УДК: 004.021;004.048

НЕЙРОСЕТЕВОЙ ПОДХОД ДЛЯ КЛАССИФИКАЦИИ НЕНАВИСТНИЧЕСКОГО КОНТЕНТА В СОЦИАЛЬНЫХ СЕТЯХ

М. А. КОРНЕЕВЕЦ,
специалист Центра перспективных исследований в сфере цифрового развития ОАО «Гипросвязь», магистр соц. н.

Н. Г. ЮНЕВИЧ,
науч. сотр. Центра перспективных исследований в сфере цифрового развития ОАО «Гипросвязь», аспирант кафедры «Информационные технологии в управлении» МИДО, БНТУ

B. AGARWAL,
кандидат философских наук в области компьютерных наук и инженерии (PhD CSE), доцент кафедры компьютерных наук и инженерии Центрального университета Раджастхана, Индия

N. KESSWANI,
доктор философских наук (Post-Doc), доцент факультета компьютерных наук и инженерии Центрального университета Раджастхана, Индия

В статье рассматривается проблема распространения ненавистнического контента в социальных сетях и необходимость его регулярного мониторинга. Авторы предлагают решение этой проблемы с использованием искусственного интеллекта и нейронных сетей, в частности предобученной модели ruBERT. Разработанный алгоритм способен автоматически классифицировать текстовые сообщения по тональности и типу ненависти. Для обучения модели применяется метод автоматической разметки данных, основанный на языковой модели GPT-3.5-turbo. Модель состоит из двух модулей: классификации тональности и класса ненависти, что повышает точность классификации. В статье подробно описаны этапы разработки алгоритма, его архитектура и планируемые дальнейшие исследования.

Ключевые слова: ненавистнический контент, социальные сети, нейронные сети, обработка естественного языка, ruBert, GPT-3.5-turbo, разметка данных, классификация текста, определение тональности текста, k-кратная кросс-валидация.

АКТУАЛЬНОСТЬ

ЮНИСЕФ и рядом исследователей (P. Núñez-Gómez, K.P. Larrañaga, C. Rangel, F. Ortega-Mohedano и др.) указано, что проблема кибербуллинга до сих пор остается актуальной. Несмотря на то, что все наиболее популярные социальные платформы оснащены продвинутыми системами модерации контента с использованием как классических полуавтоматических методов, так и нейросетевого подхода, региональная специфика языка препятствует качественной классификации контента [1]. Это касается и наиболее популярных среди белорусского сообщества социальных сетей (Pinterest, Instagram, Facebook, YouTube и др. [2]). На данный момент существует множество моделей выявления негативного контента, в том числе построенные на искусственном интеллекте, включая машинное обучение, глубокое обучение и гибридное обучение. Однако для качественного выявления негативного контента с учетом семантической и морфологической региональной специфики языка необходима тонкая настройка моделей и их дообучение на соответствующих наборах данных.

Статья посвящена поиску и построению подходов (алгоритмов) к выявлению негативного текстового контента с ориентацией на молодежь как наиболее активную и подверженную кибербуллингу группу. Актуальные исследования демонстрируют, что данная группа интернет-пользователей в большей степени сталкивается с кибербуллингом. Последствия – психологическая подавленность, депрессивные и тревожные состояния, суицидальные мысли и проблемы с социализацией [3, 4, 5, 6]. По данным ЮНИСЕФ, более 30 % белорусских детей и подростков так или иначе сталкивались с угрозами и травлей в интернете, при этом часть из них не сообщали об этом своим родителям [7].

Работа ведется совместно с представителями Центрального университета Раджастхана. Фокусом их исследований является разработка решений для повышения точности моделей, работающих с хинди, что особенно актуально в контексте кибербуллинга, где региональная специфика языка играет важную роль. Объединение усилий в рамках этого проекта позволит разработать эффективные алгоритмы для выявления негативного контента с учетом семантических и морфологических региональных особенностей языка.

ОСНОВНАЯ ЧАСТЬ

На сегодняшний день существует несколько подходов, лежащих в основе распознавания негативного содержания текстовой информации, в число которых входит: морфологический анализ – выделение грамматических характеристик слов и их изменяемых форм; синтаксический анализ, предполагающий анализ структуры предложения и выявление связей между словами в нем; семантический анализ, суть которого заключается в выявлении смысловых связей между словами в тексте [8].

Алгоритмизация и комбинация данных подходов, реализованные в архитектурах нейросетевых моделей NLP (Natural language processing), позволяют автоматизировать процессы распознавания и анализа текстовой информации. Однако стандартные рекуррентные нейронные сети RNN (Recurrent Neural Network) сталкиваются с проблемой затухающего градиента при обработке длинных последовательностей, что ограничивает их эффективность. LSTM (Long Short-Term Memory) представляет собой тип рекуррентной нейронной сети, специально разработанной для решения проблемы затухающего градиента, характерной для стандартных RNN-моделей при обработке длинных последовательностей. В отличие от классических RNN, LSTM обладает механизмом долговременной памяти, который позволяет ей эффективно сохранять информацию и использовать ее для обработки текущих последовательных данных. Ключевым элементом архитектуры LSTM является использование «гейтов» – механизмов, контролирующих поток информации внутри сети. Три основных гейта – входной, выходной и гейт забывания – регулируют добавление новой информации в состояние ячейки, передачу информации в следующие ячейки и удаление ненужной информации [9].

Благодаря этой архитектуре LSTM способны учитывать контекст при обработке последовательностей, что делает их эффективными для решения задач, требующих понимания временных зависимостей, таких как генерация текста, машинный перевод, анализ настроений и распознавание речи. Несмотря на это, модель может испытывать трудности с сохранением информации из далекого прошлого, если она не активируется достаточно часто в процессе обучения. Такое явление обусловлено ограниченной способностью LSTM-модели хранить и извлекать информацию, относящуюся к далеким временным интервалам.

Архитектура Seq2seq (Sequence-to-Sequence) смогла решить данную проблему. Данная архитектура модели способна эффективнее преобразовывать входные текстовые последовательности в новые текстовые последовательности, используя два ключевых компонента: энкодер и декодер. Энкодер последовательно обрабатывает входной текст, сводя его к вектору фиксированной длины, называемому вектором контекста. Этот вектор, содержащий сжатую информацию о входной последовательности, передается декодеру, который использует его для предсказания следующего слова в выходной последовательности.

В отличие от LSTM модель Seq2seq способна генерировать целые последовательности токенов (слов) за один запуск, что позволяет учитывать полный контекст входного текста. Декодер использует словарь слов, сформированный во время обучения модели, для генерации выходного текста. Несмотря на широкое применение в продуктах, использующих естественный язык, модель Seq2seq в своем изначальном виде имеет некоторые недостатки: она склонна генерировать грамматически неверные тексты и отличается низкой скоростью обучения из-за последовательного алгоритма обработки входного текста [9].

Последней и самой эффективной на сегодняшний день архитектурой NLP-моделей является Transformer. В отличие от Seq2seq-модели, Transformer отказывается от рекуррентных нейронных сетей в пользу механизма многоканального внимания (Multi-Head Attention), что позволяет ей устанавливать связи между разными словами в тексте, учитывая разный контекст. Архитектура Transformer, подобно Seq2seq, состоит из энкодера и декодера, но с радикально отличающейся внутренней структурой. Механизм многоканального внимания позволяет модели формировать вектор контекста для каждого отдельного токена в последовательности, анализируя взаимосвязи между словами.

Кроме этого, Transformer использует «поиск по лучу» (Beam Search) для генерации выходных последовательностей. В отличие от изначального подхода, при котором выбирается слово с максимальной вероятностью на каждом шаге, «поиск по лучу» анализирует множество вариантов, храня наиболее вероятные последовательности и продолжая генерацию от каждого из них. Эти особенности архитектуры позволяют осуществлять предварительное обучение моделей на неразмеченных данных, что способствует адаптации весовых коэффициентов к специфическим языковым особенностям конкретного естественного языка. Такой подход позволяет модели экстраполировать общие закономерности и структуру языка, что делает ее более эффективной [10]. Среди наиболее известных предобученных моделей можно выделить: GPT (OpenAI), BERT (Google), ruT5 (Sberbank).

Самостоятельная разработка и обучение модели Transformer для классификации текстовых сообщений на предмет наличия ненавистнических высказываний представляет собой ресурсоемкий процесс, требующий значительных инвестиций в данные, вычислительные мощности и временные ресурсы. Это обстоятельство в первую очередь сопряжено с необходимостью ручного сбора и аннотации больших объемов текстовых данных. Кроме этого, аннотация данных требует привлечения экспертов с глубоким пониманием нюансов языка

и особенностей ненавистнических высказываний. Также обучение модели Transformer требует значительных вычислительных ресурсов с использованием высокопроизводительных графических процессоров (GPU) и вычислительных кластеров, способных обрабатывать большие объемы данных. Обучение может занимать несколько недель или даже месяцев, в зависимости от сложности модели, размера датасета и доступных вычислительных ресурсов.

В качестве альтернативы можно использовать предобученные модели Transformer, которые уже обучены на огромных объемах текстовых данных. Эти модели обладают базовым пониманием языка и могут быть адаптированы для решения конкретных задач, в число которых входит и классификация. Данный подход позволяет значительно сократить затраты на обучение модели, так как не требуется собирать и аннотировать большие объемы данных.

В контексте русского языка существует дообученная нейронная сеть ruBERT, демонстрирующая лучшие результаты в задачах классификации, по сравнению с имеющимися на сегодняшний день аналогами [11, 12, 13]. Модель ruBERT представляет собой модель глубокого обучения, которая основана на архитектуре трансформера и является русской адаптацией модели BERT. Данная модель обучена на обширном корпусе русскоязычных текстов, собранных из разнообразных источников, таких как новостные ленты, литературные произведения, научные публикации и другие, что обеспечивает широкое покрытие тематик и стилей русского языка. Использование этой модели и ее дальнейшее дообучение легли в основу разработанной нами авторской архитектуры алгоритма автоматической классификации текстовых сообщений.

В рамках исследования прежде всего было необходимо разработать методологические основы определения и классификации ненавистнического контента, которые лягут в основу разработки и обучения алгоритма лингвистического анализа текста. Под «ненавистническим контентом» в социальных сетях нами понимается любое сообщение или поведение, которое принижает достоинство других пользователей и выражается в форме дискриминации, оскорблений, угроз, распространения ненависти и в других формах неприемлемого поведения. Следующим необходимым этапом является определение валидной классификации, отражающей специфичный для белорусского веб-пространства характер онлайн-коммуникаций. Выбор качественной классификации злоупотреблений позволяет повысить точность идентификации и анализа неприемлемого поведения, создает целевую стратегию противодействия злоупотреблениям и позволяет разрабатывать более эффективные алгоритмы детектирования. По результатам предварительного анализа имеющегося массива тестовых данных, нами были определены следующие виды ненавистнического контента в интернет-среде: сексизм, гомофобия, лукизм, ксенофобия, а также другие формы дискриминации (например, возрастная, религиозная, социальная и т. д.).

Дальнейшая реализация подхода, основанного на использовании обученной нейронной сети с последующим ее дообучением, требует наличия размеченного набора данных, используемого для обучения модели и оценки ее предсказательной способности. В рамках данной задачи существует несколько подходов к разметке данных: ручная (осуществляется человеком, который вручную присваивает метки каждой единице данных), автоматическая (использует алгоритмы для автоматического присвоения меток данным) и полуавтоматическая разметка (представляет собой комбинацию ручного и автоматического подходов).

В процессе выбора оптимальной стратегии разметки данных, учитывая масштабность задачи и стремление к минимизации ресурсных затрат, было принято решение о приоритетном использовании автоматического подхода. Анализ существующих подходов выявил ряд существенных ограничений ручной разметки, среди которых наиболее значимыми являются: высокая трудоемкость и ресурсоемкость, требующая значительных временных и финансовых затрат, особенно при работе с масштабными наборами данных, а также субъективность интерпретации, приводящая к несоответствиям в разметке и снижению точности обучаемой модели из-за различных интерпретаций данных экспертами. Эти ограничения ручной разметки делают ее менее привлекательной в контексте настоящего исследования, особенно при учете масштаба задачи и ограниченности ресурсов.

Нами был разработан алгоритм автоматизации разметки данных, использующий языковую модель GPT-3.5-turbo. В основе алгоритма лежат две специально сформулированные инструкции, которые используются для взаимодействия с моделью. Первая из них направлена на определение тональности текста, вторая – для классификации текста по классу злоупотреблений. Для увеличения точности разметки используется метод калибровки уверенности с помощью вычисления моды и определения доли совпадений ответов модели при пяти запусках с одним и тем же текстом (рисунок 1).

Разработанная нами модель представляет собой систему, основанную на предобученной языковой модели ruBERT, дополнительно обучающейся на размеченном массиве текстовых данных. На первом этапе размеченные данные передаются на вход модели, после чего полученное векторное представление, содержащее информацию о контексте и смысловых связях слов, передается в два специализированных модуля: модуль классификации тональности и модуль классификации класса ненависти.

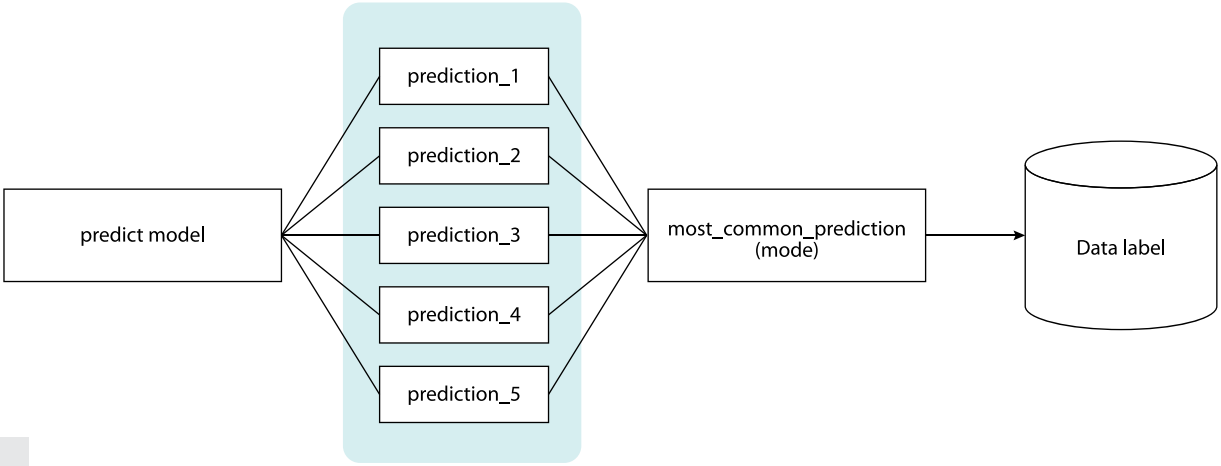

Рисунок 1. Схема алгоритма автоматической разметки данных

Модуль классификации тональности состоит из полносвязного слоя с тремя нейронами, каждый из которых соответствует одному из трех классов тональности: -1 (отрицательная), 0 (нейтральная), 1 (положительная). Особый нюанс этой системы состоит в том, что тексты, классифицированные как нейтральные или положительные по тональности, автоматически получают класс «нейтральный» или «позитивный» соответственно и не передаются в модуль классификации класса ненависти. Таким образом, в модуль классификации класса ненависти попадают только тексты, имеющие негативную тональность (-1).

Модуль классификации класса ненависти использует полносвязный слой с пятью нейронами, соответствующими пяти классам ненависти: «сексизм», «гомофобия», «ксенофобия», «лукизм» и «другое». Оба модуля используют функцию активации Softmax для вычисления вероятностей принадлежности входного текста к каждому из классов (рисунок 2).

Обучение модели происходит методом обратного распространения ошибки на наборе данных, включающем тексты с метками тональности и класса ненависти. Для оптимизации процесса обучения используется функция потери кросс-энтропии, а оптимизатор Adam настраивает веса модели во время обучения, чтобы минимизировать функцию потери и улучшить точность классификации. Для предсказания тональности и класса ненависти нового текста, он проходит через модель ruBERT, полученные векторные представления передаются в классификаторы и выбираются классы с наибольшими вероятностями.

Данная архитектура обладает рядом преимуществ: использование предобученной модели ruBERT значительно улучшает точность классификации, т. к. модель уже обладает знаниями о русском языке; разделение на два классификатора позволяет эффективно решать две независимые задачи. Однако точность модели зависит от качества и количества данных обучения. Кроме того, модель может быть чувствительна к нестандартным словам, которые не встретились в данных, использованных для обучения. Мы полагаем, что данная модель может выступать эффективным инструментом для классификации сообщений по тональности и классу ненависти, а также может быть применена в дальнейшем для разработки систем модерации контента и борьбы с онлайн-ненавистью.

На данный момент разработанная архитектура находится на стадии тестирования и рассматриваются варианты ее оптимизации. В дальнейшем планируется полноценное обучение модели на размеченном наборе данных и оценка качества работы модели. Для оценки качества классификации алгоритма будет использоваться метод k-кратной кросс-валидации. Данный метод позволяет разбить исходный набор данных на k частей, использовать k-1 частей для обучения модели, а оставшуюся часть – для тестирования. Процесс повторяется k раз, каждый раз используя другую часть данных для тестирования. В результате получается k наборов метрик качества, которые затем усредняются для получения окончательной оценки качества работы алгоритма. Для более детального анализа точности классификации будут использоваться стандартные метрики precision (доля правильно классифицированных злоупотреблений среди всех текстов, классифицированных моделью как «злоупотребление»), recall (доля правильно классифицированных злоупотреблений среди всех действительных злоупотреблений в исходном наборе данных) и F1-score (гармоническое среднее precision и recall). F1-score позволяет сбалансировать точность и полноту классификации и дает более полное представление о качестве работы алгоритма. Применение этих метрик позволит получить более глубокое понимание работы модели и оценить ее эффективность с учетом особенностей данных.

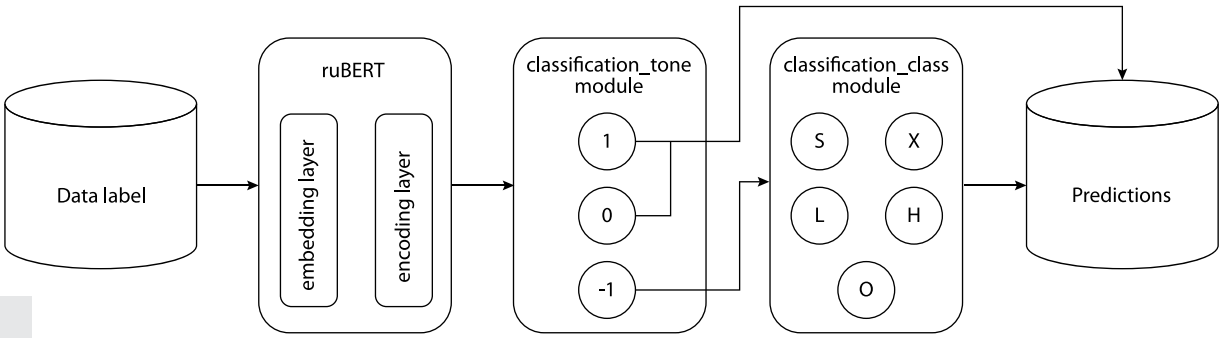

Рисунок 2. Схема архитектуры нейросетевой модели

ЗАКЛЮЧЕНИЕ

В работе была представлена нейросетевая архитектура для автоматической классификации текстовых сообщений в социальных сетях по тональности и типу ненависти. Ключевым элементом модели стала предобученная модель ruBERT, дополнительно обучаемая на размеченном наборе данных. Для автоматической разметки данных использовалась языковая модель GPT-3.5-turbo, что позволило снизить трудоемкость процесса. Модель состоит из двух модулей: классификации тональности и класса ненависти, что повышает точность классификации. В настоящее время модель находится на стадии тестирования, и в дальнейшем планируется ее полноценное обучение и оценка качества работы с использованием метода k-кратной кросс-валидации. Результаты исследования позволят разработать эффективные инструменты для модерации контента и борьбы с онлайн-ненавистью в Республике Беларусь с учетом региональной специфики языка и особенностей культуры интернет-коммуникаций.

ЛИТЕРАТУРА

1. Gongane, V. U., Munot, M. V., Anuse A. D. Detection and moderation of detrimental content on social media platforms: current status and future directions [Электронный ресурс]. – Режим доступа: https://link.springer.com/article/10.1007/s13278-022-00951-3. – Дата доступа: 29.07.2024.
2. Datareportal / Digital 2024: Belarus [Электронный ресурс]. – Режим доступа: https://datareportal.com/reports/digital-2024-belarus. – Дата доступа: 25.07.2024.
3. Gaffney, H., Farrington, D.P. Cyberbullying in the United Kingdom and Ireland // International Perspectives on Cyberbullying: Prevalence, Risk Factors and Interventions / Eds. A.C. Baldry, C. Blaya; D.P. Farrington. Cham: Palgrave Macmillan. – 2018. – pp. 101–143.
4. Ефимов Е.Г. Социальные Интернет-сети (методология и практика исследования) / Е. Г. Ефимов // Волгоград: Волгоградское научное издательство. – 2015. – 168 с.
5. John, A. Self-Harm, Suicidal Behaviours, and Cyberbullying in Children and Young People: Systematic Review / A. John [et al.] // Journal of medical internet research. – 2018. – № 4. – pp. 113–129.
6. Chang, E. The mixed effects of online diversity training / E. Chang [et al.] // PNAS. – 2019. – № 16. – pp. 778–783.
7. #ИнтернетБезБуллинга. Тренируем навыки безопасного поведения в интернете [Электронный ресурс]. – Режим доступа: https://www.mts.by/unicef/. – Дата доступа: 29.07.2024.
8. Исаева, М. З. Методы и технологии обработки естественного языка / М.З. Исаева, А.В. Алисултанова, М-А. Дасаев и др. // «МИЛЛИОНЩИКОВ-2023». – 2023. – с. 90–95.
9. Науменко, В. И., Петров С. А. Обзор методов обработки естественного языка для автоматической генерации тестовых заданий / В. И. Науменко, С. А. Петров // Вестник Московского энергетического института. – 2024. – №3. – с. 112–126.
10. Qiu X. et al. Pre-trained models for natural language processing: A survey // Science China Technological Sciences. – 2020. – Т. 63. – №10. – pp. 1872–1897.
11. Shaheen, Z., Mouromtsev, D.I., Postny, I. Rulegalner: a new dataset for russian legal named entities recognition // Научно-технический вестник информационных технологий, механики и оптики – 2023. – № 4. – pp. 854–857.
12. Vychegzhanin, S., Milov, V., Kotelnikov E. Comparative analysis of machine learning methods for news categorization in Russian [Электронный ресурс]. – Режим доступа: https://ceur-ws.org/Vol-2922/paper012.pdf. – Дата доступа: 25.07.2024.
13. Kotelnikova, A., Paschenko, D., Bochenina K., Kotelnikov, E. Lexicon-Based vs. Bert-Based Sentiment Analysis / A. Kotelnikova, D. Paschenko, K. Bochenina, E. Kotelnikov // Analysis of Images, Social Networks and Texts : 10th International Conference, AIST 2021. – Georgia, 2021 – pp. 71–83.

The article discusses the problem of disseminating hateful content on social networks and the need for its regular use. The authors propose a solution to this problem using artificial intelligence and neural networks, in particular, the pre-trained ruBERT model. The developed algorithm is able to automatically classify text messages by tone and type of hatred. To train the model, an automatic data labeling method is used, based on the GPT-3.5-turbo language model. The model consists of two modules: sentiment classification and hate class, which improves classification accuracy. The article describes in detail the stages of development of the algorithm, its architecture and planned further research.

Key words: hateful content, social networks, neural networks, natural language processing, ruBert, GPT-3.5-turbo, data tagging, text classification, text sentiment detection, k-fold cross-validation.