ИЗДАТЕЛЬСТВО ГРАМОТА

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2024. Том 17. Выпуск 5 | 2024. Volume 17. Issue 5
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Выявление «токсичности» в социальных сетях на основании критерия семантической близости

Курганская Е. В., Степанова Н. В.

Аннотация. Цель исследования заключается в проверке действенности метода автоматического выявления «токсичных» комментариев пользователей в социальных сетях на основании семантической близости. В статье проводится лингвистический анализ примеров «токсичного» поведения, определяются критерии «токсичности» и основные лексические и стилистические особенности «токсичных» текстов. Исследование последних работ по теме дает общее представление об актуальных методах выявления «токсичности». Выполняется тестирование решения для определения «токсичных» комментариев, основанного на идее отсутствия семантической близости между текстом поста и «токсичным» комментарием. Научная новизна состоит в том, что в работе впервые предлагается использовать критерий семантической близости для выявления «токсичных» комментариев, что представляет собой довольно простое и эффективное решение. Более того, в рамках наиболее популярной русскоязычной социальной сети «ВКонтакте» исследования такого рода ранее не проводилось. В результате исследования установлено, что определение семантической близости между постом и комментарием является достаточно эффективным способом определения релевантности комментария и, следовательно, его вероятного «токсичного» оттенка. Также было выяснено, что метрика косинусной близости подходит для проведения экспериментов по выявлению «токсичности», но для улучшения результатов может быть дополнена другими методами машинного обучения.

EN

Identification of “toxicity” in social networks based on the semantic proximity criterion

E. V. Kurganskaia, N. V. Stepanova

Abstract. The aim of the research is to check the effectiveness of the method of automatic identification of “toxic” comments of users in social networks based on semantic proximity. The article carries out a linguistic analysis of examples of “toxic” behavior, defines the criteria of “toxicity” and the main lexical and stylistic features of “toxic” texts. The analysis of the latest works on the topic gives a general idea of the current methods of identifying “toxicity”. A solution for identifying “toxic” comments based on the idea of the lack of semantic proximity between the text of the post and the “toxic” comment is tested. The scientific novelty lies in the fact that the work proposes for the first time to use the criterion of semantic proximity to identify “toxic” comments, which is a fairly simple and effective solution. Moreover, such studies have not been conducted earlier within the framework of the most popular Russian-language social network VKontakte. As a result of the research, it was found that determining the semantic proximity between a post and a comment is a fairly effective way to determine the relevance of a comment and, consequently, its probable “toxic” connotation. It was also found that the cosine similarity metric is suitable for conducting experiments to identify “toxicity”, but to improve the results, it can be supplemented with other machine learning methods.

Введение

Актуальность данного исследования обусловлена быстрыми темпами цифровизации общества, требующими разработки новых эффективных подходов к анализу интернет-дискурса. Повсеместное распространение интернет-коммуникации приводит к необходимости повышения качества сетевого общения. С этой точки зрения проблема выявления «токсичных» комментариев является несомненно актуальной для современной дискурсологии, что и определяет востребованность настоящего исследования. «Токсичное» и агрессивное поведение представляет большой интерес для зарубежных и отечественных ученых, что отражено в работах

Научная статья (original research article) | https://doi.org/10.30853/phil20240231

(Платонов, Руденко, 2022; Aken, Risch, Krestel et al., 2018; Andrusyak, Rimel, Kern, 2018; Khieu, Narwal, 2019; Hao, Weiguan, Hanyan, 2018; Risch, Krestel, 2020; Smetanin, 2020).

Для достижения указанной цели исследования необходимо решить следующие задачи:

– проанализировать существующие подходы к определению термина «токсичность»;
– провести лингвистический анализ примеров «токсичного» поведения в социальных сетях;
– определить основные критерии «токсичности» пользователя;
– рассмотреть актуальные решения для автоматического выявления «токсичности» в социальных сетях;
– изучить методы определения семантической близости между текстами, выбрать метод, соответствующий целям эксперимента;
– провести эксперимент по определению «токсичности» комментариев на основании их семантической близости с постами;
– оценить работоспособность метода на основании метрик оценки эффективности алгоритма;
– предложить действия по улучшению работы алгоритма выявления токсичности.

Материалом исследования послужил корпус размером 9642 комментария к 417 постам из следующих сообществ социальной сети «ВКонтакте»: «РИА Новости» (https://vk.com/ria), «Подслушано Уфа» (https://vk.com/ufa_overhear), «AstroAlert | Наблюдательная астрономия» (https://vk.com/astro.nomy), «Пикабу» (https://vk.com/pikabu), «Телеканал ТНТ» (https://vk.com/tnt), “IGM” (https://vk.com/igm), «Авторадио» (https://m.vk.com/avtoradio), “Rosetta | Destiny 2” (https://vk.com/rosettad2).

Теоретической базой исследования послужили труды Н. Д. Арутюновой (1990) и В. В. Красных (2002), посвященные трактовке дискурса как специфического коммуникативного события, где дискурс – это одновременно и процесс языковой деятельности (коммуникация, контекст), и ее результат (текст). Кроме того, также учитывались работы Е. Н. Галичкиной (2001), Е. Г. Грибовод (2013), О. В. Лутовиновой (2009), Е. К. Русанова (2016), А. А. Ушакова (2010) и Е. С. Юртаевой (2016), представляющие основные подходы к трактовке терминов «интернет-дискурс», «компьютерный дискурс», «виртуальный дискурс», «сетевой дискурс». Важным аспектом исследования явилась дифференциация понятий «сетевой дискурс» и «дискурс социальных сетей», что стало возможным благодаря трудам А. А. Ефановой и А. А. Осокина (2022), В. И. Карасика (2019), М. А. Павлова (2017) и А. С. Рябовой (2020). Большое значение имеют работы, представляющие подходы к трактовке понятия «токсичность» (Буряковская, Дмитриева, 2022; Ионова, 2018; Сундиев, Смирнов, 2020; Овинова, Шрайбер, 2022).

Классификации «токсичного» и агрессивного поведения в зарубежных соцсетях посвящено значительное количество исследований. Авторы статьи (Aken, Risch, Krestel et al., 2018) используют нейронные сети, работая с комментариями в обсуждениях Википедии и данными из Twitter. Актуальность для нынешнего исследования имеет анализ причин ошибок первого (False Positive) и второго рода (False Negative). Причиной ошибок False Positive авторы считают использование нецензурных слов в ложных срабатываниях и цитаты. Ошибки False Negative появляются в случае токсичности предложения без использования ругательств, а также из-за сарказма и иронии.

В работе (Risch, Krestel, 2020) также применяются различные подходы глубокого обучения, такие как сверхточная нейронная сеть, сети LSTM и GRU. Обучающими данными стали Yahoo News Annotated Comments Corpus и One Million Posts Corpus, в отличии от предыдущей работы, выполняется мультиклассовая, а не бинарная классификация. Интерес представляет статья (Hao, Weiguan, Hanyan, 2018), где в качестве базовых алгоритмов классификации выбраны наивный байесовский классификатор и метод опорных векторов, которые дополняются двумя моделями: LSTM и BERT. В статье (Khieu, Narwal, 2019) также используются методы глубокого обучения, а в качестве датасета выбран Kaggle Toxic Comments Classification Challenge Dataset.

Чрезвычайно важными для текущего исследования выступают работы (Smetanin, 2020) и (Платонов, Руденко, 2022) где внимание уделяется конкретно выявлению «токсичных» высказываний и используются соответствующие датасеты. В статье (Smetanin, 2020) на основании Russian Language Toxic Comments Dataset применяются несколько языковых моделей: классификатор Multinomial Naive Bayes, нейросеть Bidirectional Long Short-Term Memory, а также модели M-BERT, ruBERT и M-USE. Оценка результатов бинарной классификации показала высокое качество настройки языковых моделей, лучший результат (F1 = 92,20%) продемонстрировала модель ruBERT. Исследование (Платонов, Руденко, 2022) представляет собой всеобъемлющий анализ методов векторного преобразования текстов, классификации текста и нейросетевых подходов (LSTM, CNN).

Выбор методов исследования обусловлен целью и совокупностью поставленных задач. Прежде всего для отбора материала из социальной сети «ВКонтакте» использовались такие методы для сбора данных, как wall.getComments (https://vk.com/dev/wall.getComments) и groups.getById (https://vk.com/dev/groups.getById). В работе также использован метод систематизации информации, структурирующий собранные посты и комментарии при помощи базы данных (PostgreSQL. https://www.postgresql.org/).

В качестве методов также применяются дискурсивный анализ и стилистический анализ, позволяющие составить речевой портрет токсичного пользователя на основании разбора коммуникативных ситуаций и анализа коннотаций использованных лексем. Кроме этого, в работе применены квантитативные методы для предобработки собранных текстов и такие инструменты, как библиотека pymystem3 (https://pypi.org/project/pymystem3/) и модуль NLTK stopwords (https://pypi.org/project/nltk/).

Анализ отобранных данных осуществлен методами машинного обучения и дистрибутивной семантики. Внедрение метода моделирования производится при работе с векторными представлениями слов и алгоритмами

word2vec, реализованными в библиотеке gensim (https://pypi.org/project/gensim/). «Токсичность» комментариев определяется при помощи метода компонентного анализа и меры косинусной близости (Scikit-learn. https://pypi.org/project/scikit-learn/). Усовершенствование работы моделей производится методом взвешивания термов tf-idf, реализованным в модуле tfidfvectorizer библиотеки Scikit-learn.

Практическая значимость исследования заключается в эмпирической проверке алгоритма для выявления «токсичных» сообщений в текстах онлайн-сообществ. Результаты исследования должны способствовать решению ряда задач прикладной лингвистики и социолингвистики, в особенности оптимизации коммуникативной и социальной функции языка. Выявление «токсичных» комментариев и пользователей посредством оценки релевантности комментариев по отношению к постам позволит улучшить качество общения в социальных сетях. Результаты исследования могут быть использованы при разработке лекционных и семинарских курсов по дисциплинам «Методы прикладной лингвистики» и «Квантитативная лингвистика и компьютерные технологии».

Обсуждение и результаты

Высокая технологичность современного общества оказывает непосредственное воздействие на качество коммуникации, позволяя переносить речевое взаимодействие из реальной жизни в виртуальную. Межличностное общение, погруженное в речевую ситуацию в пределах интернет-пространства, а также его процесс и одновременно результат носят название интернет-дискурса (Юртаева, 2016).

Несмотря на возможности, которые предоставляет Интернет, пользователи социальных сетей сталкиваются с уязвимостью личной информации, социальным давлением в соцсетях, вытеснением реального общения виртуальным, девиантным поведением и большими объемами нерелевантной информации. Одной из актуальных тем для исследования является «токсичное» поведение пользователей. Под «токсичным» контентом подразумевается информация, оказывающая деструктивное психологическое воздействие на личность, социальные группы и общество в целом (Сундиев, Смирнов, 2020). Основным критерием «токсичных» отношений считается дискомфорт, который вызван общением с тем или иным человеком. В работе Л. Н. Овиновой, Е. Г. Шрайбер «токсичное» общение трактуется как «общение, которое отравляет участников, подавляя их коммуникативную энергию, и создает негативный эмоциональный фон» (2022, с. 37). О «вредности» для окружающих говорится и в статье С. В. Ионовой: «…понятие токсичности используется для обозначения негативных реалий социального взаимодействия, межличностной и профессиональной коммуникации для оценки их как крайне опасных и вредных для окружающих» (2018, с. 2). В рамках социального взаимодействия подчеркивается разрушительное действие на личность, затяжное, перманентное воздействие на жертву (Буряковская, Дмитриева, 2022).

В последнее время понятие «токсичность» ассоциируется с конкретной личностью, по своим характеристикам напоминающей фигуру манипулятора. По мнению Л. Н. Овиновой, Е. Г. Шрайбер (2022), «токсичность» как коммуникативная черта имеет в своей основе сознательную направленность у говорящего на возбуждение у адресата негативных эмоций в процессе общения. «Токсичный» человек-манипулятор стремится контролировать других, не слышит собеседника и пренебрегает его чувствами, легко становится агрессивным, всегда считает себя правым. В виртуальной среде противостоять такому человеку сложнее, чем в реальной жизни, поскольку отсутствует прямой контакт с ним. Возможность выявления «токсичных» сообщений и «токсичных» пользователей представляет собой актуальную задачу для современных исследователей.

С целью исследования «токсичности» и поиска решений для выявления «токсичных» сообщений был использован следующий алгоритм действий: собрать и проанализировать выборку примеров из социальной сети, где в каждом случае представлен пост и набор комментариев к нему; определить языковые особенности «токсичного» общения; рассмотреть существующие решения автоматического выявления токсичных сообщений; опробовать один из методов определения «токсичности» сообщений.

Рассмотрим пример коммуникативной ситуации в социальной сети «ВКонтакте», где можно наблюдать «токсичное» поведение пользователей (Пост сообщества «Авторадио». 2024. https://m.vk.com/wall-383476_773553):

Пост: А вам легко отказаться от мобильного телефона?

Пользователь 1: Легко, я и интернетом вообще не пользуюсь, даже сейчас, *надиктовываю через кофеварку* 🧐

Пользователь 2: Вот я сейчас сижу в Египте люди прилетели за пять за 10.000 км вышли в лобби-бар и сидят вдвоём в телефонах Спрашивается *Нахрена сюда приехали сидеть в телефонах*

Пользователь 3: *А вам легко вытирать ж*** трусами или рукой*, а не туалетной бумагой? Кто автор этих "животрепещущих" вопросов?

Пользователь 4: *Вопрос дебильный*, давайте откажемся от авто, стиральных машин, водоканала, отопления централизорованого (здесь и далее орфография и пунктуация авторов сохранены).

В примере пользователи не отвечают серьезно на вопрос из поста, предпочитая иронию, встречные вопросы, критические замечания. Такое поведение типично для «токсичных» пользователей: желание вступить в спор часто преобладает над стремлением к продуктивной коммуникации. Так, например, в комментариях под другим постом два пользователя обмениваются оскорблениями, которые никак не связаны с объявлением о выходе руководства по новой игре (Пост сообщества “Rosetta | Destiny 2”. 2019. https://vk.com/wall-100460387_231359):

Пользователь 2: ясно, я понял, ты *недоразвитый*, прости куда деньги на лечение скидывать?

Пользователь 1: что тебе от меня надо, *чучело*?

Пользователь 2: вот это *ты токсичный*.

Проанализированная выборка (32 поста, 118 комментариев, из них 24 токсичных) позволяет составить речевой портрет «токсичного» пользователя, которому свойственно:

– применение лексем с отрицательной коннотацией (*плохая*, *позор*);
– избегание основной темы беседы, использование оскорбительной лексики (*чучело*, *недоразвитый*);
– иронизирование (*интернетом вообще не пользуюсь*, *надиктовываю через кофеварку*);
– обвинение в «токсичности» других пользователей (*вот это ты токсичный*);
– использование сленга (*рофлю*, *kappa*);
– употребление медиаинструментов для выражения критики и принижения других пользователей (эмодзи, оскорбительные картинки).

Интересно употребление слова *kappa*: первоначально так назывался смайл, представляющий собой черно-белое ехидное лицо, используемое для демонстрации саркастического отношения к обсуждаемому вопросу на стриминговом сервисе «Твич». В настоящее время пользователи часто просто пишут название смайла, если хотят передать, что текст написан с сарказмом (употребление мета-инструментов свойственно интернет-дискурсу вообще).

Сложность задачи выявления «токсичных» комментариев в рамках русскоязычного исследования заключается в отсутствии надлежащего датасета, который содержал бы достаточное количество примеров «токсичных» и адекватных комментариев. Интерес представляет Russian Language Toxic Comments Dataset (https://www.kaggle.com/datasets/blackmoon/russian-language-toxic-comments) – коллекция аннотированных комментариев с сайтов «Двач» и «Пикабу». Несмотря на то, что датасет не содержит пояснений о процессе аннотирования, в статье (Smetanin, 2020) предоставляются достоверные результаты работы моделей глубокого обучения на Russian Language Toxic Comments Dataset.

Большая часть рассмотренных в рамках исследования работ предполагает внедрение глубокого обучения и, следовательно, наличие большого набора размеченных данных для тренировки модели. В рамках настоящей работы поднимается вопрос о вероятности получения адекватных результатов при использовании базовых алгоритмов машинного обучения.

Так, например, авторы исследования (Andrusyak, Rimel, Kern, 2018) используют исходный словарь оскорбительных терминов, который дополняется при помощи итеративного неконтролируемого присвоения меток (оскорбительных или не оскорбительных) комментариям в социальных сетях. Несмотря на применение базовых инструментов предобработки, анализ результатов показывает, что мера полноты и F-мера превышают 0,6, а максимальный показатель меры точности – 0,87.

Особый интерес для текущего исследования представляет работа (Bakarov, Gureenkova, 2017), целью которой является автоматическое выявление нерелевантных постов имиджборда «Двач». В статье предлагается определять релевантность сообщений путем вычисления семантической близости самого поста-комментария и стартового поста. Задача решается путем обучения классификатора на основе дистрибутивных векторных моделей, в работе сравниваются результаты обучения 7 моделей (Word2Vec, Glove, Word2Vec-f, Wang2Vec, AdaGram, FastText, Swivel), оценивается работа полученных векторных вложений слов на датасете, собранном на платформе «Двач», сравнивается их эффективность касательно задачи автоматического выявления нерелевантных комментариев. Работа является новаторской для проблемы автоматического выявления нерелевантных сообщений в русскоязычных социальных сетях, а полученные результаты указывают на эффективность изложенного метода.

Анализ отобранных примеров «токсичного» поведения выделяет смысловое несоответствие с постом как один из критериев принадлежности комментария к «токсичным». В связи с этим выдвинута гипотеза: выявить «токсичные» комментарии и, следовательно, «токсичного» пользователя можно при помощи критерия семантической близости. Если между постом и комментарием семантическая близость достаточно высокая, значит, комментарий релевантен. Если критерий семантической близости ниже порогового значения, значит, комментарий может содержать признаки «токсичности».

Безусловно, не все комментарии, отобранные таким образом, будут действительно «токсичными». Возможны ситуации, когда комментарий может отклониться от темы, например в случае, если оставлен по ошибке под другим постом или пользователь сам изменил тему. Также критерий семантической близости может быть неэффективным способом, если комментарий содержит вставки на языках, отличных от языка поста. Тем не менее в рамках нынешнего исследования предполагается, что пользователи редко по ошибке оставляют комментарии не под тем постом, как и то, что в русскоязычной социальной сети превалирующим в общении является русский язык. Основная задача работы – проверка действенности подхода, основанного на семантической близости, для проведения первого, базового отбора комментариев, претендующих на то, чтобы называться «токсичными».

Экспериментальный дизайн выглядит следующим образом:

1) создание и предобработка корпуса постов и комментариев;
2) выбор экспериментальной модели;
3) создание и разметка валидационной выборки для проверки работы модели;
4) проведение эксперимента по определению семантической близости;
5) оценка полученных результатов.

В качестве платформы для создания корпуса была выбрана социальная сеть «ВКонтакте», популярная среди русскоязычных интернет-пользователей. При помощи сервисного ключа доступа было создано приложение в социальной сети. Далее при помощи методов wall.get и wall.getComments были собраны последние 50 постов и комментарии к ним из следующих сообществ: «РИА Новости», «Подслушано Уфа», «AstroAlert | Наблюдательная астрономия», «Пикабу», «Телеканал ТНТ», "IGM". Для этого также был использован метод groups.getById, передающий ID сообщества. Полученные посты и комментарии были последовательно записаны в базу данных PostgreSQL, для упрощения доступа к которой позже использовалось объектно-реляционное отображение Peewee (устанавливается разработчиками Python как обычная библиотека).

Для приведения текстового массива в цифровой была выбрана концепция дистрибутивных векторных моделей или эмбедингов (word embeddings), для ее реализации в рамках исследования требуется метод, который можно внедрять на больших объемах данных. В качестве экспериментального метода выбрана word2vec – искусственная нейронная сеть, которая обрабатывает текст, преобразуя его в числовые векторизованные слова. Word2vec преобразовывает большой текстовой корпус в пространство векторов, где каждое уникальное слово в корпусе представлено вектором из сгенерированного пространства. Векторы слов расположены в пространстве векторов таким образом, что близкие по значению слова располагаются в непосредственной близости друг от друга (модель фиксирует синтаксическое и семантическое сходство между словами).

В качестве предобработки для работы с эмбедингами проводятся следующие действия:

1) токенизация и лемматизация теста: морфологический анализатор Pymystem3;
2) убираются стоп-слова: NLTK, модуль stopwords;
3) избавление от шума: регулярные уравнения;
4) перевод в нижний регистр: PyCharm (https://www.jetbrains.com/pycharm/), метод lower();
5) добавление частеречного тега для каждого слова: морфологический анализатор Pymystem3.

После предобработки на вход модели можно подать лемматизированный текст без шума с частеречными тегами.

Среди способов вычисления семантической близости было выбрано определение косинусной близости, которую можно подсчитать и для обычного текста. Однако данная мера не учитывает ни синонимы, ни грамматические формы, ни порядок слов, а только лишь наличие определенных токенов в самом предложении. Так, косинусная близость между предложениями «Я купил кролика в Петербурге» и «Я выступил перед публикой в Петербурге» будет достаточно высокой при довольно разном смысле предложений.

Векторные вложения слов содержат контексты, в которых встречается конкретное слово, их использование позволяет вычислить косинусную близость наиболее точно. Для получения вектора предложения усредняются векторы всех слов предложения. Затем можно найти косинусную близость между векторами поста и комментария.

В качестве модели для работы с эмбедингами была выбрана gensim, в которой реализованы алгоритмы word2vec. С сайта RusVectores (https://rusvectores.org/ru/models/) были скачаны три дистрибутивных векторных модели (размерность векторов моделей – 300):

1) Araneum, обученная на корпусе 10 млрд слов;
2) векторная модель, обученная на корпусе русскоязычных новостей размером 2,6 млрд слов;
3) дистрибутивная векторная модель, обученная на Web-корпусе объемом 900 млн слов.

Косинусная близость измеряется числом от 0 до 1: чем ближе значение близости к 1, тем ближе семантически значения двух предложений. Для эксперимента в качестве порогового значения релевантности была выбрана косинусная близость 0,5. Для вычисления косинусной близости была выбрана библиотека Scikit-learn – один из наиболее используемых пакетов Python для машинного обучения. Метод metrics.pairwise.cosine_similarity позволяет вычислять косинусную близость между двумя векторами.

Для проверки гипотезы из созданного ранее корпуса вручную были отобраны в подкорпус 448 пар «пост – комментарий», на основании субъективной оценки комментариям приписана релевантность (релевантен/нерелевантен). Затем для каждой пары «пост – комментарий» вычислена косинусная близость и автоматически проставлена предполагаемая метка релевантности (0 – если косинусная близость меньше 0,5, 1 – если косинусная близость больше или равна 0,5).

При предварительной проверке модели, обученной на корпусе русскоязычных новостей, был получен результат, удовлетворяющий целям исследования, затем был проведен пробный эксперимент, где результаты были скорректированы с помощью весов TF-IDF – статистической меры, используемой для определения важности слова в контексте документа как части коллекции (корпуса). При вычислении меры TF-IDF в Python с помощью модуля tfidfvectorizer библиотеки Scikit-learn на основе собранного корпуса была построена матрица, которая накладывается на предобученные векторные вложения слов и корректирует их работу в соответствии с новыми весами. Так, векторные вложения слов подстраиваются под работу экспериментального корпуса. В ходе эксперимента для работы с эмбедингами были использованы все три модели, действие которых проверялось для всех 448 пар «пост – комментарий».

Для анализа эффективности моделей полученные результаты сравнивались с оригинальными данными и определялся процент правильных или неправильных выводов:

1) True positive (TP) – если программа правильно определила релевантный комментарий;
2) True negative (TN) – если программа правильно определила нерелевантный комментарий;
3) False positive (FP) – если программа ошибочно приняла релевантный комментарий за нерелевантный;
4) False negative (FN) – если программа ошибочно приняла нерелевантный комментарий за релевантный.

Расчет матрицы ошибок выполнялся при помощи модуля metrics библиотеки Scikit-learn. Так как в работе проводилась бинарная классификация, использовалась функция confusion_matrix(), предназначенная для задач с двумя классами.

Матрица ошибок представляет четыре индивидуальных показателя, на основании которых можно рассчитать другие метрики. Дополнительными метриками оценки эффективности алгоритма стали:

– Precision (точность) – доля документов, действительно принадлежащих данному классу, вычисляется при помощи функции precision_score() модуля sklearn.metrics.

– Recall (полнота) – доля найденных классификатором документов, принадлежащих классу, относительно всех документов этого класса в тестовой выборке, вычисляется при помощи функции repl_score() модуля sklearn.metrics.

– F1 (F-мера) – гармоническое среднее между полнотой и точностью, которое придает одинаковый вес обеим метрикам, вычисляется при помощи функции f1_score() модуля sklearn.metrics.

– Accuracy – доля документов, по которым классификатор дал правильное решение, вычисляется при помощи функции precision_score() модуля sklearn.metrics.

Итогом анализа стали следующие предварительные результаты (Табл. 1).

Таблица 1. *Оценка эффективности алгоритма*

	araneum	**web**	**news**
True positive	51,12	17,90	47,71
True negative	18,08	21,92	18,66
False positive	4,02	0,00	1,61
False negative	26,79	60,18	32,02
Precision	**0,93**	**1,00**	**0,97**
Recall	**0,66**	**0,23**	**0,60**
F1	**0,77**	**0,37**	**0,74**
Accuracy	**0,69196**	**0,398206018**	**0,6637**

Эксперимент можно считать успешно пройденным для двух моделей: Araneum и дистрибутивной векторной модели, предобученной на новостном корпусе. Модель, обученная на Web-корпусе, не прошла пороговое значение эксперимента. Возможно, это связано с качеством сборки корпуса или слишком большим разнообразием жанров.

На основании первых результатов исследования можно утверждать, что внедрение предобученных векторных моделей эффективно с точки зрения экономии ресурсов и получения качественного результата. Также мера семантической близости может использоваться как один из критериев выявления «токсичности» комментария по отношению к посту, однако при комплексном исследовании проблемы следует учесть и действенность других мер.

Анализ ошибок первого (False positive) и второго (False negative) ряда позволил определить наиболее слабые места алгоритма, основанного на определении семантической близости. Рассмотрим посты и комментарии, где наблюдаются оба типа ошибок.

Пост 1: Концерта Канье Уэста в Лужниках не будет. Информацию о подготовке шоу опровергли в концертном агентстве TCI (Пост сообщества «РИА Новости». 2024. https://vk.com/ria?w=wall-15755094_45458277).

Комментарий: Мне кажется в российской эстраде своих дибило В хватает (здесь причиной ошибки первого ряда могло стать ругательство, написанное с опечаткой).

Комментарий: Да и уэст с ним… Других насущных новостей у РИА Новости, нет…? (комментарий выражает недовольство автора подборкой новостей и не имеет почти ничего общего с постом, однако использование фамилии заявленного в посте исполнителя, хоть и написанной с маленькой буквы, дало основание алгоритму определить комментарий как релевантный).

Пост 2: Хорошие новости для тех кто пропустил прошлые сезоны, начиная с этой неделе можно лутать пушки с рамками в сезонных активностях (Пост сообщества “Rosetta | Destiny 2”. 2024. https://vk.com/wall-100460387_664037).

Комментарий: А зачем, если у вендоров можно получить все эти ганы?

Комментарий: Народ хотелось бы спросить стоит ли возвращаться в дестени ради порции сюжета с новым длц (оба комментария по теме игры были приняты за нерелевантные, потому что пользователи пользуются сленгом, который не определяется алгоритмом).

Пост 3: Запад завидует крепкому союзу России и Белоруссии, поэтому хочет вбить клин между государствами. Об этом в интервью РИА Новости заявил глава белорусского МИД Сергей Алейник (Пост сообщества «РИА Новости». 2024. https://vk.com/wall-15755094_45456180).

Комментарий: Пока Батька у руля, миру быть (здесь релевантный комментарий был ошибочно воспринят как нерелевантный, так как алгоритму неизвестно, что Батька – это распространенное прозвище президента А. Г. Лукашенко, от белор. «Бацька»).

Комментарий: А то… «разделяй и властвуй»….старо как мир (использование цитаты также сработало в пользу ошибки первого ряда).

Пост 4: Бывший завод Toyota в Санкт-Петербурге передали автопроизводителю Aurus. Об этом сообщил Денис Мантуров. Предприятие будет выпускать автомобили бизнес-класса. Первые экземпляры ожидаются в конце года. На конвейер встанут четыре модели (Пост сообщества «РИА Новости». 2024. https://vk.com/wall-15755094_45452530).

Комментарий: Отлично. Надеюсь Аурус выпустит повседневную машину....В районе 10 миллионов (нейтральное на вид предложение содержит явную иронию, но определяется как нерелевантное из-за того, что транслитерированное название не сходится с упомянутой в посте маркой автомобиля).

На основании рассмотренных примеров можно предположить, что причинами ошибок первого ряда являются: использование нецензурных слов и цитат, сленга и вставок на других языках, а также упоминание отдельных культурных реалий. Ошибки второго ряда появляются в случае, когда комментарий токсичен (например, содержит иронию), однако автор использует формулировки, близкие по значению с исходным постом, поэтому семантическая близость предложений оказывается высокой.

Заключение

«Токсичность» пользователей социальных сетей представляет серьезную проблему для современной онлайн-коммуникации. Популярность социальных сетей обуславливает необходимость разработки алгоритмов по выявлению нерелевантных сообщений, что определяет актуальность проведенного исследования.

Целью работы было определение действенности алгоритма выявления «токсичных» сообщений на основании семантической близости. Проведенный анализ соответствующей терминологии и актуальных работ по теме дал представление об основных критериях «токсичности» и методах выявления токсичных сообщений. Изучение примеров токсичного поведения позволило выделить смысловое несоответствие с постом как один из критериев принадлежности комментария к «токсичным», в качестве меры оценки семантической близости выбрана косинусная близость.

К выводам исследования можно отнести эффективность использования предобученных дистрибутивных векторных моделей для задач бинарной классификации. Более того, вычисленная при помощи программных средств косинусная близость сделала возможным проведение первого, базового отбора комментариев, претендующих на то, чтобы называться «токсичными». Расчет матрицы ошибок и других метрик эффективности алгоритма свидетельствует об эффективности данного метода для достижения целей эксперимента: метрика Accuracy для двух моделей из трех составила 0,69 и 0,66 соответственно. Работа с третьей дистрибутивной векторной моделью не была настолько эффективной: доля Accuracy 0,39 не является удовлетворительной для эксперимента.

Анализ ошибок алгоритма демонстрирует необходимость его доработки, примером может стать создание дополнительных «фильтров», которые позволили бы избегать ошибок первого и второго рода, вызванных использованием нецензурных слов и цитат, сленга и вставок на других языках.

В качестве перспектив дальнейшего исследования заявленной проблематики можно назвать как использование других дистрибутивных векторных моделей и меры семантической близости, так и выбор нового экспериментального дизайна. На основании выделенных стилистических особенностей можно предложить другие механизмы выявления токсичных пользователей, например с применением методов сентимент-анализа, что и станет материалом для дальнейших исследований по этой теме.

Источники | References

1. Арутюнова Н. Д. Дискурс // Лингвистический энциклопедический словарь / отв. ред. В. Н. Ярцева. М.: СЭ, 1990.
2. Буряковская В. А., Дмитриева О. А. Квазинаучный термин «токсичный» в современной блогосфере (на материале русского, английского и французского языков) // Известия Волгоградского государственного педагогического университета. 2022. № 5 (168).
3. Галичкина Е. Н. Специфика компьютерного дискурса на английском и русском языках: на материале жанра компьютерных конференций: дисс. … к. филол. н. Астрахань, 2001.
4. Грибовод Е. Г. Дискурс // Дискурс-Пи. 2013. Т. 10. № 3.
5. Ефанова А. А., Осокин А. А. Дискурс социальных медиа: к проблеме интерпретации // Вопросы теории и практики журналистики. 2022. Т. 11. № 3.
6. Ионова С. В. Токсичный руководитель: лингвоэкология речевого поведения // Экология языка и коммуникативная практика. 2018. № 4.
7. Карасик В. И. Жанры сетевого дискурса // Жанры речи. 2019. № 1 (21).
8. Красных В. В. Этнопсихолингвистика и лингвокультурология: курс лекций. М.: Гнозис, 2002.
9. Лутовинова О. В. Лингвокультурологические характеристики виртуального дискурса. Волгоград: ВГПУ; Перемена, 2009.
10. Овинова Л. Н., Шрайбер Е. Г. «Токсичное» педагогическое общение: анализ состояния, причины и признаки // Вестник Южно-Уральского государственного университета. Серия: Образование. Педагогические науки. 2022. Т. 14. № 3.
11. Павлов М. А. Понятие сетевого дискурса в современной лингвистике // Наука и образование: новое время. 2017. № 1.

12. Платонов Е. Н., Руденко В. Ю. Выявление и классификация токсичных высказываний методами машинного обучения // Моделирование и анализ данных. 2022. Т. 12. № 1.
13. Русанов Е. К. Интернет-дискурс в дискурсивной парадигме // Гуманитарные юридические исследования. 2016. № 1.
14. Рябова А. С. Лингвистические особенности англоязычного дискурса социальных сетей // Огарёв-Online. 2020. № 6 (143)
15. Сундиев И. Ю., Смирнов А. А. «Токсичный» контент в сети Интернет и его влияние на радикализацию молодежи // Научный портал МВД России. 2020. № 4 (52).
16. Ушаков А. А. Интернет-дискурс как особый тип речи // Вестник Адыгейского государственного университета. Серия 2: Филология и искусствоведение. 2010. № 4.
17. Юртаева Е. С. Характеристики виртуальной языковой личности в коммуникативном пространстве Интернет-дискурса // Иностранные языки в контексте межкультурной коммуникации: материалы докладов VIII международной конференции. Саратов, 2016.
18. Aken B. van, Risch J., Krestel R., Löser A. Challenges for Toxic Comment Classification: An In-Depth Error Analysis // Proceedings of the 2nd Workshop on Abusive Language Online (ALW2) / ed. by D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, J. Wernimont. Brussels, 2018. https://doi.org/10.18653/v1/W18-5105
19. Andrusyak B., Rimel M., Kern R. Detection of Abusive Speech for Mixed Sociolects of Russian and Ukrainian Languages // Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018. Karlova Studánka, 2018.
20. Bakarov A., Gureenkova O. Automated Detection of Non-Relevant Posts on the Russian Imageboard “2ch”: Importance of the Choice of Word Representations // Analysis of Images, Social Networks and Texts. AIST 2017 / ed. by W. M. P. van der Aalst, D. I. Ignatov, M. Khachay, S. O. Kuznetsov, V. Lempitsky, I. A. Lomazova, N. Loukachevitch, A. Napoli, A. Panchenko, P. M. Pardalos, A. V. Savchenko, S. Wasserman. Cham: Springer, 2017. https://doi.org/10.1007/978-3-319-73013-4_2
21. Hao L., Weiguan M., Hanyan L. Toxic Comment Detection and Classification. 2018. https://cs229.stanford.edu/proj2019spr/report/71.pdf
22. Khieu K., Narwal N. Detecting and Classifying Toxic Comments. 2019. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6837517.pdf
23. Risch J., Krestel R. Toxic Comment Detection in Online Discussions // Deep Learning-Based Approaches for Sentiment Analysis / ed. by Dr. B. Agarwal, Dr. R. Nayak, Dr. N. Mittal, Prof. S. Patnaik. Singapore: Springer, 2020.
24. Smetanin S. Toxic Comments Detection in Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2020” (Moscow, June 17-20). Moscow, 2020.

Информация об авторах | Author information

RU

Курганская Екатерина Владимировна[1]
Степанова Наталия Валентиновна[2], к. филол. н., доц.
[1, 2] Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им В. И. Ульянова (Ленина)

EN

Ekaterina Vladimirovna Kurganskaia[1]
Natalia Valentinovna Stepanova[2], PhD
[1, 2] Saint Petersburg Electrotechnical University “LETI”

[1] *katrinkurg26@gmail.com,* [2] *nathalie.tresjolie@icloud.com*

Информация о статье | About this article

Дата поступления рукописи (received): 26.02.2024; опубликовано online (published online): 27.05.2024.

Ключевые слова (keywords): токсичность в социальных сетях; релевантность комментариев; семантическая близость; векторные вложения слов; toxicity in social networks; relevance of comments; semantic proximity; word vector embeddings.