Выявление слаботоксичных текстов в схожих наборах с использованием модифицированной нейронной сети XLM-RoBERTa и параметров достоверности токсичности

Я. А. Селиверстов
ФГБУН Институт проблем транспорта
им. Н.С. Соломенко РАН;
Университет 20.35
silver8yr@gmail.com

Э. Д. Пословская
Санкт-Петербургский государственный университет
el.poslovskaya@gmail.com

А. А. Комиссаров
Университет 20.35
seliverstov_s_a@mail.ru

А. А. Лесоводская
Университет 20.35
a.lesovodskaya@2035.university

А. В. Подтихов
Университет 20.35;
Национальный исследовательский университет «Высшая школа экономики»
a.podtikhov@2035.university

Аннотация. **В статье рассмотрена задача классификации слаботоксичных текстов с использованием модифицированной нейронной сети трансформерной архитектуры XLM-RoBERTa, обученной на сильнотоксичных текстах. В качестве наборов для выявления слаботоксичных текстов использовались комментарии школы МИФИ и школы педагогического дизайна Университета 20.35. Переобучение сети на слаботоксичных текстах не производилось. Вместо этого классификация слаботоксичных текстов осуществлялась путем варьирования параметра достоверности токсичности. Была построена аппроксимационная зависимость количества слаботоксичных текстов от параметра достоверности токсичности и получено пороговое значение параметра достоверности токсичности, при котором качество классификации слаботоксичных текстов максимально. Также была сформулирована и подтверждена гипотеза схожести токсичности однородных информационных ресурсов.**

Ключевые слова: классификация, машинное обучение, Интернет контент, токсичность, XLM-RoBERTa

I. Введение

Коммуникация является одним из важнейших процессов взаимодействия между людьми в современном обществе. Стремительное развитие информационных технологий, средств связи и социальных сетей за последнее десятилетие трансформировало среду коммуникационного взаимодействия человека, переместив процесс повседневного общения в виртуальное Интернет пространство [1]. Виртуальный мир меняет специфику межличностного общения. Виртуальная коммуникация носит глобальный характер и отличается от реального взаимодействия анонимностью, мультиязычностью, опосредованностью, неконтролируемостью, снижением пределов нравственных и социальных границ, что может приводить к широкому классу враждебных коммуникативных действий – травле, угрозам, умаления деловой репутации, насмешкам, вымогательству, клевете, унижению чести и достоинства, оскорблениям, проявлениям вражды и дискриминации на почве социальной, религиозной, гендерной и национальной нетерпимости, пропаганде наркотических средств, призывам к террористической и экстремистской деятельности, суициду, гражданскому неповиновению и различным формам девиантного поведения. Распространение подобной токсичной информации, может нанести непоправимый ущерб не только отдельному человеку или компании, но и стране в целом. Вот почему решение задач, связанных с выявлением и недопущением распространения информации, приводящей к негативным последствиям, является шагом в высшей степени актуальным и своевременным [2].

II. Анализ предметной области

В настоящее время на основе современных методов глубокого обучения [3, 4, 5] активно разрабатываются и апробируются новые программные модели, способные

распознавать и фильтровать противоправный и вредоносный контент. С 2020 года для выявления токсичного контента начали активно использоваться предобученные нейронные сети трансформерной архитектуры, такие как mBERT [6], XLM [7], ruBert [8], M-USE [9] которые в основном применяются для решения задач межъязыкового понимания [10].

В статье [11] выявляется агрессия в сообщениях пользователей с использованием бинарного классификатора на основе алгоритма случайного леса и сверточной нейронной сети.

В работе [12] предложено использовать неконтролируемый вероятностный метод с исходным словарем для выявления оскорбительных комментариев в социальных сетях на русском и украинском языках.

В работе [13] на основе нейросетевых моделей трансформерной архитектуры разрабатывается новый метод многоклассовой классификации угроз на русском языке, позволяющий снизить количество ложно-положительных предсказаний.

Для разработки качественных многоклассовых классификаторов содержимого вредоносного контента веб-страниц необходимо иметь размеченные соответствующими классами корпуса. Для русского языка в настоящий момент существует всего несколько открытых наборов, а именно:

1. набор оскорбительных комментариев на русском и украинском языках размером около 2000 [12];
2. открытый набор русскоязычных токсичных комментариев, опубликованный на ресурсе Kaggle [14];
3. тот же набор [14], доразмеченный и проверефицированный асессорами с Толоки [8];
4. размеченный набор собранных в социальной сети Вконтакте угроз [13].

Среди последних научных исследований в области токсичности, выполненных международными научными коллективами хочется отметить следующие работы.

В работе [15] разрабатывается модель для обнаружения и классификации враждебных постов и их дальнейшей классификации на фэйки, оскорбления, ненависть и клевету с использованием сверточного реляционного графа сети. Предложенная авторами модель работает на уровне XLM-RoBERTa от Google в данном наборе данных.

В работе [16] авторский коллектив представил новый набор данных для обнаружения враждебности на языке хинди, состоящий из 8200 онлайн-сообщений. Аннотированный набор данных охватывает четыре аспекта враждебности: фейковые новости, язык вражды, оскорбительные и клеветнические сообщения.

В статье [17] авторы представили мультимодальный набор данных ALONE[1] о токсичных взаимодействиях в социальных сетях между учащимися старших классов, вместе с описательными объяснениями. Каждый случай взаимодействия включает твиты, изображения, смайлики и связанные метаданные.

В работе [18] представлен подход, основанный на трансфертном обучении предобученных нейронных сетей, для классификации сообщений в социальных сетях (таких как, Twitter, Facebook и т. д.) на хинди деванагари как враждебные или недружественные.

Анализ предметной области свидетельствует об актуальности исследований по обнаружению сильнотоксичного контента в социальных сетях и веб-ресурсах с использованием предобученных нейронных сетей трансформерной архитектуры.

III. Постановка задачи

Общий тренд исследований негатива в Интернет-пространстве сосредоточен в области выявления токсичного и сильнотоксичного контента, который, как правило, распространен в социальных сетях и часто посещаемых веб-ресурсах. Между тем существует большое количество образовательных и профильных веб-ресурсов, которым свойственен иной тип пользователя. Как правило, это образованные люди, с высшим образованием, вовлеченные в трудовую деятельность различного вида. Таким пользователям свойственны хорошие манеры, сдержанность в высказываниях и проявлениях эмоций. Несмотря на этот факт, на данных веб-ресурсах также возникают острые дискуссии, характеризуемые не сильнотоксичными, а слаботоксичными высказываниями – насмешками, острыми шутками, провокационными высказываниями и скрытыми уколами. К сожалению, в настоящий момент в открытом доступе отсутствуют размеченные корпуса и наборы подобных слаботоксичных текстов необходимых для построения классификаторов.

Основная цель работы заключается в исследовании возможности выявления слаботоксичных текстов с помощью модели бинарного классификатора на основе модифицированной нейронной сети трансформерной архитектуры XLM-RoBERTa, обученной только на токсичных и сильнотоксичных текстах и регулирования параметра достоверности токсичности с учетом принципа схожести однородных Интернет-ресурсов.

В качестве слаботоксичных текстов используются комментарии школы педагогического дизайна Университета 20.35 и школы МИФИ.

IV. Методы и модели

Анализ предметной области показал, что модель классификатора целесообразно разрабатывать на основе многоязычного трансформера XLM-RoBERTa[2] [12].

В модель XLM-RoBERTa-m[3] были внесены следующие архитектурные изменения – masked-Im слой был заменен

[1] ALONE - AdoLescents ON twittEr

[2] fairseq/examples/xlmr at master · pytorch/fairseq · GitHub

на полносвязанный слой. В качестве вычислительной среды использовалась открытая платформа Kaggle.

Алгоритм выявления слаботоксичных отзывов на основе модели XLM-RoBERTa-m и параметра достоверности токсичности представлен на рис. 1.

Модель XLM-RoBERTa-m обучалась на данных соревнования Jigsaw Multilingual Toxic Comment Classification competition[4], проводимого на платформе Kaggle. Данные состояли из обучающего (jigsaw toxic comment train.csv), валидационного (validation.csv) и тестового (test.csv) наборов. Обучающая выборка содержала 223549 аннотированных пользовательских комментариев, взятых со страниц обсуждения в Википедии. Данный набор является крупнейшими общедоступным корпусом на Kaggle.

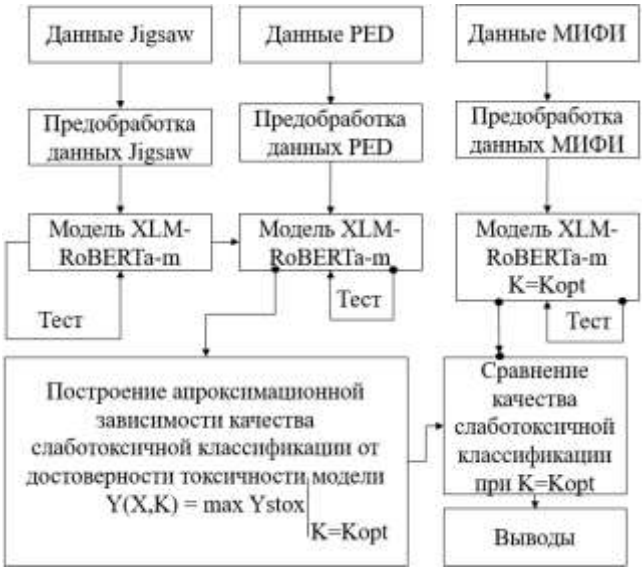

Рис. 1. Алгоритм выявления слаботоксичных отзывов на основе принципа схожести однородных Интернет-ресурсов

Представленные комментарии были размечены экспертами на шесть классов: «Токсичный», «очень токсичный», «оскорбление», «угроза», «непристойный» и «ненависть к идентичности». Комментарии могут быть связаны с несколькими классами одновременно, что образует многокомпонентность классификации.

Перед обработкой текстовых данных используются следующие методы предварительной обработки текста: удаление знаков препинания, лемматизация и удаление стоп-слов.

Качество бинарной классификации обученной модели на тестовых данных составило AUC ROC = 0.9459.

Распределение достоверности токсичности с использованием XLMRoberta-m для набора тестовых данных представлено на гистограмме на рис. 2.

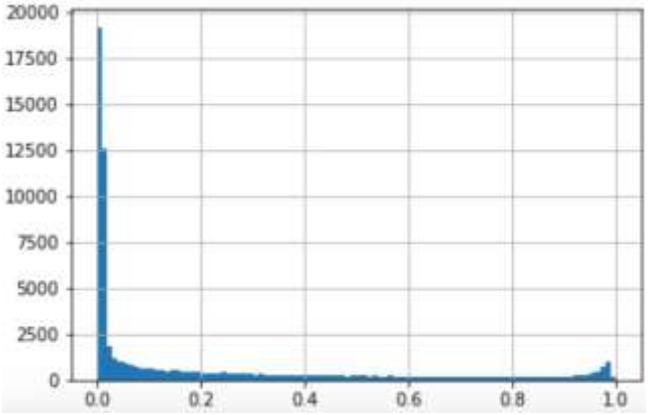

Рис. 2. Распределение достоверности токсичности на модели XLMRoberta-m для набора тестовых данных Jigsaw

Пример токсичных комментариев, отфильтрованных моделью XLM-RoBERTa-m представлен на рис. 3.

id			
216	216	ru	непонимаю...каким идиотом нужно быть...чтобы в...
219	219	ru	Ты , пиз[redacted], ещё ты мне будешь папки в колёса ...
412	412	ru	Абсолютно бредовая часть статьи ! нет там ника...
428	428	ru	Ты не прав потому что ты дурак(с). О господи т...
481	481	ru	Мне года, и...? Я, конечно, многое могу понять...
...	...	...	...
63620	63620	ru	Какой идиот статью писал? Counter-Strike базир...
63672	63672	ru	Дарт - интернетовский дурачок. Где в правилах ...
63677	63677	ru	Меня никогда не интересовали ни гомофобы, ни г...
63730	63730	ru	И желательно еще те, которые завершили службу ...
63769	63769	ru	За время Вашей работы в Википедии Вы проявили ...

Рис. 3. Пример токсичных комментариев

Далее на вход модели в качестве тестового набора подавался корпус школы педагогического дизайна Университета 20.35 со слаботоксичными высказываниями. Корпус состоял из 54352 записей.

Распределение достоверности токсичности с использованием XLMRoberta-m для набора тестовых данных школы педагогического дизайна представлено на гистограмме на рис. 4.

[3] Индекс «m» означает модифицированность модели нейронной сети XLM-RoBERTa

[4] https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/data

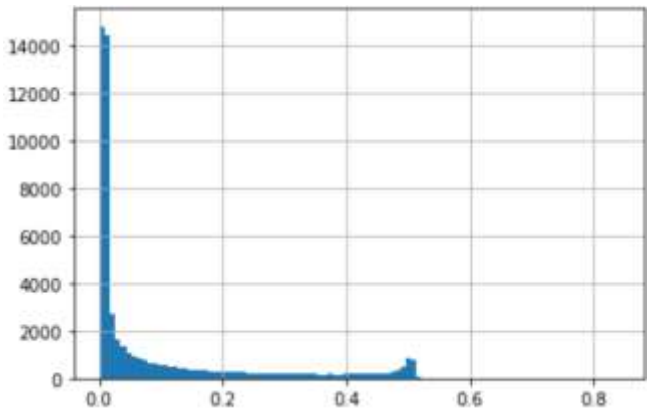

Рис. 4. Распределение достоверности токсичности на XLMRoberta-m для набора тестовых данных школы педагогического дизайна

На рис. 4 видно, что в диапазоне значения от 0.4 до 0.6 параметра достоверности токсичности модели наблюдается рост комментариев. Ручная проверка тестового набора подтвердила, что часть комментариев носит слаботоксичный характер.

Сформулируем гипотезу схожести токсичности однородных информационных ресурсов применительно к образовательным Интернет-ресурсам.

Гипотеза. Комментарии пользователей на схожих Интернет-ресурсах (например, образовательных) обладают схожим уровнем токсичности (слаботоксичности[5]).

Для проверки гипотезы выполним следующие шаги:

1) *построим апроксимационную зависимость количества отзывов от величины достоверности модели по контрольным точкам отобранных экспертным путем, ориентируясь на рис. 4.*

Контрольные точки: 0.03; 0.52; 055; 0.6; 0.75.

В качестве функций аппроксимации рассматривались степенная, квадратичная, показательная, экспоненциальная, кубическая и логарифмическая регрессия. Анализ проводился с использованием ресурса PLANETCLAC[6].

В результате анализа лучшую точность аппроксимации показала кубическая регрессия (Таблица 1).

Уравнение кубической регрессии имеет вид:

$$Y = -103642.5563x3 + 195274.8384x2 - 121590.6997x + 25089.7540$$

ТАБЛИЦА I ИСХОДНЫЕ (Y) И АППРОКСИМИРУЮЩИЕ (YРЕГ) ЗНАЧЕНИЯ КОЛИЧЕСТВА ОТЗЫВОВ ОТ ВЕЛИЧИНЫ ДОСТОВЕРНОСТИ МОДЕЛИ (X)

Аппроксимирующие значения					
X	0.03	0.52	0.55	0.6	0.75
Y	21615	85	54	42	15
Yрег	21614.98	91.93	41.97	47.48	14.62

2) *экспертным путем определим пороговое значение параметра достоверности токсичности, при котором количество слаботоксичных отзывов максимально, количественно и качественно сравнивая наборы слаботаксичных отзывов выявленных при разных значениях достоверности токсичности с эталонным набором.*

В результате сравнительного анализа пороговое значение параметра достоверности токсичности было определено равным K = 0.52.

График аппроксимирующей функции с отмеченным пороговым значением параметра достоверности токсичности представлен на рис. 5.

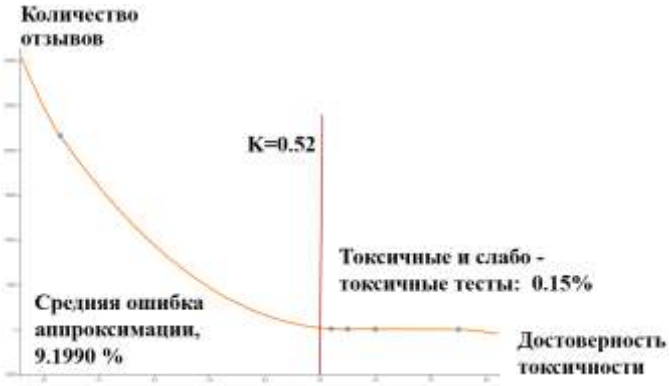

Рис. 5. График аппроксимирующей функции

Количество отфильтрованных отзывов при этом составило 85 (рис. 6).

id	lang	comment_text	id
12735	ru	trello	12735
18488	ru	trello	18488
37176	ru	чушь собачья - второкурсники, за которыми охот...	37176
45188	ru	Алот	45188
48358	ru	Зачем падлет	48358
49353	ru	Ну, столько защитников - должен же кто-то стре...	49353
49664	ru	да.. я там всем отвечаю нашим потеряшкам	49664
50034	ru	сейчас дергающийся глаз на место поставлю	50034
50459	ru	Пксть плюсики ставят	50459
51417	ru	блин. Ну, тогда надо меня там удалить, оно уже...	51417
51441	ru	Злыдни	51441
52551	ru	вкоючите здесь	52551
52902	ru	как правило, они в рефлексии пишут о том как в...	52902
53784	ru	создание рубрикаторов, сор цифрового следа	53784

Рис. 6. Пример выявленных слаботоксичных комментариев в наборе школы педагогического дизайна

[5]Под схожим уровнем слаботоксичности будем понимать схожее количество слаботоксичных комментариев.

[6]https://planetcalc.ru/5992/?xstring=0.03%200.52%200.55%200.60%200.75&ystring=21615%2085%2054%2042%2015&dolinear=0&doquadratic=1&dopower=1&docubic=1&doexponential=1&dologarithmic=1&dohyperbolic=0&dooexponential=1

Таким образом, в наборе из 54352 комментариев педагогической школы Университета 20.35 присутствуют слаботоксичные комментарии в объеме 0.156 %

$$Tox\ (Ped) = (85 * 100) / 54352 = 0.156\ \%$$

3) *оценим работу нашей модели на новом тестовом наборе данных схожей образовательной тематики и полученном на шаге 2 пороговом значении достоверности ткоксичности равным K=0.52.*

В качестве нового тестового набора будем использовать комментарии школы МИФИ. Корпус включает 9929 записей.

Количество отфильтрованных отзывов при заданных значениях составило 14 (рис. 7).

Таким образом, в наборе из 9929 комментариев школы МИФИ присутствуют слаботоксичные комментарии в объеме 0.141 %

$$Tox\ (МИФИ) = (14 * 100) / 9929 = 0.141\ \%$$

id	lang	comment_text	id
12	ru	Посадить это от слова сало	12
2097	ru	я, это еще и менеджеры, айчары и тп, у них с математикой и информатикой, как правило, такая	2097
3272	ru	Вера и вся остальные <ФБ> мы не упарываемся. Провели встречу установочную в удобное вре	3272
3531	ru	Kahoot	3531
4461	ru	Вообще обилие инструментов - это и хорошо, и отвратительно. Хорошо, что почти на любую ид	4461
4984	ru	Сколько можно все по одному и тому же месту завуда на завуда	4984
5297	ru	Коллеги, добрый день. Не пришло приглашение на сегодняшнее мероприятие. Помогите	5297
5518	ru	Тейо, Manti тебе не надоело болтать	5518
6059	ru	Я с тел в наушниках	6059
6229	ru	nleanina	6229
8891	ru	это система для выращивания отдельных ГЕНИЕВ или ВСЕ станут ГЕНИЯМИ	8891
9097	ru	отсталость системы российского образования	9097
9212	ru	Ну Александр Николаевич, Вы то с проектной деятельностью столько лет работаете, а тут люда	9212

Рис. 7. Пример выявленных слаботоксичных комментариев в наборе школы МИФИ

Величина расхождения количества выявленных слаботоксичных отзывов двух различных образовательных школ составляет менее 1 %.

V. Заключение

В результате научно исследовательской работы решена задача выявления и классификации слаботоксичных текстов с использованием модифицированной нейронной сети трансформерной архитектуры XLM-RoBERTa-m обученной на сильнотоксичных текстах и регулирования значения параметра достоверности токсичности.

Сформулирована и в первом приближении подтверждена гипотеза схожести токсичности однородных информационных ресурсов и работоспособность алгоритма выявления слаботоксичных отзывов на основе принципа схожести. Последнее утверждение особо актуально, так как в случае анализа на слаботоксичность схожих Интернет ресурсов и отсутствия размеченных слаботоксичных наборов можно попытаться выявить слаботоксичные комментарии, опираясь только на обученную нейронную сеть трансформерной архитектуры и параметр достоверности модели.

Благодарность

Авторский коллектив благодарит ведущего специалиста ООО «ОЦРВ», группы обработки естественного языка Алексея Шоненкова за ценные рекомендации и замечания в ходе работы над моделью нейронной сети XLM-RoBERTa.

Список литературы

[1] Brignall Thomas & Valey Thomas. (2005). The impact of Internet communications on social interaction. Sociological Spectrum. 25. 335-348. 10.1080/02732170590925882.

[2] Naslund J.A., Bondre A., Torous J. et al. Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice. J. technol. behav. sci. 5, 245–257 (2020). https://doi.org/10.1007/s41347-020-00134-x

[3] Seliverstov Y.; Seliverstov S.; Malygin I.; Korolev O. Traffic safety evaluation in Northwestern Federal District using sentiment analysis of Internet users' reviews. Transp. Res. Procedia 2020, 50, 626–635. DOI: 10.1016/j.trpro.2020.10.074

[4] Aken B. van et al.: Challenges for toxic comment classification: An in-depth error analysis. In: Proceedings of the 2nd workshop on abusive language online (ALW2). pp. 33–42. Association for Computational Linguistics, Brussels, Belgium (2018).

[5] Risch J., Krestel R.: Toxic comment detection in online discussions. In: Deep learning-based approaches for sentiment analysis. pp. 85–109. Springer (2020).

[6] Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // NAACL-HLT (2019). DOI:10.18653/v1/N19-1423

[7] Lample G., Conneau A. Cross-lingual language model pretraining // 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, arXiv preprint arXiv:1901.07291, 2019

[8] Smetanin S. Toxic Comments Detection in Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference, Dialogue 2020, 10.28995/NNNN-NNNN-2020-19-1-11.

[9] Yang Y et al. 2019. Multilingual universal sentence encoder for semantic retrieval. In arXiv preprint arXiv:1907.04307

[10] Conneau Alexis et al. Unsupervised Cross-lingual Representation Learning at Scale // ACL (2020). DOI:10.18653/v1/2020.acl-main.747

[11] Potapova R., Gordeev D. Detecting State of Aggression in Sentences Using CNN. 240-245. 10.1007/978-3-319-43958-7_28.

[12] Andrusyak B. et al.: Detection of abusive speech for mixed sociolects of russian and ukrainian languages. In: The 12th workshop on recent advances in slavonic natural languages processing, RASLAN 2018. pp. 77-84.

[13] Zueva N., Kabirova M., Kalaidin P. Reducing Unintended Identity Bias in Russian Hate Speech Detection // Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 65–69, https://doi.org/10.18653/v1/P17, arXiv preprint arXiv:2010.11666

[14] Belchikov A. Russian language toxic comments, https://www.kaggle.com/ blackmoon/russian-language-toxic-comments.

[15] Davidson Thomas & Warmsley, Dana & Macy, Michael & Weber, Ingmar. (2017). Automated Hate Speech Detection and the Problem of Offensive Language.

[16] Mohit Bhardwaj, et al. 2020. Hostility Detection Dataset in Hindi. arXiv:2011.03588.

[17] Wijesiriwardene Thilini & Inan Hale & Kursuncu Ugur & Gaur Manas & Shalin Valerie & Thirunarayan Krishnaprasad & Sheth Amit & Arpinar Ismailcem. (2020). ALONE: A Dataset for Toxic Behavior among Adolescents on Twitter.

[18] Gupta Ayush & Sukumaran Rohan & John Kevin & Teki Sundeep. (2021). Hostility Detection and Covid-19 Fake News Detection in Social Media.