DOI 10.36622/VSTU.2022.25.2.002
УДК 004.02 : 004.05

МЕТОДЫ СНИЖЕНИЯ ШУМОВЫХ ФАКТОРОВ ПРИ ВЫЯВЛЕНИИ КОНТЕНТА ЭКСТРЕМИСТСКОГО ХАРАКТЕРА В СОЦИАЛЬНЫХ МЕДИА

В.А. Минаев, Е.С. Поликарпов, А.В. Симонов

Цель исследования состоит в поиске методов снижения влияния шумовых эффектов при выявлении текстового контента экстремистского характера в социальных медиа. Осуществлен обзор методов, применяемых для выявления экстремистского контента, проанализированы их преимущества и недостатки. Сформирован экспериментальный текстовый корпус, моделирующий структуру реальных текстовых данных из социальных медиа и содержащий контент экстремистского характера о радикальном исламе. Описан процесс предобработки корпуса. Проанализированы методы классификации текста. Представлена работа моделей-трансформеров на примере трансформера BERT. Разработаны архитектуры классификаторов на основе BERT, используя различные методы многоклассовой и бинарной классификации. Проведены эксперименты по выявлению и классификации текстовых данных, содержащих экстремистский контент. Оценена точность и F-мера классификаторов в каждом эксперименте, обоснован выбор наиболее качественного классификатора для снижения шумовых факторов при выявлении контента экстремистского характера. Результаты работы классификатора на основе BERT превосходят результаты использования глубинных нейронных сетей (GRU, LSTM, CNN). Полученную модель классификатора и подход к снижению влияния шумовых факторов целесообразно применять при разработке систем мониторинга и выявления деструктивной информации в социальных медиа.

Ключевые слова: снижение шума, текстовый контент, экстремизм, социальные медиа, трансформер BERT, глубокое обучение, классификация.

Введение

Социальные медиа (СМ) являются одним из самых популярных средств распространения информации. К СМ относятся социальные сети, различные сетевые форумы, а также системы обмена мгновенными сообщениями (мессенджеры). В связи с высокой доступностью и широким охватом аудитории СМ используют различные запрещенные в Российской Федерации сообщества и организации для распространения, популяризации и пропаганды деструктивных идей и действий.

Государственные органы активно противодействуют распространению деструктивного контента и рассматривает его, согласно Доктрине информационной безопасности Российской Федерации, утвержденной Указом Президента Российской Федерации № 646 от 05.12.2016, как одну из основных информационных угроз. Выявлением и пресечением указанного контента активно занимаются как государственные структуры, так и российские компании, подписавшие хартию безопасности детей в Интернет.

Как и прежде, основным форматом распространения запрещенного контента в СМ является текст в связи с легкостью передачи, простотой распространения и изменяемости.

Выявление деструктивного текстового контента осуществляется как силами экспертов, так и специализированными автоматизированными системами. Но поскольку ежемесячно создается более одного миллиарда публичных публикаций нового текстового контента в наиболее популярных СМ, и это число только возрастает, способы экспертного анализа малоэффективны, вследствие чего основной акцент в перспективе делается на автоматизированном подходе.

При этом автоматизированные системы поиска деструктивного контента, как правило, не обладают возможностью отличать его от похожего по синтаксису, но отличного по семантике так называемого

«шумового» контента.

Снижению влияния шумовых эффектов при выявлении деструктивного контента автоматизированными системами, а также способам минимизации ошибки ложных срабатываний в таких системах посвящена настоящая статья.

Методы и этапы исследования

Автоматизированные методы обнаружения деструктивного контента условно можно разделить на два класса: статистические методы и методы машинного обучения.

При *статистическом подходе* применяется как поиск по уникальным словам, присущим только деструктивным текстам, так и по частотности используемых слов в материалах, находящихся в Федеральном списке экстремистских материалов.

Данный подход относительно прост в реализации, однако не учитывает расположение слов в предложениях, в связи с чем невозможно определить контекст тех или иных слов, что напрямую влияет на возможность удаления шумовых корпусов. Во многом именно в связи с этим все большее предпочтение отдается методам машинного обучения.

Существует два основных варианта применения *машинного обучения* при выявлении деструктивных текстов: тематическое моделирование и классификация текста.

Тематическая модель текстовых документов определяет, какую тему раскрывает тот или иной текст или документ, и какие слова (термины) являются образующими в каждой теме. Указанные модели могут присваивать несколько тем (классов) одному тексту, что, в свою очередь, далеко не всегда позволяет точно определить его деструктивную составляющую, так как присвоение одной и той же тематики может осуществляться как применительно к незаконному, так и легальному контенту.

Классификация текста – это задача отнесения объектов к заданным классам (группам). С помощью классификации решаются многие задачи обработки текстов на естественном языке, например, выявление токсичных комментариев в СМ [1] или распознавание спама в СМ «Twitter» [2]. Исходя из анализа результатов решаемых задач, методы классификации, как показывают эксперименты, являются наиболее подходящими для выявления деструктивного контента.

Так как модели машинного обучения должны получать на входе последовательность чисел, то при решении задач классификации необходимо отобразить текст в числовом и векторном виде.

Для этого осуществляется операция токенизации – разделения последовательности слов текста на более мелкие составляющие, размером не более 2-3 символов, которые в зависимости от применяемого токенизатора далее представляются в виде числового значения. После операции токенизации текста получают числовую последовательность, используемую для дальнейшей обработки и классификатора.

Существует ряд моделей классификаторов, начиная от статистических моделей, наиболее популярная из которых – логистическая регрессия, до различных архитектур глубоких нейронных сетей [3]. Наиболее распространенными среди последних выступают модели-трансформеры, представленные в 2017 году Google Brain [4].

Трансформеры используются как в задачах машинного перевода (наиболее известная модель GNMT [5]), так и в задачах генерации текста (модели серии GPT-3 [6]).

Наиболее популярным трансформером для классификации текста является BERT и его производные, состоящий в некоторых случаях из 24 искусственных нейронных сетей [7].

Основные элементы трансформера – это элементарные блоки искусственных нейронных сетей, называемые энкодерами и декодерами.

В моделях машинного перевода используются оба блока, с помощью декодера последовательность сначала отображается в общее для всех языков векторное пространство, а далее с помощью энкодера преобразовывается на том или ином языке.

В трансформерах, используемых для

классификации текста, в том числе трансформере BERT применяются только блоки энкодеров.

Любой классификатор, основанный на моделях-трансформерах, состоит из трех частей: токенизатора, базы и головы классификатора.

Токенизатор преобразует текстовую последовательность в числовую, база классификатора преобразует полученную числовую последовательность в векторное представление предложения.

В качестве головы классификатора, на вход которого подаются выходные данные из базы в виде вектора, могут выступать как статистические алгоритмы, так и различные архитектуры искусственных нейронных сетей.

Для тонкой настройки трансформера под конкретную задачу классификации [8] необходимо добавлять слой из нейронных сетей с обратной связью к последнему слою базы классификатора.

В соответствии с изложенным для реализации методов снижения влияния шумовых факторов выбран классификатор на основе трансформера BERT. Для чего использовалась библиотека для работы с трансформерами HuggingFace и модель RuBERT, адаптированная под русский язык [9].

Так как основной нашей задачей является поиск наиболее результативной архитектуры, позволяющей бороться с ошибками первого рода, вызванными шумовым контентом, в качестве головы использовался однолинейный слой нейронной сети.

В качестве обучающей выборки выбраны следующие корпуса текстов – СМ корпус, в который вошли русскоязычные публикации в СМ «Twitter» и новости интернет-издания «Lenta.ru». Данный корпус выбран для моделирования основного текстового наполнения СМ.

Радикальный корпус состоял из запрещенных к распространению в России текстов по теме радикального ислама, находящихся в списках ФСЭМ, и использован как класс, точность обнаружения которого необходимо повышать.

Шумовой корпус, составленный из религиозной литературы и одобрен муфтиями России к изучению и распространению, имеет большое количество похожих слов-пересечений с радикальным корпусом, но содержащим другую смысловую составляющую.

Каждый корпус сформирован из ста тысяч предложений со средним количеством слов в них, равном 13,7. После формирования каждый корпус подвергнут предварительной обработке, а именно: удалению специальных символов, приведению к общему виду некоторых имен собственных, а также установке по всем корпусам нижнего регистра.

Все модели используемых классификаторов подверглись так называемой «тонкой настройке» – адаптации работы классификатора под конкретные корпуса.

Во всех экспериментах для обучения классификатора отводится 80% выборки, а оставшиеся 20% используются для оценки его работы.

Для оценки работы классификатора обычно используют соотношение [10]:

$$Ac=P/N\,, \qquad (1)$$

где P – количество правильно определенных текстов; N – количество всех исследуемых текстов.

Необходимо отметить, что (1) не учитывает структуру текстов. Чтобы избежать данного недостатка, для оценки работы классификатора будем использовать F-меру [10]:

$$F = (Prec \cdot Rec)/(Prec + Rec) \qquad (2)$$

где $Prec$ – определенность работы классификатора, рассчитываемая как:

$$Prec = TP/(TP + FP)$$

Rec – полнота работы классификатора, рассчитываемая по формуле:

$$Rec = TP/(TP + FN)\ ,$$

где TP – истинно-положительное решение классификатора;

TN – истинно-отрицательное решение классификатора;

FP – ложноположительное решение классификатора;

FN – ложноотрицательное решение классификатора.

F-мера учитывает как полноту, так и определенность классификации, что позволяет оценивать качество модели более объективно.

Перейдем к описанию результатов пяти проведенных экспериментов, направленных на оценку распознавания классов исследуемых текстов в их различных сочетаниях.

Эксперимент № 1 – моделирование ситуации, при которой отсутствуют шумовые текстовые корпуса, а присутствует только СМ-контент и радикальные тексты.

Эксперимент № 2 – моделирование ситуации, в которой применяется классификатор, обученный на выборке из эксперимента № 1, то есть только на СМ и радикальном корпусе. Но классифицирует в том числе и тексты из шумового корпуса, добавленные в СМ класс. Целью эксперимента является задача показать, насколько снижаются результаты классификатора, если при его обучении не учитывать шумовые составляющие.

Эксперимент № 3 – расширение классификатора текстов путем добавления шумового класса. В данном случае применяется F-мера по трем рассматриваемым классам с целью учесть влияние шумового корпуса на радикальный.

Эксперимент № 4 – применение бинарной классификации, когда тексты разделяются на 2 класса – нерадикальный и радикальный. Причем в нерадикальном корпусе смешиваются корпус СМ с шумовым корпусом.

Эксперимент №5 – создание сложной модели, состоящей из двух последовательных классификаций. Первая отделяет СМ от совместного класса шумового и радикального корпуса, а вторая разделяет шумовой и радикальный класс. Тем самым проверяется гипотеза снижения ошибок первого рода за счет двойной классификации.

Результаты экспериментов

Результаты эксперимента № 1, в рамках которого проводилась классификация по классам СМ и радикальный, представлены в табл. 1, где указана рассчитанная Ac для каждого класса.

В таблицах использованы следующие обозначения классов:

СМК – класс текстов социальных медиа;

РК – класс радикальных текстов;

ШК – класс шумовых корпусов текстов.

По строках расположены значения истинного класса, а в столбцах точность предсказания данного класса.

Таблица 1

Результаты эксперимента № 1

Истинный класс			
	СМК	**0,96**	**0,04**
	РК	**0,05**	**0,95**
		СМ	РК
		Предсказанный класс	

Полученные результаты, без влияния шумовых факторов, можно рассматривать, как «идеальные», к которым необходимо стремиться. Точность отделения РК от СМ достаточно высокая, что говорит об имеющихся больших различиях в текстовом контенте исследуемых корпусов.

Результаты эксперимента № 2 представлены в табл. 2. Использовался классификатор из первого эксперимента в режиме оценки, без обучения.

Таблица 2

Результаты эксперимента № 2

Истинный класс			
	СМК+ШК	**0,81**	**0,19**
	РК	**0,15**	**0,85**
		СМК+ШК	РК
		Предсказанный класс	

Как видно из табл. 2, классификатор из эксперимента №1 при добавлении шумового

корпуса существенно снижает точность классификации – в случае СМК снижение произошло на 15%, в случае РК на 10%.

Исходя из соотношения возросших ошибок можно сделать вывод о действительно высоком влиянии синтаксически похожих шумовых корпусов на тексты из РК. Данный результат является условно нижней границей точности классификации, которую необходимо превышать.

Результаты эксперимента №3, где применена классификация на три класса (СМ, ШК и РК), представлены в табл. 3.

Таблица 3

Результаты эксперимента № 3

Истинный класс				
	СМК	**0,99**	**0,011**	**0,0037**
	РК	**0,02**	**0,87**	**0,11**
	ШК	**0,0048**	**0,08**	**0,92**
		СМК	РК	ШК
		Предсказанный класс		

Как видно из результатов третьего эксперимента точность определения класса СМК выше на 3%, чем в условно «идеальном» первом эксперименте, достигая почти 100%. Точность распознавания РК, в свою очередь, снизилась на 8%, но при этом на 2% превысила оценочный результат эксперимента № 2. ШК распознается достаточно высоко. Однако стоит обратить внимание на F-меру всего классификатора, результаты которой приведены в конце раздела.

Результаты эксперимента № 4 представлены в табл. 4.

Таблица 4

Результаты эксперимента № 4

Истинный класс			
	СМК+ШК	**0,94**	**0,057**
	РК	**0,063**	**0,94**
		СМК+ШК	РК
		Предсказанный класс	

Результаты, представленные в табл. 4, схожи с результатами первого эксперимента, в котором тексты ШК не фигурировали.

В эксперименте № 5 использован составной классификатор, состоящий из двух последовательных бинарных, поэтому рассмотрим три таблицы.

Результаты работы первичного классификатора, отделяющего СМ от ШК+РК, представлен в табл. 5.

Таблица 5

Результаты работы первого классификатора в эксперименте № 5

Истинный класс			
	СМК	**0,996**	**0,014**
	РК+ШК	**0,013**	**0,997**
		СМК	РК+ШК
		Предсказанный класс	

Первый классификатор показывает высокую точность работы и способность отделять обычное наполнение СМ от специфических текстов (ШК или РК).

Результаты работы второго классификатора, отделяющего ШК от РК, представлены в табл. 6.

Таблица 6

Результаты работы второго классификатора в эксперименте № 5

Истинный класс			
	ШК	**0,925**	**0,075**
	РК	**0,1**	**0,9**
		ШК	РК
		Предсказанный класс	

Однако результаты второго составного классификатора несколько хуже первого.

Для получения общего результата точности классификатора в эксперименте № 5 перемножим полученные точности ШК и РК на точность РК+ШК. Результаты представлены в табл. 7.

Таблица 7

Суммарные результаты в эксперименте № 5

Класс	Суммарная *Ac*
СМК	0,996
ШК	0,9222
РК	0,8973

Как видно из табл. 7, результаты суммарного составного классификатора не отличаются высокой точностью, хотя первый классификатор показывал практически 100% точность при бинарной классификации.

На рис. 1 в виде диаграммы представлены результаты расчета *F*-меры классификаторов, используемых в экспериментах.

В соответствии с полученными результатами наиболее близкой к эксперименту № 1 является модель классификатора в эксперименте № 3, в котором проводится бинарная классификация.

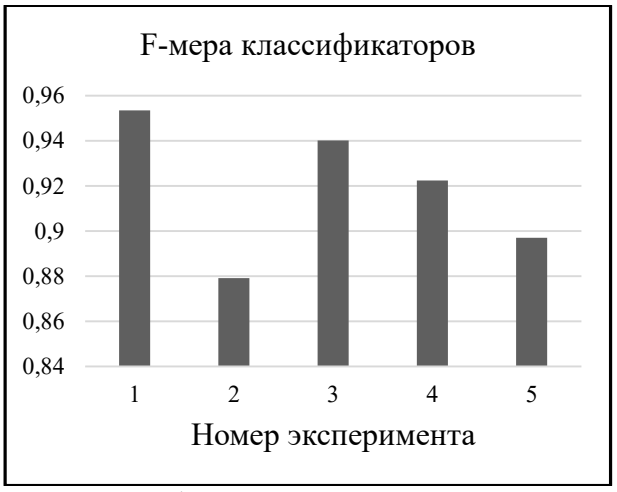

Рис. 1. Результаты расчета *F*-меры классификаторов

Обсуждение и выводы

В исследовании проведены эксперименты по поиску методов снижения шумовых эффектов (реализуемых в частотно и синтаксически похожих корпусах текстов) в процессе выявления в СМ контента экстремистского характера, в частности – радикального ислама.

Для решения указанной задачи сформированы корпуса текстов, моделирующие реальное наполнение СМ.

Исследованы наиболее распространенные методы классификации и выявления контента экстремисткой направленности, а также проведен анализ моделей, показывающих высокие результаты при решении задач классификации текста.

Показано, что эффективнее всего снижает шумовые факторы при выявлении в СМ контента экстремистского характера модель-классификатор из эксперимента №3. Она показывает наиболее высокую точность F-меры как показателя эффективности работы классификатора.

Результаты работы данного классификатора превосходят результаты использования глубинных нейронных сетей (GRU, LSTM, CNN), приведенных в исследовании [3].

Полученную модель классификатора и подход к снижению влияния шумовых факторов целесообразно применять при разработке систем мониторинга и выявления деструктивной информации в СМ.

Дальнейшее развитие и перспективы исследования методов распознавания деструктивного контента в социальных медиа видятся в:

- поиске методов классификации текстов, превышающих по объему токенизированный вектор для модели BERT (не более 256 токенов);
- определении наиболее эффективной модели BERT, используемой в качестве базы, а также обосновании головы ее классификатора.

В заключение подчеркнем, что современный этап развития общества характеризуется ускоряющейся цифровизацией, всё более широким использованием информационно-аналитических систем в управлении социальными процессами, включая перспективные технологии искусственного интеллекта, постоянным совершенствованием средств связи и возрастающей ролью информационно-коммуникационной среды, оказывающей влияние на все аспекты жизни российских граждан.

Поэтому целевую основу дальнейших исследований тематики, рассматриваемой в статье, составляют проблемы, связанные с ключевыми угрозами кибербезопасности, в числе которых в Доктрине информационной

безопасности Российской Федерации названы информационно-психологические воздействия (ИПВ) на индивидуальное, групповое и общественное сознание.

При расширяющейся палитре методов обеспечения кибербезопасности страны, общества, государства новые разработки в области анализа и прогнозирования ИПВ могут стать серьёзным вкладом в борьбу с проявлениями терроризма и экстремизма, информационно всё активнее и заметнее проявляющихся в социальных медиа.

Учитывая многоплановость, обширный спектр информационно-психологических аспектов проблемы кибербезопасности, её серьёзный научный и практический объём и потенциал, исследования в этом направлении предполагают решение многих новых задач анализа, оценки и прогнозирования распространения деструктивной информации в информационных сетях.

Для практики ценность получаемых при этом научных результатов заключается в новых возможностях методологического и методического обеспечения аналитической деятельности, связанной с обработкой и анализом больших данных (Big Data).

Список литературы

1. Van Aken B. Challenges for toxic comment classification: An in-depth error analysis / Van Aken B. et al. // Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). 2018. С. 33-42.
2. Wang X. Drifted Twitter spam classification using multiscale detection test on KL divergence / Wang X. et al. // IEEE Access. 2019. Т. 7. С. 108384-108394.
3. Минаев В. А., Поликарпов Е. С., Симонов А. В. Применение глубинных нейронных сетей для выявления деструктивного контента в социальных медиа // Информация и безопасность. 2021. Т. 24. № 3. С. 361-372.
4. Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. 2017. Т. 30. С. 1-15.
5. Wu Y. et al. Google's neural machine translation system: Bridging the gap between human and machine translation // arXiv preprint arXiv:1609.08144. 2016. С. 1-23.
6. Brown T. B. et al. Language models are few-shot learners // arXiv preprint arXiv:2005.14165. 2020. С. 1-10.
7. Devlin J. et al. Bert: Pretraining of deep bidirectional transformers for language understanding //arXiv preprint arXiv: 1810. 04805. 2018. С. 1-16.
8. Sun C. et al. How to fine-tune BERT for text classification? //China National Conference on Chinese Computational Linguistics. – Springer, Cham, 2019. С. 194-206.
9. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language // arXiv preprint arXiv:1905.07213. 2019. С 1-7.
10. Forman G. et al. An extensive empirical study of feature selection metrics for text classification //J. Mach. Learn. Res. 2003. Т. 3. №. Mar. С. 1289-1305.

Московский университет МВД России им. В.Я. Кикотя
V. Ya. Kikot Moscow University of the Russian Internal Affairs Ministry

Московский государственный технический университет им. Н. Э. Баумана
Bauman Moscow State Technical University

Поступила в редакцию 15.01.2022

Информация об авторах

Минаев Владимир Александрович – д-р техн. наук, профессор, профессор кафедры специальных информационных технологий, Московский университет МВД РФ им. В.Я. Кикотя, e-mail: m1va@yandex.ru

Поликарпов Евгений Сергеевич – канд. техн. наук, доцент, начальник кафедры специальных информационных технологий, Московский университет МВД РФ им. В.Я. Кикотя, e-mail: binox@mail.ru

Симонов Александр Валерьевич – аспирант кафедры защиты информации, Московский государственный технический университет им. Н. Э. Баумана, e-mail: san.siman@yandex.ru

METHODS OF REDUCING NOISE FACTORS IN THE PROCESS OF IDENTIFICATION EXTREMIST CONTENT IN SOCIAL MEDIA

V.A. Minaev, E.S. Polikarpov, A.V. Simonov

The aim of the study is to find methods to reduce the impact of noise effects when identifying extremist text content in social media. A review of the methods used to identify extremist content is carried out, their advantages and disadvantages are analyzed. An experimental text corpus has been formed that simulates the structure of real text data from social media and contains extremist content about radical Islam. The process of preprocessing the case is described. The methods of text classification are analyzed. The work of transformer models is presented on the example of the BERT transformer. BERT-based classifier architectures have been developed using various methods of multiclass and binary classification. Experiments were conducted to identify and classify text data containing extremist content. The results of the BERT-based classifier are superior to the results of using deep neural networks (GRU, LSTM, CNN). The resulting classifier model and approach to reducing the influence of noise factors should be used when developing monitoring systems and identifying destructive information in social media

Keywords: noise reduction, text content, extremism, social media, transformer BERT, deep learning, classification.

Submitted 15.01.2022

Information about the authors

Vladimir A. Minaev – Dr. Sc. (Technical), Professor, Professor of the Department of Special Information Technologies, Moscow University of the Ministry of Internal Affairs of Russia, e-mail: m1va@yandex.ru

Evgeny S. Polikarpov – Cand. Sc. (Technical), Associate Professor, Head of the Special Information Technologies Department, V. Ya. Kikot Moscow University of the Internal Affairs Ministry of the Russian Federation, e-mail: binox@mail.ru

Alexander V. Simonov – Post-graduate student of the Information Security Department, Bauman Moscow State Technical University, e-mail: san.siman@yandex.ru