DOI 10.36622/VSTU.2022.25.3.010
УДК 004.048

АВТОМАТИЗИРОВАННАЯ БАЗА ДАННЫХ ДЕСТРУКТИВНЫХ КОНТЕНТОВ: МЕТОДИЧЕСКОЕ И МАТЕМАТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ТЕКСТОВОГО АНАЛИЗАТОРА

Е.Ю. Чапурин, Н.М. Лантюхов, О.Ю. Макаров

В статье рассматривается создание автоматизированной базы данных деструктивного контента с возможностью тонального анализа текста в мессенджере «Telegram» с учетом структурно-функционального и методического обеспечения. Разработан алгоритм анализа контента с текстовым содержанием, в котором решение о деструктивности выносится непосредственно после глубокого анализа текста. На основе полученного алгоритма было разработано программное обеспечение, которое позволяет не просто добавлять контент, по ключевым словам, заранее предопределенным в базе данных, а именно с использованием тонального анализа текста. Разработанное методическое, алгоритмическое и программное обеспечение актуализации базы данных может быть применено для анализа различных мессенджеров, таких как «WhatsApp» и «Viber», а математическое обеспечение позволит произвести усовершенствование системы на основе полученных метрик.

Ключевые слова: база данных, деструктивный контент, социальные сети.

Введение

Исходя из последнего анализа ВЦИОМ в 2022 году можно сделать вывод, что наиболее популярными платформами для общения являются: «WhatsApp» назвали 87 % россиян, «YouTube» – 75 %, «ВКонтакте» – 62%, «Telegram» – 55%, а «Одноклассники» – 42% респондентов [1]. Однако «WhatsApp» и «Telegram» являются мессенджерами, первый не подлежит анализу на наличие деструктивного контента, так как нет общедоступных групп, в которых могли бы состоять пользователи. Актуальными и наиболее интересными для исследования представляются социальная сеть «ВКонтакте» и мессенджер «Telegram», ибо именно у них наблюдается наибольший прирост аудитории, причем достаточно юной, восприимчивой к противоречивому контенту.

Регулирование противоправного, деструктивного контента является первостепенной задачей в информационном пространстве, важно защитить население от влияния фейковой и «токсичной» информации. Важность законодательного регулирования поддерживается и в Общественной палате, где эксперты, сенаторы и депутаты 17 февраля 2022 года дали старт созданию реестра деструктивного контента для защиты детей, а один из авторов проекта сообщил, что работа с Роскомнадзором в этом направлении уже ведется и перечень тем, подлежащих блокировке, будет только расширен. Действительно, на данный момент деструктивной информацией, подлежащей немедленному удалению, являются только остросоциальные проблемы – терроризм, призыв к суициду, пропаганда наркотиков, однако на уровне закона формально законным является контент о романтизации насилия, извращений, ненависти к семье, поощрение шок-контенту.

Большинство средств, направленных на обнаружение и фильтрацию деструктивного контента, работают лишь поверхностно – они проводят лишь поверхностный анализ – по ключевым словам, которые есть в базе данных. В данном случае вероятность ошибочного распознавания контента очень высока, ведь не учитывается контекстная окраска текста. Поэтому целесообразно создание и использование методического и программного обеспечения для тонального анализа текста.

Тональный анализ текста, а в особенности русскоязычного текста, является довольно сложной и трудоемкой задачей в рамках машинного обучения, ведь сам по себе русский язык очень сложный и многогранный в отличии от, например, английского языка. В русском языке слишком

много подтекстов, интонирования и прочего, а также само строение языка и его грамматика сильно отличаются от «логического» английского с его короткими словами и предложениями, понятной искусственному интеллекту логикой построения и считывания смыслов. Большинство созданных приложений, систем глубокого анализа текста создано также в большей части для англоязычной аудитории, исследованию русского языка ученые уделяли гораздо меньше внимания, хотя есть ряд факторов, благодаря которым актуальность таковых исследований становится очевидной:

– русский язык располагается на восьмом месте по количеству носителей во всем мире;

– согласно опросу, проведенным компанией Omnibus GFK, распространение «Интернета» в России больше 75%, из них 92 миллиона – это россияне, чей возраст старше 16 лет;

– русскоязычные сайты распространены на всех континентах, но наибольшая концентрация приходится на Содружество Независимых Государств (СНГ) и, особенно, на Россию и Украину;

– русский язык является вторым по распространенности языком в «Интернете» после английского. По состоянию на апрель 2020 года 8,6 % из 10 миллионов самых популярных интернет-сайтов в мире используют русский язык [2-4].

Таким образом, анализ текстового контента для русскоязычной аудитории является актуальной и слабоизученной задачей.

Выбор средств и способов, используемых для создания автоматизированной базы данных деструктивных контентов

Чтобы приступить к непосредственному методическому обеспечению программных средств работы систем управления базами данных, стоит изначально выбрать: база данных будет реляционная или же нереляционная. Особенности выбора будут описаны далее. К нереляционным в данном случае относятся абсолютно все виды баз данных, которые отличны от реляционных, например, графические базы данных и базы данных временных рядов, которые на данный момент выделены в отдельные классы.

Реляционная система управления базами данных – это табличный набор программ и возможностей, который обеспечивает интерфейс между пользователями, приложениями и базой данных, предлагая систематический способ создания, обновления, удаления, управления и извлечения данных. Большинство систем управления реляционными базами данных используют язык программирования SQL для доступа к базе данных. Большинство баз данных соблюдают свойства ACID (Atomicity – атомарность, Consistency – согласованность, Isolation – изоляция, Durability – долговечность).

Реляционные базы данных зачастую хорошо подходят для любых данных, которые являются регулярными, предсказуемыми и выигрывают от способности гибко составлять информацию в различных форматах. Поскольку реляционные базы данных работают по схеме, то возникают проблемы с изменением структуры данных после того, как они были загружены систему. Тем не менее, схема также помогает обеспечить целостность данных, убедившись, что значения соответствуют ожидаемым форматам, и что необходимая информация включена. В целом, реляционные базы данных являются надежным выбором для многих приложений, потому что приложения часто генерируют хорошо упорядоченные структурированные данные.

Положительные стороны реляционных баз данных:

– отлично подходит для структурированных данных;

– использование существующего языка запросов (SQL);

– отлично подходит для сложных запросов;

– простая навигация по данным;

– высокий уровень интеграции данных благодаря связям и ограничениям между таблицами;

– транзакции безопасны;

– высокая надежность.

Отрицательные стороны реляционных баз данных:

– предварительное определение схемы;

– нет адаптации к меняющимся требованиям: динамические изменения элемента влияют на все остальные элементы.

Рассмотрим современные альтернативы для данных, которые не соответствуют реляционной парадигме – нереляционные базы данных (NoSQL).

Базы данных «NoSQL» появились в эпоху мэйнфреймов и экспоненциального развития веб-приложений. Когда стоимость хранения резко снизилась, возникла необходимость в создании сложной модели данных для уменьшения дублирования данных. Разработчики были основной стоимостью разработки программного обеспечения, поэтому базы данных NoSQL оптимизированы для производительности разработчиков.

Базы данных «NoSQL», также иногда называемые нереляционными или не только базами данных SQL, представляют собой широкую категорию, которая охватывает любой тип систем баз данных, которые отклоняются от общей модели реляционных баз данных. Хотя нереляционные базы данных уже давно доступны, эта категория обычно используется для обозначения новых поколений баз данных, использующих альтернативные модели, такие как хранилища ключ-значение, ориентированные на документы, графы и семейства столбцов. Термин «NoSQL» является несколько неправильным, поскольку базы данных в этой категории являются скорее реакцией против реляционного архетипа, а не языка запросов SQL. Говорят, что «NoSQL» означает либо «non-SQL», либо «не только SQL».

Существует несколько основных подтипов в категории нереляционных или «NoSQL» баз данных, два из которых включают:

– базы данных, ориентированные на документы, именованные строковые поля связаны со значениями объектных данных в структуре данных, известной как «документ», который может быть закодирован с помощью таких технологий, как JSON, XML, YAML или BSON. Документы не должны поддерживать идентичные структуры, что обеспечивает очень высокую гибкость;

– хранилища ключ-значений: очень простая форма базы данных «NoSQL», которая, как следует из названия, использует структуру, в которой полные документы хранятся в виде значений, соответствующих уникальному ключу, который обычно представляет собой хэшированную строку. В этом типе систем доступ к документам возможен только через их уникальный ключ, что обеспечивает очень быстрый поиск.

Подводя итог, нереляционные базы данных хранят данные не в табличной форме, а как объекты с произвольными атрибутами: это могут быть пары «ключ-значение», документы в формате JSON, графы и так далее.

Положительные стороны нереляционных баз данных:

– гибкая модель данных;

– быстрая адаптация к меняющимся требованиям: динамические изменения элемента не влияют на другие элементы;

– хранение огромного количества данных с небольшой структурой;

– высокая производительность.

Отрицательные стороны нереляционных баз данных:

– язык запросов производится вручную;

– трудно проверить целостность и согласованность данных.

Итак, для определения подхода, приведем таблицу сравнений реляционной и нереляционной баз данных (табл. 1).

Исходя из табл. 1, для решения поставленных задачи в большей степени подходит реляционная база данных, а именно: данные и их взаимосвязь заранее известны, что в свою очередь улучшает навигацию по данным, а также высокая возможность интеграции с другими сервисами. Поэтому при разработке методического обеспечения будет использоваться именно она. Среди всех реляционных баз данных была выбрана «MySQL», так как она имеет широко проработанный модуль с языком программирования «Python». Для самой базы данных предлагается следующая структура, изображенная на рис. 1.

Таблица 1

Достоинства и недостатки различных типов баз данных

Особенность	Реляционная БД	Нереляционная БД
Структура	Заранее определенная схема	Отсутствие заранее выстроенной схемы
Возможность интеграции с другими сервисами	Высокая	Средняя
Хранение данных	Оптимизирован для больших данных	Оптимизирован для больших данных
Навигация по данным	Простая	Сложная
Надежность	Высокая	Средняя
Масштабируемость	Высокая	Высокая

Рис. 1. Детальная блок-схема обработки запросов в SQL

Создание лингвистического обеспечения

Важно не просто создать базу данных с применением текстового анализатора, но и производить его дальнейшую категоризацию по тематикам.

Для начала необходимо определиться с ключевыми тематиками деструктивного контента, а только потом переходить к набору терминов, ключевых слов, которые могут его идентифицировать. Нижеизложенные тематики деструктивного контента были выбраны на основе анализа пользовательских соглашений социальных сетей и нормативно-правовых актов Российской Федерации:

– разжигание этнической и религиозной ненависти;

– пропаганда экстремизма и терроризма;

– пропаганда наркотических средств;

– пропаганда суицида;

– оправдание и популяризация террористических актов;

– подрыв государственного суверенитета и стабильности;

– пропаганда нетрадиционных отношений;

– пропаганда сепаратизма;

– нецензурная лексика;

– нарушение прав и свобод человека;

– дискредитация вооруженных сил Российской Федерации (далее – ВС РФ).

Список слов, относящихся к каждой из категорий, достаточно широкий, поэтому в качестве примера приведу лишь некоторые тематики деструктивного контента.

Так, например, по тематике дискредитации ВС РФ ключевыми словами для поиска будут: война, Украина, Россия, Донецк, Луганск, Киев, рашисты, русские, украинцы, москали, хохлы, орки, поражение и т.д., а для пропаганды наркотиков: зависимость, сознание, дорожка, алкоголь, шприц, передозировка, легализация, кайф, ломка, игла, притон, доза, наркомания, вещество, ощущение и т.д. Естественно, в дальнейшем будет происходить более глубинный анализ выявленных текстов на основе семантического анализа.

Разработка методического обеспечения

Глубокий анализ текста сводится к анализу тональной окраски. В свою очередь, тональный анализ – процесс обнаружения положительных или отрицательных настроений в тексте. Он часто используется компаниями для обнаружения настроений в социальных данных, оценки репутации бренда и понимания клиентов, а вот в работе специальных служб для выявления

контентов, деструктивной направленности, в социальных сетях и мессенджерах этого нет.

Для решения задач по выявлению деструктивных контентов возможно использовать два алгоритма:

– статистический метод. Для функционирования статистического метода нам необходимы тексты, заранее сформированные по тональной окраске. Корпуса необходимы для обучения модели, благодаря которой определяется тональность текста или отдельной фразы;

– метод, основанный на словарях и правилах. В данной ситуации необходимо составить словари позитивных и негативных слов и словосочетаний. Данный метод имеет возможность применять как шаблоны, так и правила соединения тональной лексики.

В рамках реализации создания методического обеспечения был выбран непосредственно статический метод, который основан на так называемом «обучении с учителем». Главный преимуществом этого метода является то, что он имеет возможность быстрого масштабирования. Таким образом, работа на данном этапе сводится к поиску подходящего датасета.

Исследования по анализу настроений в настоящее время основаны на применении подходов глубокого обучения, что требует обучения и тестирования на специализированных наборах данных. Для английского языка популярные наборы данных для анализа настроений включают: StanfordSentimentTreebankdatasets «SST», набор данных «IMDB» с обзорами фильмов, наборы данных о настроениях в «Twitter» и многие другие. Для других языков было создано гораздо меньше наборов данных. Например, для русского языка собрано не так много датасетов, которые могут быть использованы в анализе настроений. Среди наиболее популярных и распространенных можно выделить следующие: «RuSentiment» [5], «RuBERT-RuSentiment» [6], «RuReviews» [7], «Russian Hotel Reviews Dataset» [8], «SentuRuEval-2016» [9], «RuTweetCorpx» [10] и проект «Natasha», который поддерживается отечественными разработчиками [11].

Рассмотрим одни из самых крупных датасетов в Ру-сегменте и определим наиболее подходящий.

«RuSentiment» – это датасет, созданный для анализа настроений сообщений в социальных сетях на русском языке на основе публикаций в социальной сети «ВКонтакте». «RuSentiment» содержит новый набор всеобъемлющих рекомендаций по аннотированию, которые можно распространить на другие языки. «RuSentiment» в настоящее время является крупнейшим в своем классе для русского языка. Чтобы разнообразить набор данных, было предварительно отобрано 6 950 сообщений с использованием стратегии активного обучения. Данные были собраны с личных страниц пользователей социальных сетей, которые были членами сообществ майдана и антимайдана во время конфликта на Украине в 2014 году. Отсутствие предварительного отбора по темам делает «RuSentiment» в настоящее время крупнейшим вручную аннотируемым набором данных о настроениях в общей предметной области для русского языка, превышающим по размеру только автоматически аннотируемый набор данных «silver от Rubtsova».

Чтобы удалить сообщения, которые не несут информативной ценности, были использованы следующие критерии отбора в датасет: текст должен быть длиной 10-800 символов, не менее 50 % из которых были алфавитными. Были исключены кириллица и URL-адреса. Чтобы обеспечить осмысленность сообщений, были также исключены любые сообщения с более чем 4 хэштегами. «RuSentiment» распространяется без идентификаторов постов в «ВКонтакте» и включает только те посты, которые были опубликованы публично. Средняя скорость аннотирования составляла 250-350 сообщений в час. Проект сделал категоризацию по трехбалльной шкале («негативные», «нейтральные» и «позитивные»). Также был определен класс «пропустить» для исключения сообщений, которые были неясными или не на русском языке. Существенным отличием от других наборов данных, в том числе многих английских, является то, что датасет старается учесть неявные формы выражения внутреннего эмоционального состояния. Однако набор данных больше не доступен из-

за запроса от «ВКонтакте», но при этом существует готовая к использованию модель анализа настроений для русского языка, которая была предварительно обучена на «RuSentiment» [5].

«RuBERT» был обучен на русскоязычной части «Википедии» и новостных данных. Обучающие данные для создания словаря русских подтекстов взяли из многоязычной версии «BERT-base» в качестве инициализации для «RuBERT». Многоязычный «BERT» использовался в качестве инициализации для «SlavicBERT» [6].

«RuReviews» – это датасет, основанный на тональном анализе настроения отзывов в крупнейшем российском Интернет-магазине. Он включает в себя:

– 112 049 отзывов на русском языке;

– 6 756 отзывов на казахском языке;

– 243 необнаруженных языковых обзора (цифры типа «10/10», английский или казахский на латинице).

Количество категорий в датасете – 17. Некоторые из них: парфюмерия, смартфоны, шины, часы и т.д. Формат рейтинга – это целые числа от 1 до 5. Основными языками являются русский, казахский и другие (однако они не несут никакой полезной информации). Спустя 1,5 года, а именно в 2021 году, данный датасет был переработан, т.к. в нем были удалены стоп-слова, используемые в русском и казахском языках [7]. Данный датасет не подходит для реализации поставленных задач по анализу деструктивного контента.

«RussianHotelReviewsDataset». Исходя из названия уже становится очевидным его происхождение – это набор с примерами настроений, который был собран на основе анализа 50 329 отзывов о отелях среди русскоязычных текстов. Следующим шагом была компиляция лексики аспектов и настроений в полученном векторном пространстве. Подход к построению лексики был основан на итеративном расширении небольшого набора первоначально заданных терминов. Наконец, настроение аспектов в реальных отзывах было рассчитано с учетом терминов аспекта и настроения, встречающихся в тексте, и их весов, т.е. косинусного сходства с исходными условиями. Модель была протестирована на корпусе из 6 876 текстов из одного домена. Однако здесь все тот же минус – для оценки отзывов отелей этот датасет может принести пользу, однако деструктивный контент он будет определять гораздо хуже [8].

«SentuRuEval-2016» – датасет результатов тонального анализа текста в русскоязычном «Twitter» о качестве работы крупнейших телекоммуникационных компаний и банков. Набор данных «SemRuEval-2016» содержит 75 000 записей в формате текстового файла [9].

«RuTweetCorp» – это русскоязычный датасет коротких текстов. «RuTweetCorp», который состоит более чем из 17 миллионов записей. Если необходим непосредственно тональный анализ, то существуют классы, которые в автоматическом режиме поделены на две категории: положительные и отрицательные, каждая из категорий содержит порядка 110 тысяч записей. Данные собраны на основе анализа русскоязычных записей в «Twitter» [10].

Алгоритм форматирования датасетов в автоматическом режиме выглядит следующим образом (рис. 2):

Рис. 2. Алгоритм формирования датасета

Подводя итог вышеописанным датасетам можно сделать вывод о том, что большинство доступных наборов данных принадлежат какой-то предметной области. Часть датасетов обучены на текстах из Википедии.

Датасетов, основанных на общении в какой-то социальной сети, очень мало, что нашло отражение в сравнительной таблице (табл. 2).

Таблица 2

Сравнительная таблица русских датасетов

Датасет	Описание	Аннотирование	Классы	Область сбора данных
«RuSentiment»	Набор аннотированных в ручном режиме данных публичных сообщений из социальной сети «ВКонтакте», учитывающий явно и неявно выраженные настроения	Ручное	5	Социальная сеть «ВКонтакте»
«RuBERT-RuSentiment»	Набор данных, собранный в русскоязычной части «Википедии» и новостных данных	Автоматическое	3	«Википедия», новости
«RuReviews»	Набор данных, основанных на тональном анализе настроения отзывов в крупнейшем российском Интернет-магазине	Автоматическое	3	Отзывы на товары
«SentuRuEval-2016»	Набор данных результатов тонального анализа текста в русскоязычном «Twitter» о качестве работы крупнейших телекоммуникационных компаний и банков	Ручное	4	Отзывы в социальной сети «Twitter»
«RuTweetCorp»	Самый крупный корпус текстов, собранный в автоматическом режиме из русскоязычного «Twitter» с частичной ручной фильтрацией данных	Автоматическое	3	Посты в «Twitter»

Исходя из сравнительной таблицы, можно сделать вывод, что наиболее подходящим датасетом для решения поставленных задач является «RuSentiment», а именно:

– один из самых крупных датасетов, сформированных в ручном режиме;

– наиболее удобен для исследования, так как был изначально сформирован на основе открытых данных из социальной сети;

– существуют успешные примеры использования данного датасета в исследованиях, например, при оценке настроений на выборах в 2020 году;

– высокая оценка F1.

Для работы с данным датасетом практичнее всего использовать библиотеку Dostoevsky, которая была успешно апробирована в проекте: «Анализ настроений россиян в отношении конституционных поправок».

Всего существует 5 классов в «RuSentiment» и, как следствие, эти же 5 классов реализованы в библиотеке: позитивный, негативный, нейтральный посыл, речевой акт и неясные случаи, когда программа не знает, к какому классу отнести сообщение или текст.

В рамках реализации функции поиска деструктивного контента с использованием средств интеллектуального анализа текста требуются лишь те посты, которые несут негативное настроение, все остальные не представляют исследовательского интереса.

Перед переходом к самим метрикам необходимо ввести важную концепцию для описания этих метрик в терминах ошибок классификации – confusionmatrix (матрица

ошибок). Матрица ошибок – это таблица, в которой представлены различные прогнозы и результаты тестов и сопоставлены они с реальными значениями. Они используются в статистике, интеллектуальном анализе данных, моделях машинного обучения и других приложениях искусственного интеллекта (ИИ). Сама по себе матрица достаточна проста в понимании, каждое предсказание может быть одним из четырех результатов, основанных на том, насколько оно соответствует фактическому значению. Непосредственному созданию матрицы предшествует небольшое ветвление – прогнозируемые значения описываются как положительные и отрицательные, а фактические значения как истинные и ложные. Схематически это выглядит следующим образом (рис. 3):

Рис. 3. Схема фактических и прогнозируемых значений

Объединив это в более удобный и работоспособный формат данных, получим матрицу, изображенную на рис. 4.

		ФАКТИЧЕСКИЕ ЗНАЧЕНИЯ	
		Исинные (True)	Ложные (False)
ПРОГНОЗИРУЕМЫЕ ЗНАЧЕНИЯ	Положительные (Positive)	TP True Positive	FP (False Positive, Ошибка 1 типа)
	Отрицательные (Negative)	FN (False Negative, Ошибка 2 типа)	TN True Negative

Рис. 4. Матрица ошибок

Таким образом, получаются следующие ячейки:

– истинно положительный результат (TP): предсказанная правда и правда в реальности;

– истинно отрицательный (TN): предсказанная ложь и ложь в реальности;

– ложноположительный результат (FP): предсказанная истина и ложь в реальности;

– ложноотрицательный результат (FN): предсказанная ложь и истина в реальности.

Ошибка типа I: эквивалентно ложным срабатываниям (FP). Первый возможный тип ошибки включает отклонение нулевой гипотезы, которая является истинной.

Ошибка типа II: эквивалентно ложноотрицательным результатам (FN). Другой вид ошибки, который возникает, когда мы принимаем ложную нулевую гипотезу. Такого рода ошибки называются ошибками типа II и также называются ошибками второго рода.

Математическое обеспечение

Базовыми метриками являются точность (Precision) и полнота (Recall). Эти два принципа важны с математической точки зрения в генеративных системах и концептуально важны в ключевых аспектах, связанных с усилиями ИИ имитировать человеческое мышление. Базовым показателем, используемым для оценки модели, часто является точность, описывающая количество правильных прогнозов по отношению ко всем прогнозам:

$$Precision = \frac{True\ positive}{True\ positive + False\ positive}. \quad (1)$$

Когда модель делает много неправильных положительных классификаций или мало правильных положительных классификаций, это увеличивает знаменатель и делает точность малой. С другой стороны, точность высока, когда:

– модель делает много правильных положительных классификаций (максимально истинно положительных);

– модель делает меньше неправильных положительных классификаций (минимизирует ложноположительные результаты).

Фактически, точность (1) отражает, насколько надежна модель при классификации образцов как положительных. Точность помогает, когда затраты на ложные срабатывания высоки. Итак, давайте

предположим, что проблема связана с обнаружением рака кожи. Если у нас есть модель с очень низкой точностью, тогда многим пациентам скажут, что у них меланома, и это будет включать некоторые ошибочные диагнозы. На карту поставлено множество дополнительных тестов и стресса. Когда количество ложных срабатываний слишком велико, те, кто следит за результатами, научатся игнорировать их после бомбардировки ложными тревогами.

Следующая метрика тоже напрашивается сама по себе – полнота (Recall) или чувствительность (Sensitivity). Полнота дает нам истинный положительный показатель (TPR), который измеряет вероятность обнаружения или долю фактических положительных результатов, которые были предсказаны правильно.

$$Recall = \frac{True\ positive}{True\ positive + False\ negative}. \quad (2)$$

Полнота (2) помогает, когда стоимость ложных отрицательных результатов высока. Очень часто используется, когда нужно правильно классифицировать какое-то событие, которое уже произошло. Например, модели обнаружения мошенничества должны иметь высокий отзыв, чтобы правильно обнаруживать мошенничества.

Вышеизложенные метрики являются базовыми и они довольно чувствительны к набору данных и их количеству. Следующей метрикой, объединяющей полноту и точность, является F1.

$$F1 = 2 \times \frac{precision \times reccal}{precision + recall}. \quad (3)$$

Среднее гармоническое из списка чисел (3) сильно смещается в сторону наименьших элементов списка, оно имеет тенденцию (по сравнению со средним арифметическим) смягчать влияние больших выбросов и усугублять влияние малых. Например, точность 0,01 и полнота 1,0 дадут:

– среднее арифметическое:
(0,01 + 1,0) / 2 = 0,505,
– оценка F1 (3):
2*(0.01*1.0)/(0.01+1.0)=~0.02.

Это связано с тем, что оценка F1 гораздо более чувствительна к одному из двух входных данных, имеющих низкое значение (здесь 0.01). Это делает формулу удобной и гибкой в условиях, если необходим баланс между точностью и полнотой.

В идеале оценка F1 может быть эффективной метрикой оценки в следующих сценариях классификации:

– когда FP и FN одинаково дороги, то есть они пропускают истинные положительные результаты или обнаруживают ложные положительные результаты, оба влияют на модель;

– добавление дополнительных данных не приводит к существенному изменению результата;

– TN высок (например, при прогнозировании наводнений, прогнозах рака и т. д.)

Для наглядности построим график зависимости Precision от Recall (рис. 5):

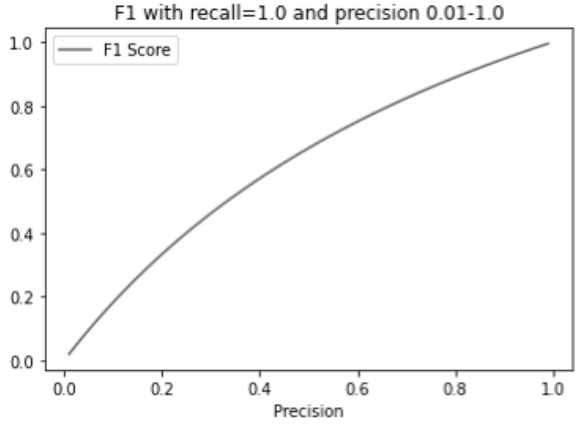

Рис. 5. График зависимости P от R, P=0…1, R=1

Стоит иметь в виду, что F1-мера предполагает одинаковую важность Precision и Recall, если одна из этих метрик для вас приоритетнее, то можно воспользоваться F_β мерой:

$$F_\beta = (\beta^2 + 1)\frac{precision \times reccal}{\beta^2 precision + recall}, \quad (4)$$

Это означает, что оценка (4) равным образом учитывает метрики точности и полноты, не отдавая предпочтения тому или иному значению. Таким образом, метрика F1 актуальна в том случае, когда имеем дело со сбалансированными исходными значениями, в противном случае метрика теряет свой смысл. Однако этого все также недостаточно и необходимо, чтобы учет ошибок

производился для обоих классов. Наряду с этим необходимо ввести еще две метрики, учитывающие ошибки на объектах обоих классов.

TPR (truepositiverate) – полнота, то есть это часть положительных результатов, которые были верно оценены средствами тонального анализа текста как положительные.

$$TPR = \frac{TP}{P} = \frac{TP}{TP+FN}, \quad (5)$$

FPR (falsepositiverate) – в свою очередь это часть отрицательных результатов, которые были ошибочно оценены как положительные.

$$FPR = \frac{FP}{N} = \frac{FP}{FP+TN}, \quad (6)$$

Таким образом, TPR (5) измеряет вероятность обнаружения, которая также называется чувствительностью. В то время как FPR (6) измеряет вероятность ложной тревоги. Для модели классификации нам нужно сбалансировать затраты и выгоды, потому что мы хотим максимизировать TPR при минимизации FPR. Обе эти величины растут при уменьшении порога. Кривая в осях TPR/FPR, которая получается при варьировании порога, исторически называется ROC-кривой (receiveroperatingcharacteristicscurve, сокращённо ROC curve). Самая «идеальная» модель имеет кривую ROC, которая достигает верхнего левого угла (координата (0, 1)) графика: FPR равен нулю, а TPR равен единице.

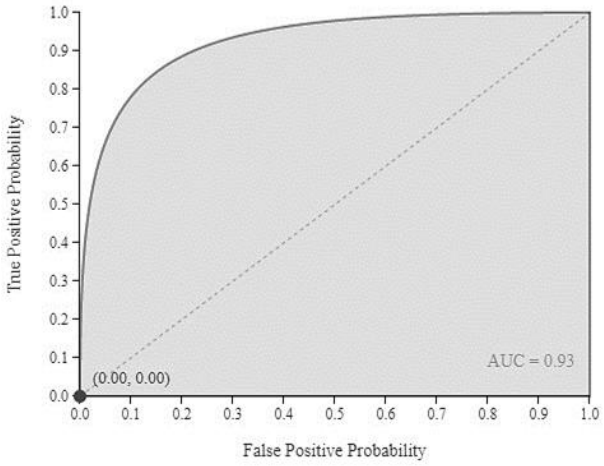

Рис. 7. ROC-кривая

Чем лучше классификатор делит два класса, тем больше значение AUC, то есть площадь под кривой. Именно эта метрика в конечном счете является наиболее однозначной и корректной.

Для наглядности продемонстрируем изменение ROC-кривой (значения AUC) и распределения классов (рис. 6). На практике AUC хорошо работает как общая мера точности прогнозирования.

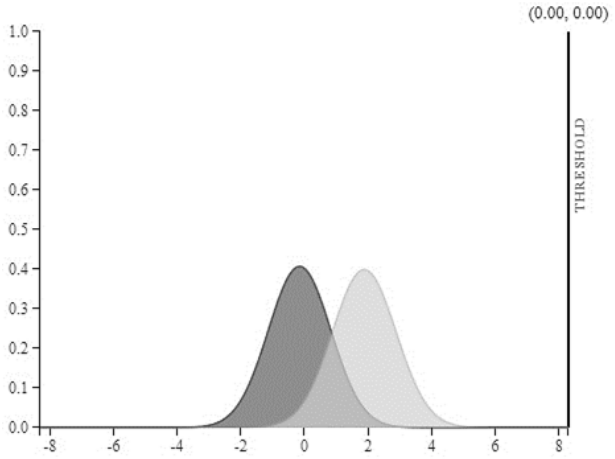

Рис. 6. Распределение классов

На рис. 6 по оси Y отображено пороговое значение, а по оси X распределение предсказаний классификатора. Кривая ROC в общем виде будет выглядеть следующим образом (рис. 7):

Заключение

Полученные результаты могут быть использованы для реализации программно-технической составляющей тонального анализа текста с целью выявления контентов с деструктивным содержанием.

Следующей ступенью в анализе деструктивного контента средствами машинного обучения может стать распознавание деструктива в фото/видео/аудио материалах с применением тонального анализа текста, рассмотренного в рамках данной работы.

Математический аппарат позволяет улучшать качество построенной модели тонального анализа текста путем оценки каждого из параметров. Наиболее эффективной метрикой является значение AUC (площади под кривой ROC), так как она

учитывает и распределение классов, и пороговые значения, да и учет ошибок в данном случае производится для обеих классов.

Список литературы

1. Аудитория Telegram превысила половину всех пользователей рунета. URL: https://habr.com/ru/news/t/580354/ (дата обращения: 22.07.2022).

2. Penetration of Internet in Russia. TheResultsof 2017, 2018. URL: https://www.gfk.com/fileadmin/user_upload/dyna_content/RU/Documents/Press_Releases/2019/GfK_Rus_Internet_Audience_in_Russia_2018.pdf (дата обращения: 22.07.2022).

3. Vīksna R., Jēkabsons G. Sentiment analysis in Latvian and Russian: A survey // Appl. Comput. Syst. 2018. P. 45-51.

4. Gatti L., Guerini M., Turchi M. SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis // IEEE Trans. Affect. Comput. 2016. P. 409-421.

5. Rogers A. RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian / A. Rogers, A. Romanov, A. Rumshisky [etc.] URL: https://text-machine.cs.uml.edu/projects/rusentiment/ (дата обращения: 22.07.2022).

6. RuBERTforSentimentAnalysis. URL: https://huggingface.co/blanchefort/rubert-base-cased-sentiment-rusentiment (дата обращения: 22.07.2022).

7. RuReviews: An Automatically Annotated Sentiment Analysis Dataset for Product Reviews in Russian. URL: https://github.com/sismetanin/rureviews (дата обращения: 22.07.2022)

8. Aspect-Based Sentiment Analysis of Russian Hotel Reviews. URL: https://publications.hse.ru/en/chapters/243615096 (дата обращения: 22.07.2022)

9. SemEval-2016 Task 4: SentimentAnalysisinTwitter. URL: https://metatext.io/datasets/semeval-2016-task-4 (дата обращения: 22.07.2022).

10. Рубцова Ю. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора // Инженерия знаний и технологии семантического веба. 2012. Т. 1. С. 109-116.

Воронежский государственный технический университет
Voronezh State Technical University

Поступила в редакцию 30.07.2022

Информация об авторах

Чапурин Евгений Юрьевич – ассистент, Воронежский государственный технический университет, e-mail: mnac@comch.ru.

Лантюхов Никита Михайлович– студент, Воронежский государственный технический университет, e-mail: mnac@comch.ru

Макаров Олег Юрьевич – д-р техн. наук, профессор, Воронежский государственный технический университет, e-mail: mnac@comch.ru.

AUTOMATED DATABASE OF DESTRUCTIVE CONTENT: METHODOLOGICAL AND MATHEMATICAL SUPPORT OF THE TEXT ANALYZER

E.Yu. Chapurin, N.M. Lantyukhov, O.Yu. Makarov

The article discusses the creation of an automated database of destructive content with the possibility of tonal text analysis in the Telegram messenger, taking into account the structural, functional and methodological support. An algorithm for analyzing content with textual content has been developed, in which the decision on destructiveness is made immediately after a deep analysis of the text. Based on the developed algorithm, software was written that allows not just adding content based on keywords predefined in the database, but using tonal text analysis. The developed methodological, algorithmic and software for updating the database can be used to analyze various messengers, such as WhatsApp and Viber, and the mathematical software will allow improving the system based on the metrics obtained.

Keywords: database, destructive content, social networks.

Submitted 30.07.2022

Information about the authors

Evgeny Yu. Chapurin – Assistant, Voronezh State Technical University, e-mail: mnac@comch.ru
Nikita M. Lantyukhov – Student, Voronezh State Technical University, e-mail: nmac@comch.ru
Oleg Yu Makarov – Dr. Sc. (Technical), Professor, Voronezh State Technical University, e-mail: mnac@comch.ru