DOI 10.37882/2223-2966.2025.04-2.30

К ВОПРОСУ ВЫЯВЛЕНИЯ ДЕСТРУКТИВНОГО ПОВЕДЕНИЯ НА ОСНОВЕ АНАЛИЗА ТЕКСТОВЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

ON THE ISSUE OF IDENTIFYING DESTRUCTIVE BEHAVIOR BASED ON TEXT DATA ANALYSIS USING MACHINE LEARNING METHODS

N. Teterin
V. Smolentseva

Summary. The problem of identifying destructive behavior of Internet users based on the analysis of text data considered in the paper is an effective tool for preventing the consequences of destructive behavior. Destructive behavior, including cyberbullying, trolling, and hate speech, is a serious problem for online communities. The paper proposes an approach based on machine learning and natural language processing (NLP) methods based on the example of identifying destructive behavior. Experiments have been conducted using various classification algorithms, such as Random Forest, as well as text preprocessing methods. The results show that the proposed approach makes it possible to effectively identify destructive behavior.

Keywords: Destructive behavior, machine learning, forecasting, natural language processing, Internet users.

Тетерин Николай Николаевич
Ассистент, ФГБОУ ВО МИРЭА — Российский технологический университет (РТУ МИРЭА)
teterin@mirea.ru
Смоленцева Владислава Владимировна
ФГБОУ ИВО «Российский государственный университет социальных технологий (РГУ СОЦТЕХ)
smoltan@bk.ru

Аннотация. Рассмотренная в работе задача выявления деструктивного поведения интернет-пользователей на основе анализа текстовых данных является эффективным инструментом для предотвращения последствий от негативного влияния. Деструктивное поведение, включая кибербуллинг, троллинг и распространение ненависти, представляет собой серьезную проблему для онлайн-сообществ. В работе предложен подход, основанный на методах машинного обучения и обработки естественного языка (NLP) на примере выявления деструктивного поведения. Проведены эксперименты с использованием различных алгоритмов классификации, таких как Random Forest, а также методов предобработки текста. Результаты показывают, что предложенный подход позволяет эффективно выявлять деструктивное поведение.

Ключевые слова: деструктивное поведение, машинное обучение, прогнозирование, обработка естественного языка, интернет-пользователи.

С развитием интернет-сообществ и социальных сетей проблема деструктивного поведения пользователей становится все более актуальной. Кибербуллинг, троллинг, распространение ненависти и другие формы агрессивного поведения негативно влияют на психологическое состояние пользователей и могут приводить к серьезным социальным последствиям. Автоматизация выявления таких случаев является важной задачей для обеспечения безопасности в онлайн-среде. В данной статье предлагается подход к прогнозированию выявления деструктивного поведения на основе анализа текстовых данных с использованием методов машинного обучения [1]. В данной области проводятся исследования Минаевым А.В. и Симоновым А.В., которые в своих работах акцентируют внимание на вопросах информационной безопасности [2–3].

Для преобразования текста в числовой формат использовался метод TF-IDF (Term Frequency-Inverse Document Frequency). Данный метод учитывает частоту слов и степень важности [4].

В качестве модели классификации использовался Random Forest. Данный алгоритм был выбран благодаря его способности работать с нелинейными зависимостями и высокой интерпретируемости результатов.

Для обучения модели использовался синтетический набор данных, пример представлен в таблице 1, содержащий 15 текстовых сообщений, разделенных на четыре категории:

- Сообщения, содержащие запросы информации (вопросы).
- Сообщения, выражающие недовольство (жалобы).
- Сообщения с благодарностями или положительными оценками (положительные сообщения).
- Сообщения, содержащие агрессию или крайний негатив (деструктивное сообщения).

Выбор именно такого числа категорий описан в работе соавторов раннее [5].

Матрица ошибок, представленная на рисунке 1, является важным инструментом для оценки качества работы модели классификации. Она наглядно демонстрирует, как модель справляется с разделением данных на классы, и позволяет понять, какие ошибки она допускает.

Таблица 1.

Набор данных

Текст	Категория
Почему в библиотеке так мало учебников?	вопрос
Преподаватель по математике просто замечательный!	положительный
Я не могу понять, как подать апелляцию на оценку.	вопрос
Этот курс ужасно организован, ничего не понятно!	жалоба
Спасибо за помощь в оформлении документов!	положительный
Почему в столовой всегда такие длинные очереди?	жалоба
Как записаться на дополнительные курсы?	вопрос
Мне не нравится, как проводят лекции, это скучно.	жалоба
Отличная библиотека, всегда нахожу нужные книги!	положительный
Зачем вообще нужны эти обязательные предметы?	деструктивное
Я ненавижу этот университет, всё здесь ужасно!	деструктивное
Какой прекрасный кампус, здесь так уютно!	положительный
Почему никто не отвечает на мои письма?	жалоба
Мне нужна помощь с выбором курсов, куда обратиться?	вопрос
Этот университет — пустая трата времени и денег!	деструктивное

В данном случае матрица ошибок показывает, что модель корректно классифицировала все сообщения в тестовой выборке.

Анализ важности признаков, представленный на рисунке 2, позволяет определить, какие именно слова или характеристики текста оказывают наибольшее влияние на процесс классификации, что важно для понимания того, как модель машинного обучения принимает решения и какие данные она считает наиболее информативными. В данном случае выявлено, что определенные слова играют ключевую роль в классификации, что может указывать на их значимость для выявления деструктивных сообщений. Такой анализ помогает:

Улучшить интерпретируемость модели для понимания того, какие слова важны, что делает модель более прозрачной и объяснимой.

Оптимизировать предобработку данных для выявления значимых слов, что позволяет сосредоточиться на наиболее релевантных признаках и ускорить процесс обработки текста.

Повысить точность модели для исключения малозначимых признаков или добавление новых, что может улучшить качество классификации.

Выявить закономерности, для анализа важности слов, помогая понять, какие темы или формулировки

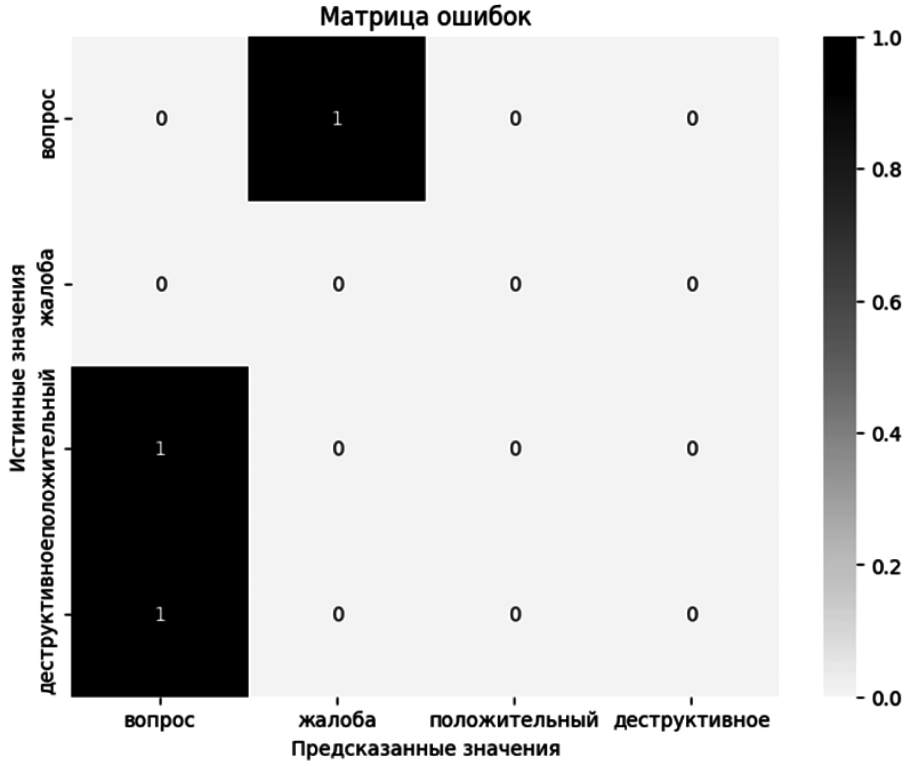

Рис. 1. Матрица ошибок

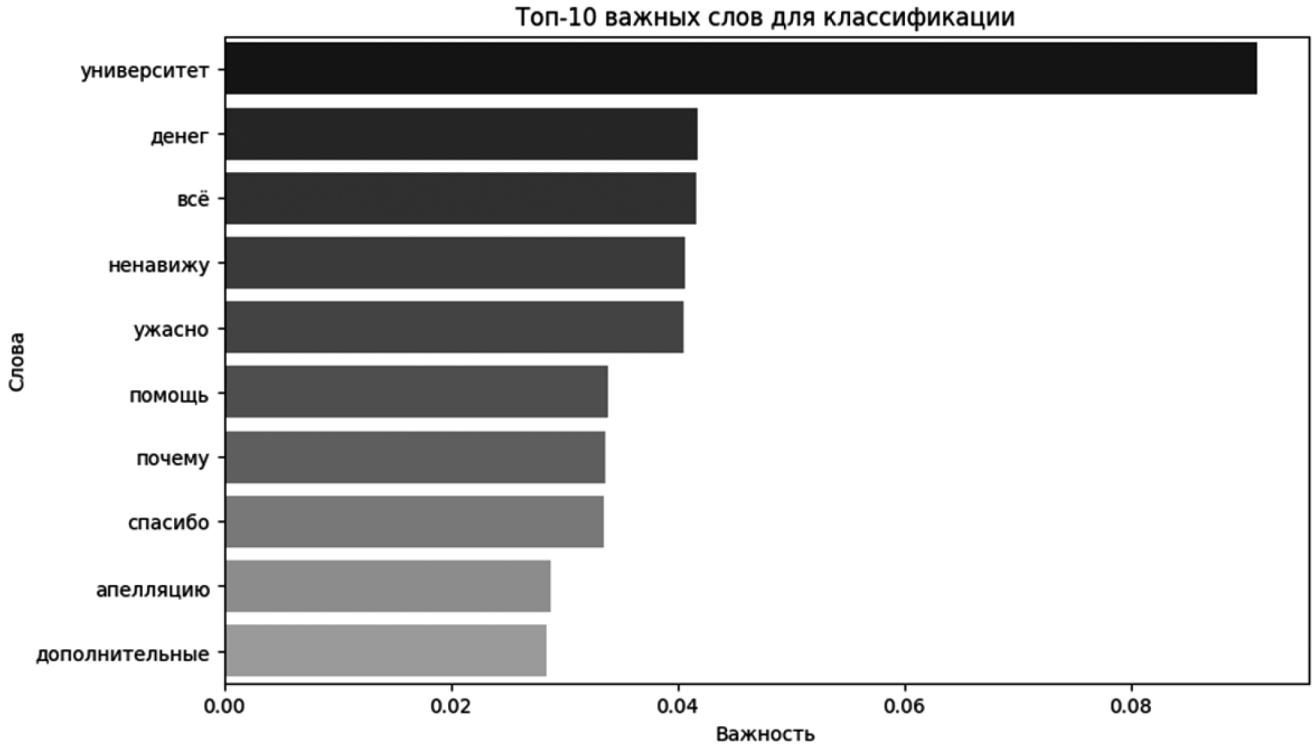

Рис. 2. Наиболее значимые слова для классификации

чаще всего связаны с деструктивными сообщениями, что может быть полезно для дальнейших исследований или разработки профилактических мер.

Использование систем автоматического выявления деструктивного поведения в текстовых сообщениях сопряжено с рядом этических и правовых вопросов, связанных с конфиденциальностью пользователей [6]. Алгоритмы, применяемые для анализа текста, должны быть прозрачными и интерпретируемыми, чтобы пользователи могли понимать, на основании каких критериев принимаются решения, что особенно важно для минимизации риска ложных срабатываний, которые могут привести к необоснованным блокировкам или ограничениям пользователей [7]. Ложные срабатывания не только нарушают права пользователей, но и снижают доверие к системе, для решения этой проблемы необходимо внедрять механизмы контроля осуществляющего ЛПР (лицо принимающее решение), а также обеспечивать возможность анализа решений, предложенных алгоритмом.

Одной из ключевых проблем в разработке моделей для выявления деструктивного поведения является недостаток размеченных данных. Качественная разметка данных требует значительных временных и ресурсных затрат, так как предполагает привлечение экспертов для аннотирования текстов. Кроме того, деструктивное поведение часто зависит от контекста, что усложняет задачу классификации. Например, одно и то же сообщение может быть воспринято как оскорбительное в одном контексте и как нейтральное или даже позитивное — в другом, что требует разработки более сложных методов анализа, учитывающих контекстную зависимость сообщений [8]. Еще одной проблемой является культурная и языковая специфика. Деструктивное поведение может по-разному проявляться в различных языковых и культурных средах. Например, выражения, которые считаются оскорбительными в одной культуре, могут быть приемлемыми в другой, что требует создания мультиязычных и мультикультурных наборов данных, а также адаптации моделей к локальным особенностям.

Для повышения качества классификации деструктивного поведения предлагаются следующие подходы:

Использование предобученных языковых моделей: Современные языковые модели, такие как BERT, GPT и их аналоги, демонстрируют высокую эффективность в задачах обработки естественного языка (NLP). Данные модели способны учитывать контекст сообщений и извлекать сложные семантические зависимости, что делает их особенно полезными для классификации деструктивного поведения. Например, BERT может анализировать контекстные связи между словами, что позволяет более точно определять намерения автора сообщения [9-10].

Для улучшения классификации контекстно-зависимых сообщений предлагается использовать методы, учитывающие не только отдельные слова или фразы,

но и их окружение, что может включать анализ последовательностей сообщений, учет тональности текста, а также использование дополнительных метаданных [11].

Для обучения моделей необходимо создавать аннотированные корпуса текстов, которые охватывают различные языки, культуры и типы деструктивного поведения, что может быть достигнуто за счет краудсорсинга, сотрудничества с социальными платформами и использования полуавтоматических методов разметки данных. Кроме того, важно учитывать баланс классов в наборах данных, чтобы избежать смещения модели в сторону одного из классов.

Язык и поведение пользователей постоянно эволюционируют, модели должны регулярно обновляться и переобучаться на новых данных, что позволит поддерживать их актуальность и эффективность в долгосрочной перспективе.

Заключение

Использование методов машинного обучения, таких как Random Forest, в сочетании с комплексной предобработкой текста, включающей очистку данных, устранение дубликатов и удаление повторов, демонстрирует высокую эффективность в выявлении деструктивных сообщений. Тем не менее, для достижения более высокой точности и надежности модели требуется дальнейшее исследование, включающее внедрение более сложных алгоритмов, таких как нейронные сети, а также расширение и улучшение качества обучающих данных.

Разработка систем автоматического выявления деструктивного поведения требует комплексного подхода, который учитывает как технические, так и этические аспекты. Использование современных языковых моделей, таких как BERT и GPT, в сочетании с методами анализа контекста и созданием качественных наборов данных, позволяет значительно повысить точность и надежность таких систем. Однако важно помнить о необходимости соблюдения прозрачности, минимизации ложных срабатываний и учета культурных особенностей, чтобы обеспечить точную интерпретацию и защиту прав пользователей.

ЛИТЕРАТУРА

1. Семериков, А.В. Классификация объектов на основе нейронной сети и методами дерева решения и ближайших соседей: учебное пособие / А.В. Семериков, М.А. Глазырин. — Ухта: УГТУ, 2022. — С. 68.
2. Минаев, В.А. Выявление деструктивного контента в социальных МЕДИА на основе Моделей машинного обучения / В.А. Минаев, А.Д. Реброва, А.В. Симонов // Информация и безопасность. — 2021. — Т. 24, № 1. — С. 7–20.
3. Минаев, В.А. Повышение точности идентификации контента экстремистского характера в социальных медиа / В.А. Минаев, А.В. Симонов // Информационная безопасность: вчера, сегодня, завтра: Сборник статей по материалам V Международной научно-практической конференции, Москва, 14 апреля 2022 года. — Москва: Российский государственный гуманитарный университет, 2022. — С. 80–86.
4. Гехт, А.Б. Гуманитарные проблемы искусственного интеллекта и его применения: монография / А.Б. Гехт, Р.В. Душкин, А.В. Неровный [и др.]. — Санкт-Петербург: СПбГУТ им. М.А. Бонч-Бруевича, 2024. С. 204–212.
5. Тетерин, Н.Н. К вопросу формализации задачи выявления деструктивного поведения с применением технологий искусственного интеллекта / Н.Н. Тетерин, В.В. Смоленцева // Тенденции развития науки и образования. — 2024. — № 114-10. — С. 81–83.
6. Ибрагимова, Н.И. Опасности в современном мире и методы обучения вопросам безопасности: учебно-методическое пособие / Н.И. Ибрагимова, А.В. Жогаль. — Сургут: СурГУ, 2024. С. 20–25.
7. Романов, В.Г. Социальная инженерия мошенничества: монография / В.Г. Романов, И.В. Романова. — Чита: ЗабГУ, 2021. С. 103–107.
8. Москвитин, А.А. Данные, информация, знания: методология, теория, технологии: монография / А.А. Москвитин. — Санкт-Петербург: Лань, 2022. С. 171–174.
9. Широких, И.В. Специфика текста как продукта искусственного интеллекта / И.В. Широких, А.Г. Фомин // Искусственный интеллект в автоматизированных системах управления и обработки данных : Сборник статей II Всероссийской научной конференции: в 5 томах, Москва, 27–28 апреля 2023 года. — Москва: КДУ, Добросвет, 2023. — С. 349–355.
10. Салып, Б.Ю. Анализ модели BERT как инструмента определения меры смысловой близости предложений естественного языка / Б.Ю. Салып, А.А. Смирнов // StudNet. — 2022. — Т. 5, № 5. — С. 33.
11. Методика сбора и обработки социологической информации из сети Интернет / А.А. Воробьев, А.М. Рыбак, Р.А. Середкин [и др.] // Известия Тульского государственного университета. Технические науки. — 2022. — № 2. — С. 208–213.

Журнал «Современная наука: актуальные проблемы теории и практики»