. 787–821.

10. Skvortsov A.A., Anur'eva M.S., Solodovnikov A.N. Programmnaya inzheneriya. 2025. V. 16. No. 11. Pp. 583–593.

11. Skvortsov A.A., Anur'eva M.S., Solodovnikov A.N. Inzhenernyj vestnik Dona. 2025. No. 5. URL: ivdon.ru/ru/magazine/archive/n5y2025/10059

12. Skvortsov A.A., Anur'eva M.S., Solodovnikov A.N. Modelirovanie sistem i protsessov. 2025. V. 18. № 1. Pp. 85–96.

Авторы согласны на обработку и хранение персональных данных.

Дата поступления: 19.01.2026

Дата публикации: 3.03.2026