УДК 004.89:004.056.5

ПРИМЕНЕНИЕ NLP-МОДЕЛЕЙ ДЛЯ ВЫЯВЛЕНИЯ «ТОКСИЧНОГО» ПОВЕДЕНИЯ В ЧАТАХ

Василевский В. В., Пущик Е. А.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Научный руководитель: Рышкель О.С. – к. с.-х н., доцент, доцент кафедры ИПиЭ

Аннотация. Апробирован подход к автоматической модерации текстовых сообщений в онлайн-чатах на основе предварительно обученной нейронной модели архитектуры *BERT*. Разработан программный прототип в виде чат-бота для платформы *Telegram*, продемонстрирована его эффективность на тематической выборке данных и определены метрики качества выявления «токсичного» поведения.

Ключевые слова: *NLP*-модели, автоматическая модерация, «токсичное» поведение, эмодзи-семантика, классификация сообщений

Введение. Онлайн-общение давно перестало быть «простым текстом». В университетских и рабочих *Telegram*-чатах смешаны языки, эмодзи, аббревиатуры, мемы, код, сарказм и *IT*-жаргон. Классические решения (частотные словари, линейные модели, статические правила и регулярные выражения) системно проигрывают из-за проблем:

– смысл сообщения зависит от предыдущих реплик, адресата и ветвления треда;

– эмодзи, реакции и «скобочки» часто заменяют лексику и меняют тон;

– сообщениям в контексте диалога присущи полисемия и сарказм (например: «круто, опять всё сломалось» - отрицательная оценка, а не похвала).

Нейронные модели типа трансформер лучше улавливают семантику, но в финальных условиях возникают ограничения: задержки, стоимость для вывода, интерпретируемость, а также неспособность на лету учитывать эмодзи, молодёжный сленг и вставки кода [1].

Цель работы - оптимизировать процесс определения грубого поведения и показать, что компактная модель, обученная под *Telegram*-среду, способна быстро и объяснимо выявлять токсичное поведение при высокой точности.

Основная часть. В рамках исследования разработан полный алгоритм обработки данных и проведено дообучение модели на специализированных корпусах с учётом: контекста; значков-эмоций, реакций и символических обозначений («скобочек»), выделяемых отдельной подсистемой разметки; профессионального жаргона и включений программного кода, предварительно нормализуемых и приводимых к единообразному виду; работы в нескольких видах данных: в том числе автоматическое распознавание речи для голосовых сообщений при необходимости уточнения смысла.

Полученная система обладает следующими чертами:

– адаптация под среду *Telegram* с учётом локального контекста и значков-эмоций;

– поддержка нескольких видов данных: распознавание речи и выборочное привлечение описаний изображений;

– формирование понятного отчёта с указанием токенов и выражений, наиболее повлиявших на решение;

– высокая скорость работы – менее 50 мс на процессорных системах, при среднем показателе качества более 95% *F*1 на тематических наборах данных.

Для классификации токсичности используется модель семейства *BERT* [2]. Формально задача работы определяется как выбор между двумя классами: нейтральное сообщение и оскорбительное сообщение.

Решение принимается на основе максимальной вероятности, а уверенность модели используется для разделения однозначно токсичных и сомнительных случаев. Для модерации вводится пороговое значение уверенности, позволяющее уменьшить количество ошибок, когда нейтральные фразы ошибочно принимаются за оскорбления.

Создан специализированный корпус: анонимизированные учебные и профессиональные

чаты, а также синтетические примеры диалогов. Сообщения были вручную размечены: «токсичное» / «нетоксичное», а также выделены подтипы – сарказм, пассивная агрессия, полушутливые пикировки, намеренное искажение слов, массовое использование значков-эмоций, реакции и «скобочки».

Значки-эмоции и реакции выделяются в отдельную область представления данных, что позволяет различать их смысл в разных ситуациях. Например, одинаковые значки могут означать насмешку в обзоре или дружескую шутку между знакомыми. Модель получает окно контекста из нескольких сообщений, что позволяет корректно трактовать смысл фраз в диалоге.

Процесс обработки сообщения включает: получение сообщения из *Telegram*, предварительную обработку текста, формирование контекста, при необходимости – распознавание речи или получение текстового описания изображения, классификацию, формирование объяснения. Схема процесса обработки представлена на рисунке 1.

Рисунок 1 – Схема процесса обработки

Предварительная обработка включает нормализацию значков-эмоций, выделение типичных сочетаний, замену фрагментов кода служебными обозначениями, унификацию ссылок и цитат, а также устранение намеренных искажений текста.

Контекст учитывается через окно из нескольких последовательных сообщений. Это позволяет различать шутки и настоящие оскорбления, что существенно уменьшает количество пропусков токсичных сообщений.

Модель выдаёт не только итоговое решение, но и объяснение, в котором указано, какие слова, знаки и элементы контекста повлияли на результат.

Для анализа устойчивости модели были собраны три независимых корпуса:

– сообщения учебных и профессиональных чатов (основной корпус);

– открытые обсуждения в других источниках;

– набор сложных примеров: сарказм, искажённый текст, избыток значков-эмоций.

Оценка проводилась по нескольким метрикам: точность, полнота, *F*1-мера, *ROC-AUC*. Модель показала высокие результаты: более 96% точности на основном корпусе и более 94% – на внешних данных. Точные показатели представлены на матрице ошибок на рисунке 2:

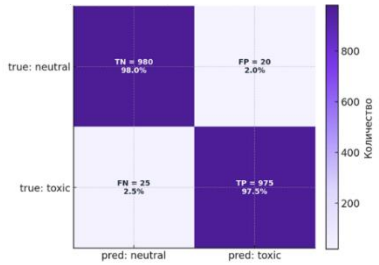

Рисунок 2 – Матрица ошибок

Для каждого сообщения строится тепловая карта вкладов слов, символов и значков-эмоций, показывающая, какие элементы больше всего повлияли на классификацию.

Даже с включёнными механизмами объяснения задержка обработки одного сообщения удерживается в пределах менее 50–80 мс. Все показатели задержки представлены на рисунке 3:

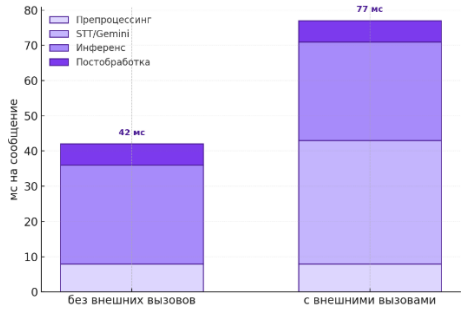

Рисунок 3 – Задержки по стадиям

Заключение. В результате выполнения работы было выявлено, что адаптированная под *Telegram* модель с учётом значков-эмоций, контекста диалога и выборочной поддержки нескольких видов данных обеспечивает высокое качество выявления грубого поведения при низкой задержке обработки.

Модель устойчива к сленгу, искажённому тексту, реакциям и неформальному стилю общения, объясняет свои решения и подходит для использования в учебных и рабочих чатах.

Список литературы

1. *Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Electronic resource]. – Date of access: March 02, 2025. – Available at: https://aclanthology.org/N19-1423/.*

2. *Wolf T. Transformers: State-of-the-Art Natural Language Processing [Electronic resource]. – Date of access: March 02, 2025. – Available at: https://arxiv.org/abs/1910.03771.*

UDC 004.89:004.056.5

APPLICATION OF NLP MODELS FOR DETECTING TOXIC BEHAVIOR IN CHATS

Vasilevsky V.V., Pushchik E.A.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Scientific supervisor: Ryshkel O.S. –Cand. of Sci., associate professor, associate professor of the Department of ICSD

Annotation. The article presents an approach to automatic moderation of text messages in online chats using a pre-trained neural network model based on the BERT architecture. A software prototype in the form of a Telegram chat-bot has been developed, its efficiency has been demonstrated on a domain-specific dataset, and key quality metrics for detecting crude and offensive behavior have been determined.

Keywords: NLP models, automatic moderation, toxic behavior, emoji semantics, message classification