ДИСКУССИОННАЯ ПЛОЩАДКА

УДК 004.89
DOI: 10.17223/19988648/47/17

А.Л. Богданов, И.С. Дуля

СЕНТИМЕНТ-АНАЛИЗ КОРОТКИХ РУССКОЯЗЫЧНЫХ ТЕКСТОВ В СОЦИАЛЬНЫХ МЕДИА

Стремительный рост популярности социальных медиа (*Facebook, Twitter,…*) *повышает интерес к проблеме сентимент-анализа – автоматического выделения в текстах сообщений пользователей эмоционально окрашенной лексики, например эмоциональных оценок авторов сообщений по отношению к обсуждаемым темам, объектам, событиям и т.п. Огромные объемы уже накопленных данных и скорость поступления новых не оставляют шансов заинтересованным лицам и компаниям на проведение анализа данных в ручном режиме, что делает разработку инструментов автоматического извлечения требуемой информации актуальной задачей. В данной работе предлагается подход к решению задачи сентимент-анализа коротких русскоязычных текстов на основе их векторного представления. При проведении исследования использовался самостоятельно подготовленный корпус коротких русскоязычных текстов, состоящий из более чем 112 тысяч записей, разметка которого была выполнена в автоматическом режиме на основе маркеров. На данном корпусе было выполнено сравнение эффективности трех алгоритмов классификации: дерева решений, многослойного персептрона и логистической регрессии. Лучший из построенных классификаторов продемонстрировал точность* (*Accuracy*) *76,2 %, что является достаточно высоким значением для задач данного класса и позволяет применять предложенный метод подготовки данных и обучения классификатора на практике при проведении маркетинговых исследований и мониторинга лояльности аудитории к конкретной теме или бренду.*

Ключевые слова: *сентимент-анализ, анализ естественного языка, машинное обучение, обучение с учителем, обучение без учителя, анализ данных.*

Введение

За последние 10 лет социальные медиа, такие как Facebook, Twitter, ВКонтакте, стали неотъемлемой частью жизни общества. Огромное количество компаний строит свой бизнес на платформе социальных медиа [1]. Сегодня уже невозможно вообразить, что более-менее крупная организация не была бы представлена в социальных медиа. Социальные медиа позволяют поддерживать отношения с потребителями, быстро реагировать на обращения и отзывы, проводить рекламные кампании.

Развитие социальных медиа не только ознаменовало переход к новым источникам получения данных, но и позволило каждому пользователю стать таким источником. Журналисты и репортеры лишились привилегии

быть единственными источниками новостей. Теперь каждый пользователь, встретив что-нибудь интересное, может опубликовать соответствующую информацию в своем микроблоге.

Такое качественное изменение способа распространения информации открывает невероятные возможности перед человеком и обществом в целом. Любое событие, информацию о котором опубликовал хотя бы один из пользователей социальных медиа, становится доступным для остальных людей [2]. По этой причине социальные медиа часто называют озерами или океанами информации, в которых скапливается информация практически обо всем, что происходит в мире.

Объемы информации в социальных медиа огромны. Например, в сети Facebook, активная аудитория которой на конец лета 2018 г. составляла 2,23 миллиарда человек, ежеминутно публикуется более полумиллиона комментариев и размещается более ста тысяч фотографий. Аудитория Twitter каждый день публикует более полумиллиарда твитов, что составляет около 200 миллиардов постов за год. Это беспрецедентно большой объем неструктурированной информации. Появляется необходимость в инструменте, который мог бы решать задачи автоматического извлечения из публикаций интересующей информации, отделения отзывов от рекламы, определения отношения пользователей к интересующей теме и т.д. [3].

Типы контента публикаций в социальных медиа сильно разнятся. Для некоторых социальных медиа основным видом контента является видео (YouTube), для других – фотографии (Instagram, Pinterest), для третьих – текст (Twitter, Facebook). В данной работе речь пойдет исключительно о текстовой информации, способах ее обработки, а именно о семантическом анализе коротких текстов.

1. Обзор социальных медиа

Под *социальными медиа* понимают совокупность интернет-площадок, которые предоставляют пользователям возможность устанавливать коммуникацию друг с другом и производить пользовательский контент. На сегодняшний день социальные медиа существенно дифференцированы. Обычно исследователи выделяют 7 видов социальных медиа: *блоги*, *социальные сети*, *вики*, *форумы*, *подкасты*, *микроблоги* и *контент-сообщества* [4]. В табл. 1 перечислены наиболее распространенные социальные медиа.

Наибольший интерес для аналитика представляет деление социальных медиа по типу основного контента: текст, изображения, видео и т.д. Например, основа Twitter – это микроблоги, т.е. текстовая информация, в Instagram пользователи делятся изображениями со своими подписчиками, Youtube позиционирует себя как сервис видеохостинга.

По данным Фонда общественного мнения [5], доля интернет-активного населения растет каждый год (рис. 1). Этот рост порождает большое количество неструктурированных данных. В публикациях пользователи остав-

ляют отзывы о товарах и услугах, формируют тренды, делятся своими предпочтениями и интересами. По этой причине крупные компании имеют высокий интерес к анализу данных социальных медиа, желая лучше понять своих потребителей [6].

Таблица 1. **Наиболее распространенные социальные медиа**

Название	Число активных пользователей в месяц
Facebook	1,79 млрд
Twitter	313 млн
LinkedIn	467 млн
Instagram	500 млн
Pinterest	150 млн

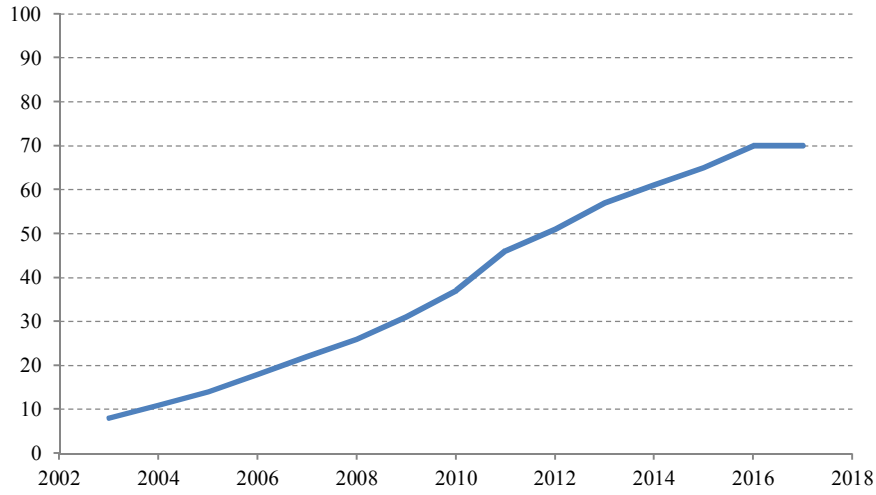

Рис. 1. Месячная доля интернет-активного населения, %

Помимо этого, компании заинтересованы в коммуникации со своими потребителями в социальных медиа, так как это существенно повышает узнаваемость бренда и лояльность аудитории. Потребители все чаще используют социальные медиа для принятия решения о покупке. Они ищут отзывы, смотрят мнение известных блогеров, нередко принимают решение на основании того, как часто они контактируют с брендом в социальных медиа.

Появление социальных медиа позволило пользователям самостоятельно создавать контент и делиться им с другими пользователями. Сегодня каждый отдельный пользователь – ценный источник информации, транслирующий окружающим свои мысли и мнения. В условиях, где почти каждый может стать источником информации, практически невозможно предположить, что упоминание о сколь угодно значимом событии не было отражено в социальных медиа.

1.1. Twitter

Twitter позиционируется как микроблогинговая среда. Он был запущен для массового использования в 2007 г. На сегодняшний день на сервисе зарегистрировано почти 1,5 млрд пользователей, среди них активных пользователей чуть более 300 млн. Отличительной особенностью сервиса на протяжении почти десяти лет было ограничение максимальной длины сообщения – 140 символов. Сегодня максимальная длина сообщений увеличена до 280 символов.

Twitter хорошо подходит для анализа естественного языка, поскольку существенная часть пользователей публикуют информационно значимые сообщения и придерживаются общепринятых норм языка. На рис. 2 приведена статистика содержания публикаций в Twitter [7].

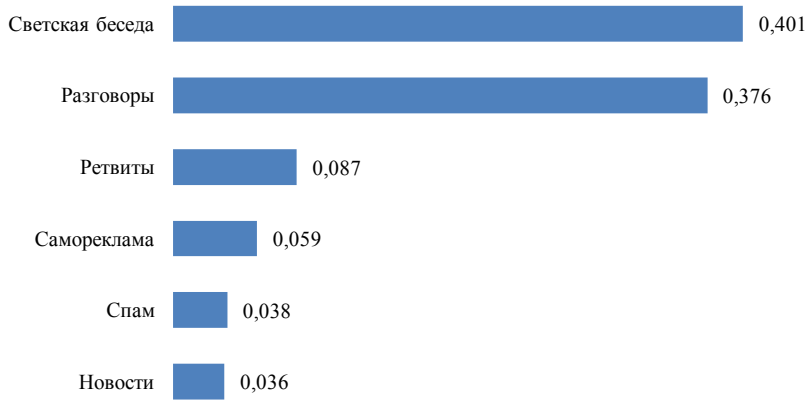

Рис. 2. Содержание публикаций Twitter

Twitter предлагает официальный *прикладной интерфейс разработчика* (API), предоставляющий доступ к публикациям по соответствующему запросу. Для получения выборки твитов используется метод `search()`. Метод имеет ряд ограничений, например, за одно обращение можно получить не более 100 твитов для пользователя и 450 для авторизованного приложения. Также в целях снижения нагрузки на серверы установлены 15-минутные интервалы ожидания при превышении числа лимита запросов. В качестве ответа метод возвращает кортеж твитов в формате JSON. Каждый твит содержит атрибуты: имя автора, текст сообщения, время публикации, количество ретвитов, количество подписчиков автора и др.

1.2. Facebook

Сервис был основан в 2004 г. Марком Цукербергом и его соседями по комнате во время обучения в Гарвардском университете. Изначально сер-

вис был предназначен только для студентов Гарварда, но начиная с 2006 г. сервис стал доступен по всему миру. Сейчас Facebook имеет более 1,7 млрд активных пользователей, количество просмотров страниц сайта превышает один триллион [8].

Facebook поддерживает публикацию текстовых сообщений, фото и видео. Существенная часть публикаций носит новостной характер или характер отзыва на различные события и товары. По сравнению с отечественными социальными медиа доля спама в Facebook относительно невысока. Аналитики, работающие с анализом естественного языка, выделяют Facebook как сервис с широким охватом аудитории и высоким качеством публикуемого контента.

Разработчики Facebook также предоставили пользователям удобный инструмент для построения запросов к социальной сети – Graph API. Данные, получаемые с его помощью, представлены в формате JSON. Graph API также имеет ограничения по количеству получаемых данных и количеству запросов в единицу времени.

2. Сентимент-анализ текста

Под *сентимент-анализом* текста понимают класс методов автоматического выделения в текстах эмоционально окрашенной лексики и эмоциональной оценки авторов (их мнений) по отношению к объектам, речь о которых идет в тексте [9]. Сентимент-анализ применяется в маркетинговых исследованиях, в задачах мониторинга лояльности аудитории к различным темам и брендам и в других случаях.

Тексты, публикуемые в социальных медиа, характерно отличаются разговорным стилем речи. С данной особенностью связана большая часть трудностей: наличие сленга, видоизмененные слова, опечатки и другие особенности, которые затрудняют анализ текста [10]. Качество текста существенно зависит от конкретной социальной сети. Тексты следует различать *субъективные* и *объективные*. Объективный текст содержит информацию о каком-либо событии без личного мнения автора, соответственно, субъективный – наоборот. Интерес представляют именно субъективные суждения, позволяющие извлечь мнение конкретного автора. Объективные публикации, как правило, исключаются, так как они обычно носят информационный характер.

В большинстве случаев для определения эмоциональной окраски текста используют бинарную классификацию публикаций на позитивные и негативные [11]. Иногда публикации классифицируют на три класса, к уже упомянутым добавляется класс нейтральных текстов [12]. Очень редко прибегают к регрессии, где эмоциональная окраска определяется на интервале от 0 до 1 (0 – негативный, 1 – позитивный). Последний подход применяется, когда есть необходимость проранжировать тексты по эмоциональной окраске.

Часто исследователи обращают внимание на различные эмоциональные маркеры, например *стикеры* или *смайлы*. При таком подходе делается предположение, что маркеры соответствуют общей тональности высказывания. Также в качестве маркеров используются различные *хэштеги*, определяющие общую тональность публикации [13]. Данный подход не лишен своих недостатков, ибо сарказм и иные средства также могут использоваться при добавлении хэштегов и смайлов.

Русскоязычные тексты хуже поддаются анализу в силу обильного использования средств выразительности и переносных значений слов или фраз [14]. В случае, если выделение подобных средств выразительности критично, то обучают дополнительные классификаторы, определяющие их, после чего результаты работы классификаторов используются в качестве дополнительных признаков.

Принято выделять четыре основных подхода к определению тональности высказывания [15]:

1. Подход, основанный на правилах.
2. Подход, основанный на словарях.
3. Машинное обучение с учителем.
4. Машинное обучение без учителя.

Первые два подхода относительно просты и легко интерпретируемы, но имеют существенные недостатки: низкую универсальность и высокую трудоемкость. Подход, основанный на использовании банка позитивных и негативных слов, заключается в подсчете позитивных и негативных слов из словаря в каждом тексте [16]. Если количество позитивных слов превышает количество негативных, то текст считается позитивным, и наоборот. Очевидно, что такой подход малоэффективен: для его осуществления необходимо создать относительно большой банк слов и периодически его обновлять. К тому же факт того, что количество позитивных слов превышает количество негативных, не является надежным критерием, по которому можно с уверенностью сказать, что весь текст имеет позитивную семантику. В наше времяболее широкое распространение получили третий и четвертый подходы [17].

2.1. Мотивация использования сентимент-анализа

Существенная доля мировых данных представлена в текстовом виде, например: электронные письма, посты в социальных медиа, статьи, документы. Текстовые данные по своей природе являются неструктурированными, что существенно затрудняет их обработку, но в то же время текстовые данные содержат массу полезных знаний. Поэтому интерес к системам автоматического анализа текстов неуклонно возрастает.

Системы анализа мнений, которые относятся к данному классу систем, позволяют компаниям в автоматическом режиме извлекать из текстовых данных полезные знания, что, в свою очередь, позволяет экономить часы ручного труда и автоматизировать многие бизнес-процессы.

К достоинствам систем сентимент-анализа относят следующие:

1. Масштабируемость. Невозможно представить, как вручную сортировать тысячи постов в социальных медиа, разговоры служб поддержи или отзывы клиентов. Анализ тональности высказываний позволяет обрабатывать данные в большом объеме эффективным и экономичным способом. Увеличение объема обработки данных приводит к незначительному увеличению стоимости, вызванному покупкой дополнительного дискового пространства и вычислительных мощностей.

2. Анализ в реальном времени. Анализ настроений можно использовать для выявления важной информации, которая обеспечивает информационную осведомленность в конкретных ситуациях в режиме реального времени. Система анализа настроений может выявлять на ранних стадиях PR-кризисы и неудовлетворенных клиентов.

3. Согласование критериев оценивания. Когда оценкой тональности высказываний занимаются люди, даже один человек в зависимости от разных факторов (настроение, отношение к теме, ...) может давать разную оценку одним и тем же сообщениям. Проблема усугубляется, когда задачей оценки мнений пользователей занимаются сразу несколько людей. В этом случае крайне тяжело согласовать рейтинги двух разных оценщиков: один человек может оценить мнение как позитивное, другой как нейтральное и т.д. Это субъективная задача, на которую сильно влияют личный опыт, мысли и убеждения. Используя централизованную систему анализа мнений, компании применяют одинаковые критерии оценивания ко всем данным. Это снижает количество ошибок и улучшает согласованность данных.

2.2. Примеры использования сентимент-анализа

Сентимент-анализ используется в большом количестве отраслей. Области применения анализа тональности высказываний взаимосвязаны и направлены на оценку изменения в общественном мнении к конкретной теме или объекту. Выделяют шесть основных применений сентимент-анализа:

1) мониторинг брендов;
2) исследование конкурентов;
3) поддержка клиентов;
4) продуктовая аналитика;
5) маркетинговые исследования, поиск инсайтов и трендов в индустриях;
6) мониторинг мнений сотрудников.

Мониторинг брендов. Люди любят делиться своими мнениями о последних новостях, местных и глобальных событиях, своем потребительском опыте. Новости о знаменитостях, предпринимателях и глобальных компаниях привлекают тысячи пользователей в течение нескольких часов после публикации. Так почему бы компаниям не использовать этот источник для мониторинга того, что публика думает и говорит о ней? Анализ мнений в социальных медиа позволит иметь представление о репутации

компании среди ее клиентов, обнаружить возникающие репутационные кризисы и быстро реагировать на них. Например, можно отслеживать изменения лояльности аудитории в динамике (по дням, неделям и месяцам) и в случае резкого изменения лояльности искать причину. Также важно не только знать общественное мнение, но и быть в курсе, кто говорит. Измерение тональности высказываний позволяет определить компаниям, упоминают ли влиятельные персоны в отрасли их бренд и в каком контексте.

Исследование конкурентов. Компании и их конкуренты имеют общую целевую аудиторию. Поэтому компании могут исследовать мнения целевой аудитории как по отношению к себе, так и по отношению к конкурентам. Подобное исследование позволяет компании ответить на ряд вопросов. Что ценят клиенты у других игроков отрасли? Есть ли у конкурентов что-то такое, чего нет у компании (технология, продукт)? Используя эти знания, компании могут улучшить свои коммуникационные и маркетинговые стратегии, обслуживание клиентов или принять решение о разработке новых продуктов. Конкурентный анализ, включающий анализ настроений, помогает оценить компаниям свои слабые и сильные стороны и найти способы выделиться на фоне конкурентов.

Поддержка клиентов. Гостиничные бренды, финансовые учреждения, предприятия розничной торговли, транспортные компании и другие предприятия используют анализ тональности мнений для оптимизации работы отдела обслуживания клиентов. Так, компании из постов в социальных медиа определяют уровень удовлетворенности клиентов работой отдела поддержки. Часто вместе с анализом тональности используются другие методы, например, определение наиболее популярных запросов и тем сообщений, анализ причин обращений в службу поддержки. Полученные результаты анализируются с целью скорейшего реагирования на проблему, допустим, решение проблем может начинаться с наименее счастливых или наиболее рассерженных клиентов.

Продуктовая аналитика. Каждый бизнес желает видеть очередь клиентов, ожидающих открытия магазина, чтобы купить новый продукт. Как вывести такой продукт на рынок? Естественный подход – спросить людей, чего они хотят. Успешные компании создают *минимально жизнеспособный продукт* (MVP), собирают отзывы о нем, проводят анализ, учитывают проблемы и исправляют их. Затем запускают промышленный вариант продукта, продолжая постоянно его улучшать. Данные обратной связи поступают из опросов, социальных медиа, форумов, историй взаимодействия со службой поддержки. Сортируя данные по темам и настроениям, компании узнают, какие функции в продукте необходимы, а от каких стоит избавиться. Полученные результаты анализа мнений дают команде разработчиков продукта и продуктовым менеджерам информацию, как сделать продукт, который понравится целевой аудитории и будет хорошо продаваться.

Маркетинговые исследования, поиск инсайтов и трендов. Как уже было сказано, социальные медиа и форумы являются источниками информации на любую тему. Люди обсуждают новости и продукты, пишут о своих

ценностях, мечтах, потребностях и событиях. Они делают это добровольно в режиме 24 часа в день семь дней в неделю. Анализ настроений решает проблему обработки больших объемов неструктурированных данных. Используя сентимент-анализ, маркетологи отслеживают и изучают модели поведения потребителей в режиме реального времени, чтобы предсказывать будущие тенденции и помогать руководству компании принимать обоснованные решения.

Мониторинг мнений сотрудников. Некоторые компании выходят за рамки использования анализа настроений для исследования рынка или оценки опыта клиентов, применяя его во внутренних процессах. Компании измеряют удовлетворенность сотрудников, выявляют факторы, которые мешают сотрудниками и в конечном итоге снижают эффективность работы компании. Специалисты автоматизируют анализ опросов сотрудников с помощью программного обеспечения, что позволяет им быстро находить и решать проблемы. HR-менеджеры могут определять тональность результатов опросов сотрудников, группировать данные по отделам и темам, отслеживать изменения настроений в динамике. Анализ настроений также помогает автоматически отслеживать психологическое состояние сотрудников на основании опросов и постов в социальных медиа. На основании этой информации HR-менеджер может принять решение предоставить сотруднику небольшой отдых, повысить бонусы или рекомендовать обратиться к психологу.

Таким образом, анализ тональности высказываний позволяет компаниям использовать огромное количество открытых данных для изучения потребностей клиентов, их отношения к своему бренду, отслеживать процесс общения клиентов со службой поддержки, сортировать запросы в порядке приоритетов, изучать настроения сотрудников в самой компании и улучшать условия труда.

3. Основные проблемы работы с текстовыми данными и методы их решения

Задача сентимент-анализа предполагает решение следующих проблем [18]:

1. Разметка текстов.
2. Очистка текстов от нерелевантных символов и слов.
3. Исправление ошибок в текстах.
4. Представление текстов в векторной или матричной форме.
5. Удаление стоп-слов.
6. Построение модели.

3.1. Разметка текстов

Использование машинного обучения с учителем предполагает наличие *меток классов* (*labels*). Небольшой набор можно разметить вручную, однако для качественной классификации необходим набор обучающих данных

большого размера. В таком случае используют либо обучение без учителя, либо размечают набор данных с помощью маркеров [19].

Для разметки текстов иногда используются слова-маркеры или специальные символы, например *смайлы*. Такой подход соответствует гипотезе, что семантическая составляющая текста соответствует используемым маркерам в нем. Предложения, в которых отсутствовали выбранные маркеры, игнорируются. Обычно используется бинарная разметка на позитивные и негативные тексты, поскольку тяжело выделить маркеры с нейтральной семантикой. В табл. 2 приведены примеры позитивных и негативных смайлов-маркеров, использующихся для разметки.

Таблица 2. **Позитивные и негативные смайлы-маркеры**

Позитивные			Негативные		
Смайл	Unicode	Название	Смайл	Unicode	Название
😁	U+1F601	Beaming face with smiling eyes	😞	U+1F61E	Disappointed face
😂	U+1F602	Face with tears of joy	😠	U+1F620	Angry face
😃	U+1F603	Grinning face with big eyes	😡	U+1F621	Pouting face
😄	U+1F604	Grinning face with smiling eyes	😢	U+1F622	Crying face
😅	U+1F605	Grinning face with sweat	😣	U+1F623	Persevering face
😆	U+1F606	Grinning squinting face	😤	U+1F624	Face with steam from nose
😇	U+1F607	Smiling face with halo	😥	U+1F625	Sad but relieved face
😈	U+1F608	Smiling face with horns	😨	U+1F628	Fearful face
😉	U+1F609	Winking face	😩	U+1F628	Weary face
😊	U+1F60A	Smiling face with smiling eyes	😪	U+1F62A	Sleepy face
😋	U+1F60B	Dace savoring food	😫	U+1F62B	Tired face
😌	U+1F60C	Relieved face	😭	U+1F62D	Loudly crying face
😍	U+1F60D	Smiling face with heart–eyes	😰	U+1F630	Anxious face with sweat
😎	U+1F60E	Smiling face with sunglasses			
😏	U+1F60C	Smirking face			

Смайлы используются только на этапе разметки набора данных. Затем они удаляются, чтобы в процессе настройки параметров классификатора не попасть в ситуацию переобучения.

3.2. Мешок слов

Текстовые данные не являются структурированными, поэтому к ним нельзя непосредственно применить методы машинного обучения. Перед использованием текстовые данные проходят предварительную обработку, кодируются и преобразуются в векторы. Существует множество способов это сделать. Одним из таких подходов является метод «мешок слов» (*bag of words*).

Суть метода заключается в следующем. Из слов, содержащихся в множестве текстов, составляется *множество слов* (*словарь*), использованных в них. Затем для каждого текста подсчитывается количество вхождений каждого слова из словаря. В результате каждому тексту ставится в соответствие вектор, где каждая координата равна числу упоминаний соответствующего слова в тексте (рис. 3).

Рис. 3. Реализация подхода «мешок слов»

К недостаткам метода «мешок слов» можно отнести то, что порядок слов в тексте не учитывается и при векторизации новых текстов, и то, что слова, отсутствующие в словаре, игнорируются. Тем не менее данный подход очень популярен.

3.3. Статистическая мера TF-IDF

Часто при классификации коротких текстов аналитики сталкиваются с проблемой зашумления текстов отдельными малозначимыми словами. Такие слова называют *стоп-словами* (*stop-words*) или *шумовыми словами*. Ранее был распространен подход удаления стоп-слов, основанный на словаре. Суть данного подхода заключается в том, что из текстов все слова, содержащиеся в словаре стоп-слов, удаляются. Недостатками данного подхода является сложность учета всех возможных стоп-слов и необходимость непрерывной поддержки словаря в актуальном состоянии.

Другой подход к решению проблемы наличия в тексте малозначимых слов основан на статистической мере *TF-IDF* (*TF – term frequency, IDF – inverse document frequency*). Данная статистическая мера используется для оценки важности слова в *контексте документа* (*текста*), который является частью *корпуса* (*множества текстов*). Эта мера каждому слову ставит в соответствие вес, который пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всем корпусе.

Статистическая мера *TF-IDF* рассчитывается по следующей формуле:

$$\boldsymbol{tf}_{\boldsymbol{idf}(\boldsymbol{t},\boldsymbol{d},\boldsymbol{D})} = \boldsymbol{tf}(\boldsymbol{t},\boldsymbol{d}) * \boldsymbol{idf}(\boldsymbol{t},\boldsymbol{D}) = \frac{\boldsymbol{n_t}}{\sum \boldsymbol{n_k}} * \boldsymbol{log}\frac{|\boldsymbol{D}|}{|\{\boldsymbol{d_i}\,|\boldsymbol{d_i}\boldsymbol{\epsilon}\,\boldsymbol{D},\boldsymbol{t}\,\boldsymbol{\epsilon}\,\boldsymbol{d_i}\}|},$$

где n_t – число вхождений слова *t* в документ; $\sum n_k$ – общее число слов в документе; $|D|$ – число документов в корпусе; $|\{d_i\,|d_i \epsilon\, D, t\,\epsilon\, d_i\}|$ – число документов корпуса *D,* где встречается термин *t* (когда $n_t \neq 0$).

Высокое значение показателя *TF-IDF* говорит о том, что слово часто встречается в пределах какого-то одного конкретного документа и редко встречается в других документах. Таким образом, слово с высоким значением показателя *TF-IDF* является важным в данном документе.

3.4. Векторизация слов

Векторизация слов – это еще один метод представления текстовых данных в виде множества векторов, путем построения специализированной модели [20]. Построенная модель должна удовлетворять естественному требованию: чем меньше расстояние между векторами, тем ближе семантическая составляющая соответствующих им слов. Например, векторы слов «машина» и «самолет» должны располагаться ближе друг к другу, чем слова «космос» и «карандаш».

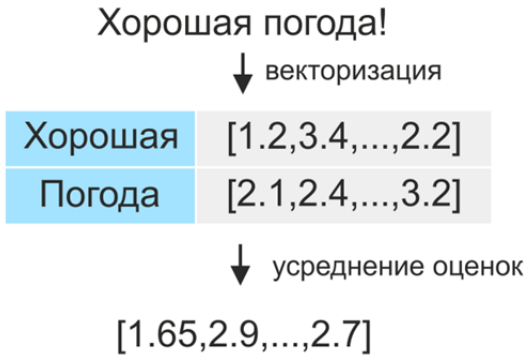

Рис. 4. Усреднение оценок Word2Vec

Векторизация слов – технически очень сложный процесс, по этой причине аналитики обычно используют уже готовые решения. Например, Word2Vec, представляющее собой технику обучения *векторизатора* (*vec-*

torizer) [26]. Обучение происходит в результате анализа большого количества текстов с запоминанием слов, которые возникают в схожих контекстах. После обучения Word2Vec позволяет представить слова в виде векторов из пространства заранее заданной размерности.

Используя данный подход, каждый текст можно представить в виде матрицы $n \times k$, где n – количество слов в тексте, а k – размерность вектора отдельного слова. Для использования классификационных моделей, требующих векторного представления входных данных, необходимо представить полученную матрицу в виде вектора, не потеряв большую часть информации. Для решения этой задачи достаточно усреднить оценки Word2Vec (рис. 4). Для работы с матричным представлением текстов используются *сверточные нейронные сети* (*convolutional neural network*).

3.5. Классификационные модели

Общепринятой практикой решения задачи построения классификационной модели является движение от простой модели к сложной. Самым распространенным и простым классификатором служит логистическая регрессия. Для оценки качества модели в данной задаче используется метрика *Accuracy* (точность), которая рассчитывается как отношение правильных предсказаний классификатора к общему числу сделанных предсказаний.

Как показывает практика, в задачах бинарной классификации (два класса: позитивный и негативный) англоязычных коротких текстов значение показателя *Accuracy* классификаторов колеблется в диапазоне 70–75% (при условии равномерного распределения классов), что считается неплохим результатом для подобной задачи. Примерно такую же точность демонстрируют многослойный персептрон (*multilayer perceptron*) и деревья решений (*decision tree*) [21]. Несомненным достоинством подхода на основе логистической регрессии является интерпретируемость коэффициентов модели: слова с наибольшими значениями коэффициентов являются наиболее значимыми.

Самую высокую точность исследователям удавалось достичь, используя синтаксис предложений [22]. Все упомянутые модели не могут по своей природе работать с порядком слов, поскольку для сохранения порядка слов необходимо представить текст в виде матрицы, где строка – это векторизованное слово. Тем не менее это под силу сверточным нейронным сетям [23]. Сверточные нейронные сети используют только при необходимости достижения наивысшей точности классификации. На практике они, как правило, используются редко, так как требуют существенно большего количества времени и ресурсов для проведения вычислений, а точность модели (значение показателя *Accuracy*) по сравнению с обычными классификаторами увеличивается в среднем только на 2%.

На рис. 5 представлена модель сверточной нейронной сети с двумя каналами представления предложений [22].

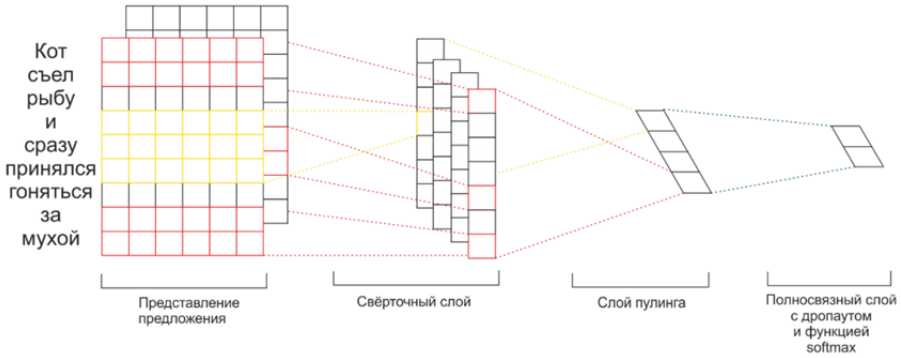

Рис. 5. Архитектура модели с двумя каналами представления предложений

В представленном случае на вход сети подается матрица, каждая строка которой – это векторизованное слово. Далее, как в простейшей сверточной нейронной сети, идет один *слой свертки* (*convolution layer*), один слой *пулинга* (*pooling layer*) и *полносвязный слой с дропаутом* (*dropout*). Автор этой работы предлагает различные вариации данной модели, например, использовать два входных канала: статический и нестатический. Статический канал сохраняется на протяжении всего обучения, а нестатический, соответственно, настраивается в ходе обучения.

4. Предлагаемое решение

4.1. Создание корпуса текстов

Для задачи классификации англоязычных коротких текстов не составляет труда найти качественный размеченный корпус. Однако корпусов для классификации русскоязычных текстов практически нет, а к имеющимся возникают вопросы относительно их качества. В связи с этим было решено создать собственный корпус коротких русскоязычных текстов.

На первом этапе был произведен сбор релевантных к текущей задаче публикаций. Наиболее легким способом получения публикаций, безусловно, является официальный Twitter API для разработчиков. Для него разработан ряд библиотек, упрощающих работу, наиболее популярная из них – tweepy [24]. Метод `search()` данной библиотеки позволяет получить публикации по соответствующему запросу.

Разметка собранных текстов проводилась в автоматическом режиме с помощью маркеров-смайлов (табл. 2) по следующему правилу: если текст содержит позитивный смайл, то он считается позитивным, если он содержит негативный смайл, то он считается негативным. В случае конфликта, т.е. когда текст содержит одновременно и позитивный и негативный смайлы, текст попадает одновременно в корпус негативных и в корпус позитивных текстов. Тексты, которые не содержали смайлы, не включались ни в один из корпусов.

В табл. 3 приведены примеры публикаций, размеченных с помощью данного метода. Как видно из таблицы, каждый маркер релевантен семантической составляющей каждого текста.

Таблица 3. **Размеченные публикации**

Позитивные	Негативные
Ну вот и еще 10 товаров, на которые дополнительно снизила цены	Рубрика «Ужасы моей профессии»
Пинтерест: у вас хороший вкус! *показывает фотку Чонгука* Я: спасибо конечно, я знаю	В очередной раз встала утром и поняла, что не буду мыть голову, потому что не успею высохнуть
Скоро релиз. Трек уже почти готов	Болею уже 2 раз за сентябрь и это только начало учебного года. А что дальше будет!?
Улицу Фрунзе открыли, можете теперь там ездить спокойно!	Сходила на тренировку и чуть не отбросила копыта
Почитала методичку по химии. Поняла немного, но уже чувствую себя на коне	Меня преследует чувство что все совместные фото с классом испорчены моим лицом

В результате выполнения данной процедуры был получен корпус коротких русскоязычных текстов, содержащий 112 994 записи, из которых 47 503 негативных записей и 65 491 – позитивных. Общее количество слов в корпусе составило 1 488 425. Словарь (множество уникальных слов), построенный из слов корпуса, содержал 173 021 слово. Минимальная длина текста в корпусе равна одному слову, максимальная – 45. На рис. 6 представлена гистограмма распределения количества слов в текстах.

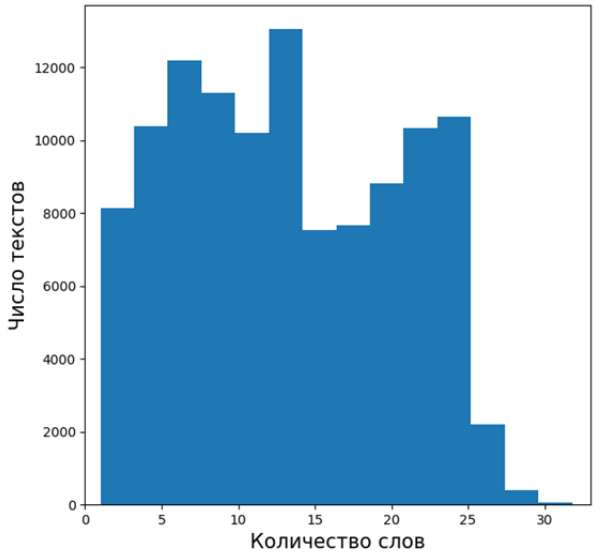

Рис. 6. Гистограмма распределения количества слов в текстах

4.2. Подготовка данных

Цель предобработки – представить данные в форме, которая будет наиболее удобна для построения модели. Предобработка текстов состоит из двух шагов: *очистки* и *преобразования в векторную форму*.

На первом шаге из текстов были удалены нерелевантные символы, не относящиеся к цифрам или буквам, нерелевантные слова, а именно: ссылки, специальные метки (метка репоста и т.п.), отметки других пользователей, числа, знаки препинания.

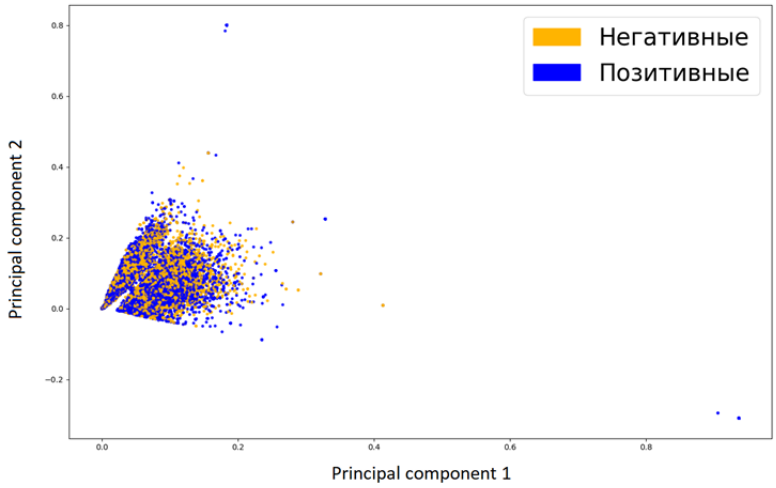

Рис. 7. Диаграмма рассеяния публикаций в двухмерном пространстве

На втором шаге тексты были *токенизированы*, т.е. было произведено разделение их на отдельные слова. Также все слова были переведены в нижний регистр. Затем тексты были представлены в векторной форме с помощью подхода «мешок слов» (*bag of words*) и для определения веса каждого слова была использована статистическая мера *TF-IDF* в целях снижения влияния стоп-слов и повышения качества классификации. Перед построением модели был применен *метод анализа главных компонент* (*principal component analysis*) для визуальной оценки возможности разделения множеств позитивных и негативных текстов (рис. 7). Как видно из диаграммы рассеяния, два множества визуально неразделимы в двухмерном пространстве.

4.3. Классификация

Для построения классификационной модели были выбраны три классификатора: логистическая регрессия, дерево решений, многослойный персептрон. Для каждого классификатора была посчитана точность методом 10-блочной кроссвалидации. Наибольшую кроссвалидационную точность

(*CV-score*) продемонстрировала логистическая регрессия (табл. 4). С учетом того, что среди рассмотренных классификаторов модель логистической регрессии легко поддается интерпретации, быстро обучается и показывает наивысшую точность, был сделан выбор в ее пользу.

Таблица 4. **Кросс-валидационная точность моделей**

Номер тестового блока	Дерево решений	Логистическая регрессия	Многослойный персептрон
1	0,7472	0,7568	0,7587
2	0,7402	0,7579	0,7551
3	0,7429	0,7611	0,7593
4	0,7398	0,7567	0,7551
5	0,7500	0,7707	0,7669
6	0,7418	0,7687	0,7561
7	0,7408	0,7625	0,7519
8	0,7394	0,7629	0,7578
9	0,7448	0,7588	0,7457
10	0,7494	0,7633	0,7597
CV–score	0,7436	0,7619	0,7566
Std	0,0040	0,0048	0,0055

Далее для логистической регрессии были посчитаны средние значения истинно отрицательных, ложноположительных, ложноотрицательных, истинно положительных результатов 10-блочной кроссвалидации. На их основании была построена нормированная матрица несоответствий (рис. 8). Согласно ей точность построенного классификатора составила 76,2%, что является отличным результатом для данной задачи. Ложноположительные и ложноотрицательные исходы распределены практически равномерно, т.е. классификатор не отдает предпочтения ни одному из классов, и так как исходная задача предполагает равные издержки классификации для ложноположительных и ложноотрицательных результатов, корректировка порогового уровня не требуется.

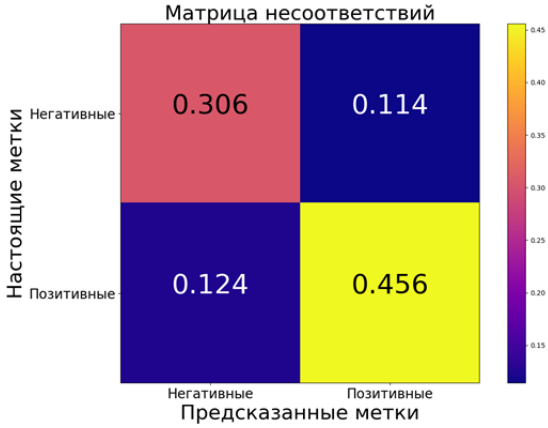

Рис. 8. Матрица несоответствий

Для лучшего понимания работы модели и ее интерпретации были отобраны признаки (слова) с наименьшими и наибольшими коэффициентами (рис. 9). Некоторые слова по этическим соображениям были заменены октоторпами (знаками решетки). Среди слов, соответствующих как наиболее низким весовым коэффициентам, так и наиболее высоким, присутствуют слова, которые по смыслу не несут соответствующую позитивную или негативную семантику. Например, слова «поэтично», «прихожанин» среди наиболее негативных или «удалилась» среди позитивных.

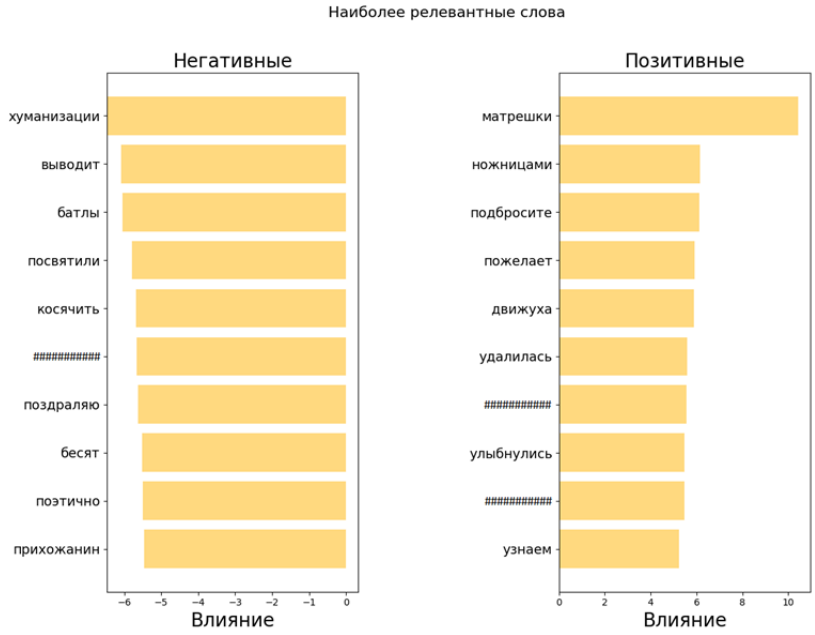

Рис. 9. Слова с наиболее значимыми коэффициентами

Заключение

Рост объема неструктурированных текстовых данных стимулирует интерес к задачам анализа естественного языка и к сентимент-анализу в частности. Доступность открытых библиотек машинного обучения позволяет использовать современные алгоритмы для решения задач данного класса. В настоящей работе был собран и размечен корпус коротких русскоязычных текстов, приведено описание процедуры очистки и подготовки исходных данных, построен классификатор тональности коротких русскоязычных текстов, имеющий точность (*Accuracy*) 76,2 %.

Применение сентимент-анализа открывает новые возможности для бизнеса. Автоматическое определение тональности высказываний позволит быстро и дешево проводить исследования в социальных медиа. Предложенный в работе подход может использоваться для проведения маркетин-

говых, социологических и политических исследований. Также он позволяет осуществлять мониторинг лояльности аудитории к конкретной теме или бренду, что дает менеджменту возможность своевременно принимать необходимые решения.

Литература

1. *Baier M., Wagner K.* User Behavior in Crowdfunding Platforms – Exploratory Evidence from Switzerland // Proceedings of Conference: Hawaii International Conference on System Sciences (HICSS), At Kauai, Hawaii, USA. 2016. Vol. 49. P. 3583–3593.

2. *Poecze F., Ebster C., Strauss C.* Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts // Proceedings of The 9th International Conference on Ambient Systems, Networks and Technologies (ANT). 2018. Vol. 130. P. 660–666.

3. *Zeroual I., Lakhouaja A.* Data science in light of natural language processing: An overview // Proceedings of The First International Confidence on Intelligent Computing in Data Science, ICDS. 2017. Vol. 127. P. 82–91.

4. *Mayfield A.* What Social Media Is // ICrossing. URL: https://www.icrossing.com/uk/sites/default/files_uk/insight_pdf_files/What%20is%20Social%20Media_iCrossing_ebook.pdf (дата обращения 21.08.2018).

5. *Интернет* в России: динамика проникновения // Фонд общественного мнения. URL: https://fom.ru/SMI-i-internet/13585 (дата обращения: 23.08.2018).

6. *Шигина Я.И., Фоменков Д.А.* Социальные медиа: современные тенденции в маркетинге // Вестник Казанского технологического университета. 2014. Т. 17, № 24. С. 453–456.

7. *Twitter* Study Reveals Interesting Results About Usage // Pear Analytics. URL: https://38r0us9g9l1438rwf2z2tcsz–wpengine.netdna–ssl.com/wp–content/uploads/2009/08/Twitter–Study–August–2009.pdf (дата обращения: 21.08.2018).

8. *The top* 500 sites on the web // Alexa Internet. URL: https://www.alexa.com/topsites (дата обращения: 23.08.2018).

9. *Pang B., Lee L.* Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. 2018. Vol. 2. P. 1–135.

10. *Boudad N., Faizi R., Oulad Haj Thami R., Chiheb R.* Sentiment analysis in Arabic: A review of the literature // Ain Shams Engineering Journal. 2018. Vol. 9, № 4. P. 2479–2490.

11. *Sokhin T., Butakov N.* Semi-automatic sentiment analysis based on topic modeling // Proceedings of 7th International Young Scientists Conference on Computational Science, YSC2018, Heraklion, Greece. 2018. Vol. 136. P. 284–292.

12. *Tartir S., Abdul-Nabi I.* Semantic sentiment analysis in arabic social media // Arabic Natural Language Processing: Models, Systems and Applications. 2017. Vol. 29, № 2. P. 229–233.

13. *Mallek F., Belainine B., Sadat F.* Arabic Social Media Analysis and Translation // Arabic Computational Linguistics, 2017. Vol. 117. P. 298–303.

14. *Al-Thubaity A., Alqahtani Q., Aljandal A.* Sentiment lexicon for sentiment analysis of Saudi dialect tweets // Arabic Computational Linguistics. 2018. Vol. 142. P. 301–307.

15. *Юсупова Н.И., Богданова Д.Р., Бойко М.В.* Алгоритмическое и программное обеспечение для анализа тональности текстовых сообщений с использованием машинного обучения // Вестник УГАТУ. 2012. Т. 16, № 6. С. 91–99.

16. *Moussa M., Mohamed E., Haggag M.* A survey on opinion summarization techniques for social media // Future Computing and Informatics Journal. 2018. Vol. 3, № 1. P. 82–109.

17. *Amrani Y., Lazaarb M., Kadiri K.* Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis // Proceedings of The First International Confidence on Intelligent Computing in Data Science, ICDS. 2017. Vol. 127. P. 511–520.

18. *Stieglitz S., Mirbabaie M., Ross B., Neuberger C.* Social media analytics – Challenges in topic discovery, data collection, and data preparation // International Journal of Information Management. 2018. Vol. 39. P. 156–168.

19. *Birjali M., Beni-Hssane A., Erritali M.* Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks // Proceedings of The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH). 2017. Vol. 113. P. 65–72.

20. *Araque O., Zhu G., Iglesias A.* A semantic similarity-based perspective of affect lexicons for sentiment analysis // Knowledge-Based Systems. 2019. Vol. 165. P. 346–359.

21. *Ankit S.N.* An Ensemble Classification System for Twitter Sentiment Analysis // Proceedings of International Conference on Computational Intelligence and Data Science. 2018. Vol. 132. P. 937–946.

22. *Yoon K.* Convolution neural networks for sentence classification // arXiv:1408.5882 [cs.CL]. 2014. URL: https://arxiv.org/abs/1408.5882 (дата обращения: 15.09.2018).

23. *Heikal M., Torki M., El-Makky N.* Sentiment Analysis of Arabic Tweets using Deep Learning // Arabic Computational Linguistics. 2018. Vol. 142. P. 114–122.

24. *Tweepy* Documentation // Tweepy. URL: https://tweepy.readthedocs.io/en/v3.5.0/ index.html (дата обращения: 10.09.2018).

25. *Srishty Jindal, Dr.* Kamlesh Sharma Intend to analyze social media feeds to detect behavioral trends of individuals to proactively act against social threats // Proceedings of International Conference on Computational Intelligence and Data Science. 2018. Vol. 132. P. 218–225.

26. *Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean.* Efficient Estimation of Word Representations in Vector Space // In Proceedings of Workshop at ICLR. 2013.

Sentiment Analysis of Short Russian Texts in Social Media

Vestnik Tomskogo gosudarstvennogo universiteta. Ekonomika – Tomsk State University Journal of Economics. 2019. 47. pp. 220–241.
DOI: 10.17223/19988648/47/17

Aleksandr L. Bogdanov, Tomsk State University (Tomsk, Russian Federation). E-mail: bogdanov.al@mail.tsu.ru

Ivan S. Dulya, Tomsk State University (Tomsk, Russian Federation). E-mail: idulya7@gmail.com

Keywords: sentiment analysis, natural language processing, machine learning, supervised learning, unsupervised learning, data analysis.

The rapid growth of the popularity of social media (Twitter, Facebook, etc.) increases interest in the sentiment analysis problem. Sentiment analysis is a method of automatic selection of an emotional component in texts, e.g., the emotional evaluation of considering themes, objects, events, etc. The large volume of accumulated data and the speed of getting new data do not leave a chance for interested people and companies to do data analysis in a manual mode. This makes the development of tools for the extraction of relevant data an important task. In this study, the author proposes an approach for sentiment analysis of short Russian text with vector representation. During the study, a self-prepared corpus of short Russian texts with 112 thousands units was used. The markup was made using markers. The efficiency of three algorithms was compared (decision tree, multilayer perception, logistic regression). The best model has an accuracy of classification equal to 76.2%, which is a high indicator of quality for the sentiment analysis task and thus allows using the approach in marketing research or monitoring audience loyalty to a particular topic or brand.

References

1. Baier, M. & Wagner, K. (2016) User Behavior in Crowdfunding Platforms – Exploratory Evidence from Switzerland. *Proceedings of Conference: Hawaii International Conference on System Sciences (HICSS)*. Kauai, Hawaii, USA. Vol. 49. pp. 3583–3593.

2. Poecze, F., Ebster, C. & Strauss, C. (2018) Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts. *Proceedings of The 9th International Conference on Ambient Systems, Networks and Technologies (ANT)*. Vol. 130. pp. 660–666.

3. Zeroual, I. & Lakhouaja, A. (2017) Data science in light of natural language processing: An overview. *Proceedings of The First International Confidence on Intelligent Computing in Data Science, ICDS*. Vol. 127. pp. 82–91.

4. Mayfield, A. (2008) *What Is Social Media?* [Online] Available from: https://www.icrossing.com/ uk/sites/default/files_uk/insight_pdf_files/What%20is%20Social%20Media_iCrossing_ebook .pdf. (Accessed 21.08.2018).

5. FOM. (2017) *Internet v Rossii: dinamika proniknoveniya* [Internet in Russian: dynamics of penetration]. [Online] Available from: https://fom.ru/SMI-i-internet/13585. (Accessed: 23.08.2018).

6. Shigina, Ya.I. & Fomenkov, D.A. (2014) Sotsial'nye media: sovremennye tendentsii v marketinge [Social media: modern tendencies in marketing]. *Vestnik Kazanskogo tekhnologicheskogo universiteta.* 17 (24). pp. 453–456.

7. Pear Analytics. (2009) *Twitter Study Reveals Interesting Results About Usage*. [Online] Available from: https://38r0us9g9l1438rwf2z2tcsz–wpengine.netdna–ssl.com/wp–content/ uploads/2009/08/Twitter–Study–August–2009.pdf. (Accessed: 21.08.2018).

8. Alexa Internet. (2018) *The top 500 sites on the web.* [Online] Available from: https://www.alexa.com/topsites. (Accessed: 23.08.2018).

9. Pang, B. & Lee, L. (2018) Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. 2. pp. 1–135.

10. Boudad, N. et al. (2018) Sentiment analysis in Arabic: A review of the literature. *Ain Shams Engineering Journal*. 9 (4). pp. 2479–2490.

11. Sokhin, T. & Butakov, N. (2018) Semi-automatic sentiment analysis based on topic modeling. *Proceedings of 7th International Young Scientists Conference on Computational Science, YSC2018.* Heraklion, Greece. Vol. 136. pp. 284–292.

12. Tartir, pp. & Abdul-Nabi, I. (2017) Semantic sentiment analysis in arabic social media. *Arabic Natural Language Processing: Models, Systems and Applications*. 29 (2). pp. 229–233.

13. Mallek, F., Belainine, B. & Sadat, F. (2017) Arabic Social Media Analysis and Translation. *Arabic Computational Linguistics,* 117. pp. 298–303.

14. Al-Thubaity, A., Alqahtani, Q. & Aljandal, A. (2018) Sentiment lexicon for sentiment analysis of Saudi dialect tweets. *Arabic Computational Linguistics*. 142. pp. 301–307.

15. Yusupova, N.I., Bogdanova, D.R. & Boyko, M.V. (2012) Algorithms and software for sentiment analysis of text messages based on machine learning. *Vestnik UGATU*. 16 (6). pp. 91–99. (In Russian).

16. Moussa, M., Mohamed, E. & Haggag, M. (2018) A survey on opinion summarization techniques for social media. *Future Computing and Informatics Journal*. 3 (1). pp. 82–109.

17. Amrani, Y., Lazaarb, M. & Kadiri, K. (2017) Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis. *Proceedings of The First International Confidence on Intelligent Computing in Data Science, ICDS*. Vol. 127. pp. 511–520.

18. Stieglitz, S. et al. (2018) Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*. 39. pp. 156–168.

19. Birjali, M., Beni-Hssane, A. & Erritali, M. (2017) Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks. *Proceedings of The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH)*. Vol. 113. pp. 65–72.

20. Araque, O., Zhu, G. & Iglesias, A. (2019) A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems.* 165. pp. 346–359.

21. Ankit, S.N. (2018) An Ensemble Classification System for Twitter Sentiment Analysis. *Proceedings of International Conference on Computational Intelligence and Data Science*. Vol. 132. pp. 937–946.

22. Yoon, K. (2014) *Convolution neural networks for sentence classification.* arXiv:1408.5882 [cs.CL]. [Online] Available from: https://arxiv.org/abs/1408.5882. (Accessed: 15.09.2018).

23. Heikal, M., Torki, M. & El-Makky, N. (2018) Sentiment Analysis of Arabic Tweets using Deep Learning. *Arabic Computational Linguistics*. 142. pp. 114–122.

24. Tweepy. (n.d.) Tweepy Documentation. [Online] Available from: https://tweepy.readthedocs.io/en/v3.5.0/index.html. (Accessed: 10.09.2018).

25. Jindal, S. & Sharma, K. (2018) Intend to analyze social media feeds to detect behavioral trends of individuals to proactively act against social threats. *Proceedings of International Conference on Computational Intelligence and Data Science*. Vol. 132. pp. 218–225.

26. Mikolov, T. Kai Chen, Corrado, G. & Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*.