УДК 004.04
DOI: http://dx.doi.org/10.21686/1818-4243-2023-4-60-71

Тапе Хабиб Жан Макс, А.А. Погуда
Национальный исследовательский Томский государственный университет, Томск, Россия

Сравнение методов анализа настроений глубокого обучения, включая LSTM и машинное обучение

Цель исследования. *Целью исследования является оценка определённых моделей машинного обучения при обработке данных на основе скорости и эффективности, связанных с анализом настроений или мнений потребителей в системе бизнес-аналитики. Для освещения уже имеющихся наработок дан обзор современных методов и моделей анализа настроений, демонстрируя их преимущества и недостатки.*
Материалы и методы. *С целью улучшения процесса анализа семестра, организованного с использованием существующих методов и моделей, необходимо внести в него корректировки в соответствии с растущими изменениями информационных потоков и на сегодняшний день. В этом случае исследователям крайне важно изучить возможности обновления определённых инструментов, либо объединить их, либо разработать, чтобы адаптировать их к современным задачам, чтобы обеспечить более чёткое понимание результатов их лечения. Мы представляем сравнение нескольких моделей глубокого обучения, включая конволюционная нейронная сеть, рекуррентные нейронные сети и долговременную и кратковременную двунаправленную память, оцененных на основе различных подходов к интеграции слов, включая трансформацию двунаправленных кодирующих представлений (BERT) и ее варианты, FastText и Word2Vec. Дополнение данных проводилось с использованием подхода простого дополнения данных.*
В этом проекте применяются методы обработки естественного языка (ОЕЯ), глубокое обучение, а также модели - LSTM, CNN, SVM TF-IDF, adaboost, naïves bayes, а затем комбинации моделей.
Результаты. *Исследования позволили получить и проверить результаты моделей с помощью пользовательских обзоров и сравнить точность моделей, чтобы увидеть, какая модель имеет наибольшую точность результатами анализа, полученными с помощью моделей, и их комбинацией CNN с LSTM-моделью, но SVM с TF-IDF векторизатором оказалась наиболее эффективной для этого несбалансированного набора данных. В построенной модели результатом стали следующие показатели: ROC AUC - 0,82, точность - 0,92, F1 - 0,82, Precision - 0,82 и Recall - 0,82. Для поиска более эффективной модели можно провести дополнительные исследования и внедрение модели.*
Заключение. *За последние годы анализ текста на естественном языке продвинулся довольно далеко вперёд, и не исключено, что в обозримом будущем подобные задачи будут полностью решены. Несколько различных моделей в ML и CNN с LSTM-моделью, но SVM с TF-IDF векторизатором оказалась наиболее эффективной для этого несбалансированного набора данных. В целом, как глубокое обучение, так и методы выбора на основе признаков могут быть использованы для решения некоторых наиболее актуальных проблем. Глубокое обучение полезно, когда наиболее значимые признаки заранее неизвестны, в то время как методы выбора на основе признаков могут помочь повысить точность и эффективность алгоритма классификации. Комбинация обоих подходов также может быть использована для дальнейшего повышения эффективности алгоритма.*

Ключевые слова: *BERT, LSTM, глубокое обучение, обработка естественного языка, TF-IDF, анализ настроений.*

Tape Habib Jean Max, Alexey A. Poguda
National Research Tomsk State University, Tomsk, Russia

Comparison of Deep Learning Sentiment Analysis Methods, Including LSTM and Machine Learning

Purpose of research. *The purpose of the study is to evaluate certain machine learning models in data processing based on speed and efficiency related to the analysis of sentiment or consumer opinions in business intelligence. To highlight the existing developments, an overview of modern methods and models of sentiment analysis is given, demonstrating their advantages and disadvantages.*
Materials and methods. *In order to improve the semester analysis process, organized using existing methods and models, it is necessary to adjust it in accordance with the growing changes in information flows today. In this case, it is crucial for researchers to explore the possibilities of updating certain tools, either to combine them or to develop them to adapt them to modern tasks in order to provide a clearer understanding of the results of their treatment. We present a comparison of several deep learning models, including convolutional neural networks, recurrent neural networks, and long-term and short-term bidirectional memory, evaluated using different approaches to word integration, including Bidirectional Encoder Representations from Transformers (BERT) and its variants, FastText and Word2Vec. Data augmentation was conducted using a simple data augmentation approach. This project uses natural language processing (NLP), deep learning, and models such as LSTM, CNN, SVM TF-IDF, Adaboost, Naive Bayes, and then combinations of models.*
The results of the study *allowed us to obtain and verify model results with user reviews and compare model accuracy to see which model had the highest accuracy results from the models and their combination of CNN with LSTM model, but SVM with TF-IDF vectoring was most effective for this unbalanced data set. In the constructed model, the result was the following indexes: ROC AUC - 0.82, precision - 0.92, F1 - 0.82, Precision - 0.82, and Recall - 0.82. More research and model implementation can be done to find a better model.*
Conclusion. *Natural language text analysis has advanced quite a bit in recent years, and it is possible that such problems will be completely solved in the near future. Several different models in ML and CNN with the LSTM model, but SVM with the TF-IDF vectorizer proved*

most effective for this unbalanced data set. In general, both deep learning and feature-based selection methods can be used to solve some of the most pressing problems. Deep learning is useful when the most relevant features are not known in advance, while feature-based selection methods can help improve the accuracy and efficiency of the classification algorithm. A combination of both approaches can also be used to further improve the efficiency of the algorithm.

Keywords: BERT, LSTM, deep learning, natural language processing, TF-IDF, sentimental analysis.

Введение

Классификация текстов в настоящее время является основной областью обработки естественного языка. Он имеет широкий спектр полезных приложений, например, фильтрацию спама, маркировку тем, анализ настроений и определение языка. Недавно эта проблема также появилась в медицинской сфере, где задача состоит в том, чтобы предсказать диагноз на основе описания самочувствия пациента. Поскольку задача относительно старая, существует множество различных подходов к решению этой проблемы. Мотивация этой статьи состоит в том, чтобы сравнить различные методы и прокомментировать их преимущества и недостатки. Конкретная проблема, с которой я решил работать, — это проблема анализа настроений. Для этого есть две основные причины. Первый из них - это большое количество открытых наборов данных с базовыми показателями. И второй - это широкий спектр коммерческих приложений, таких как мониторинг социальных сетей, маркетинговые исследования или обслуживание клиентов.

Большая часть этой работы была сосредоточена на тематической категоризации, пытаясь сортировать документы в соответствии с их темой. Однако в последние годы наблюдается быстрый рост онлайновых дискуссионных групп и сайтов обзоров, где важнейшей характеристикой опубликованных статей является их ощущение, или общее мнение по теме

– например, положительный или отрицательный отзыв о товаре. Маркировка этих статей с учетом их настроения могла бы предоставить читателям краткие резюме; действительно, эти метки являются частью привлекательности и дополнительной ценности некоторых сайтов, которые как маркируют обзоры фильмов, не содержащие явных показателей рейтинга, так и нормализуют различные схемы оценки, используемые отдельными критиками. Классификация настроений была бы также полезна в приложениях бизнес-аналитики [2] и рекомендательных системах, где комментарии и отзывы пользователей могут быть быстро обобщены; действительно, в целом, ответы на опросы в свободной форме, представленные в формате естественного языка, могут быть обработаны с использованием классификации настроений.

В последнее время глубокие нейронные сети стали наиболее популярными для решения задач классификации, поскольку они позволяют достичь наивысшей точности среди всех известных моделей машинного обучения. В частности, сверточные нейронные сети совершили прорыв в классификации изображений. В настоящее время они успешно справляются с некоторыми задачами автоматической обработки текста. Более того, как утверждается в некоторых исследованиях [3-6], сверточные сети подходят для этого даже лучше, чем рекуррентные нейронные сети, которые чаще всего используются для анализа текстовых последовательностей [7].

Существует множество исследований по установлению авторства [3–4]. В ранней работе [1] представлен подробный обзор исследований 2015–2021 гг., включая подходы на основе глубоких нейронных сетей (НС), классических методов машинного обучения (МО), аспектного анализа. В большинстве подобных публикаций применялись различные особенности стиля письма [5], включая лексические, синтаксические, структурные и специфические относительно жанра и тематики текста признаки. По состоянию на 2022 г. к моделям, успешно решающим смежные задачи текстового анализа, можно отнести LSTM, CNN, их гибриды, fastText, BERT.

При решении многих задач обработки естественного языка немало внимания уделяется качеству векторного представления текста. Созданная в 2016 г. библиотека fastText в реализации от Facebook [6] – серьезный шаг в развитии векторных семантических моделей и методов в обработке текста. например, преимущество fastText состоит в скорости работы по сравнению с другими моделями. Однако для определения авторства русскоязычных текстов fastText еще не применялся.

Анализ настроения обычно позволяет определить направленность настроения (т.е. положительное, нейтральное, отрицательное) текстовой информации, что может улучшить процессы принятия решений во множестве областей, включая бизнес, такой как финансы и фондовый рынок [8-11], цифровые платежные услуги [8], розничную торговлю [12, 13] и продукты [14, 15, 16], среди прочих. Ученые, исследующие анализ настроений на основе текстовых сообщений, также изучали или пытались определить рейтинги настроений, часто используя шкалы от 1 до 5 или 10 (т.е. более высокие баллы указывают на более

позитивные отзывы) [17]. Хотя часто для этого используются подходы машинного обучения, в последние годы глубокое обучение набирает обороты в анализе настроений, показывая многообещающие результаты [6, 17]. Кроме того, ученые исследовали различные методы встраивания слов, включая популярный Word2Vec и его варианты, а также более продвинутые и современные предварительно обученные модели на основе трансформаторов, такие как двунаправленные кодирующие представления из трансформаторов (BERT) [17-19], которые показали гораздо лучшие результаты при классификации текстов. Более того, в недавних обзорах представлены исследования, изучающие методы дополнения данных в алгоритмах глубокого обучения с супервизорами для улучшения предсказаний [20]. Эта техника, которая в целом является методом регуляризации, синтезирующим новые данные из существующих, широко используется в компьютерном зрении [20, 21]; однако работы, касающиеся текстовых данных, ограничены из-за сложности установления стандартных правил для автоматических преобразований текстовых данных при сохранении качества аннотаций [20, 22, 23], за исключением нескольких. Например, в [23] авторы исследовали различные методы предварительной обработки и регуляризации данных для анализа настроений вьетнамских пользователей в Twitter. Результаты показали, что увеличение объема данных является перспективным решением для повышения точности классификаторов. Для устранения выявленных выше пробелов данное исследование направлено на прогнозирование оценок отзывов покупателей с помощью моделей глубокого обучения на основе набора данных электронной коммерции, содержащего отзывы о женской одежде. В частности, это достигается путем предварительной обработки данных и увеличения объема данных для повышения вариативности набора данных.

Было рассмотрено несколько методов встраивания слов, включая Word2Vec, FastText, модель BERT и ее варианты (т.е. RoBERTa и ALBERT), чтобы определить лучший метод встраивания вместе с алгоритмами глубокого обучения. Затем были использованы несколько нейросетевых(NN) классификаторов, таких как рекуррентная нейронная сеть (RNN), конволюционная нейронная сеть (CNN) и двунаправленная долговременная память (Bi-LSTM), на двух различных установках, то есть 5 классов против 3 классов. Модели были оценены с помощью показателей производительности. кроме того, мы также проверили наши модели против нескольких алгоритмов машинного обучения, включая Naïve Bayes, Logistic Regression и Support Vector Machine (SVM) и др. Статья вносит вклад в обширный анализ различных известных моделей глубокого обучения наряду с более современными и продвинутыми вариантами BERT, чтобы определить лучшую модель предсказания отзывов о настроениях, используя как оригинальные, так и дополненные наборы данных.

Приложения глубокого обучения: Глубокая архитектура состоит из множества уровней нелинейных операций. Способность моделировать задачи сложного искусственного интеллекта позволяет ожидать, что глубокая архитектура будет хорошо работать в полу контролируемом обучении, таком как глубокая сеть убеждений (DBN), и достигнет значительного успеха в сообществе обработки естественного языка[24]. Глубокое обучение состоит из усовершенствованной программной инженерии, улучшенных процедур обучения и доступности вычислительных мощностей и обучающих данных [25]. Оно вдохновлено нейронаукой и имеет огромное влияние на ряд приложений, таких как распознавание речи, NLP (обработка естественного языка) и компьютерное зрение. Одной из основных проблем исследования глубокого обучения является способ изучения структуры модели, количества слоев и количества скрытых переменных для каждого слоя [26]. При работе с различными функциями архитектура глубокого обучения демонстрирует весь свой потенциал и требует большого количества помеченных образцов для сбора данных глубокой архитектурой. Сети и методы глубокого обучения широко применяются в различных областях, таких как визуальная классификация, обнаружение пешеходов, навигация внедорожных роботов, категории объектов, акустические сигналы и задачи прогнозирования временных рядов [27]. Мотивирующий подход в обработке естественного языка показал, что сложные многозадачные задачи, такие как семантическая маркировка, могут быть выполнены с использованием глубоких архитектур[27]. Что касается данных, глубокое обучение направлено на изучение высокоуровневых абстракций путем использования иерархических архитектур. Это многообещающий подход, который широко применяется в таких областях искусственного интеллекта, как компьютерное зрение, трансферное обучение, семантический разбор, обработка естественного языка и многих других. В наши дни глубокое обучение процветает по трем основным и важным причинам, а именно: улучшенные возможности чипов (GPU), значительно меньшие затраты на аппаратное обеспечение и значительное усовершенствование алгоритмов машинного обучения [28].

Важность анализа настроений

Поскольку люди выражают свои мысли и чувства более открыто, чем когда-либо прежде, анализ настроений быстро становится важным инструментом для мониторинга и понимания настроений во всех типах данных. Автоматический анализ отзывов клиентов, например, мнений в ответах на опросы и беседах в социальных сетях, позволяет брендам узнать, что вызывает у клиентов радость или разочарование, чтобы они могли адаптировать продукты и услуги к потребностям своих клиентов. Например, использование анализа настроений для автоматического анализа 4 000+ ответов в произвольной форме в ваших опросах об удовлетворенности клиентов может помочь вам узнать, почему клиенты довольны или недовольны на каждом этапе пути клиента. Возможно, вы хотите отслеживать настроение бренда, чтобы сразу же обнаруживать недовольных клиентов и реагировать на них как можно быстрее. Может быть, вы хотите сравнить настроения от одного квартала к другому, чтобы понять, нужно ли вам принимать меры [32]. Затем вы можете углубиться в качественные данные, чтобы понять, почему настроение падает или растет.

К общим преимуществам анализа настроений относятся:

Сортировка данных в масштабе Можете ли вы представить себе ручную сортировку тысяч твитов, разговоров в службе поддержки клиентов или опросов? Просто слишком много бизнес-данных, чтобы обрабатывать их вручную. Анализ настроений помогает компаниям обрабатывать огромные объемы неструктурированных данных эффективным и экономически выгодным способом.

Анализ настроений в режиме реального времени Анализ настроений позволяет выявлять критические проблемы в режиме реального времени, например, разрастается ли PR-кризис в социальных сетях? Разгневанный клиент вот-вот откликнется? Модели анализа настроений могут помочь вам немедленно определить подобные ситуации, чтобы вы могли сразу же принять меры.

Последовательные критерии - По оценкам специалистов, люди соглашаются между собой лишь в 60-65% случаев при определении настроения того или иного текста. Отметка текста по настроению очень субъективна, на нее влияют личный опыт, мысли и убеждения.

Используя централизованную систему анализа настроения, компании могут применять одни и те же критерии ко всем своим данным, что помогает им повысить точность и получить более глубокие знания. Применения анализа настроений безграничны. Поэтому, чтобы помочь вам понять, как анализ настроений может принести пользу вашему бизнесу, давайте рассмотрим несколько примеров текстов, которые можно проанализировать с помощью анализа настроений.

Затем мы рассмотрим реальный пример того, как Chewy, компания по продаже товаров для домашних животных, смогла получить гораздо более глубокое (и полезное!) понимание своих отзывов благодаря применению анализа настроений.

Методы анализа настроений

Анализ настроения – это способ автоматического извлечения субъективной информации из текста, такой как мнения или эмоции. Он используется в различных областях, таких как обслуживание клиентов, онлайн-обзоры и мониторинг социальных сетей. Анализ настроения может использоваться для выявления тенденций во мнениях клиентов, определения потребностей клиентов и помощи компаниям в реагировании на отзывы клиентов. Он также может использоваться для выявления потенциальных проблем с продуктами и услугами и для того, чтобы помочь компаниям определить возможности для улучшения. Кроме того, анализ настроений можно использовать для оценки влияния маркетинговых кампаний и выявления потенциальных областей улучшения обслуживания клиентов.

В данном разделе представлено краткое описание некоторых методов анализа настроений, изученных в данной работе. Эти методы являются наиболее популярными в литературе (т.е. наиболее цитируемыми и используемыми) и охватывают различные техники, такие как использование обработки естественного языка (NLP) для определения полярности, использование Amazon's Mechanical Turk (AMT) для создания маркированных наборов данных, использование психометрических шкал для определения настроений, использование контролируемых и неконтролируемых методов машинного обучения и т.п. Валидация этих методов также сильно варьируется, начиная от использования игрушечных примеров и заканчивая большой коллекцией помеченных данных.

Подход на основе машинного обучения является более практичным в чем другие подходы, благодаря своей полностью автоматической реализации и способности обрабатывать большие коллекции веб-данных. Методы классификации настроений на основе машинного обучения можно разделить на три типа: контролируемые, неконтролируемые и полуконтролируемые методы обучения [9].

1) Контролируемое обучение Контролируемое обучение

является зрелым и успешным решением в традиционной тематической классификации и было принято и исследовано для обнаружения мнений с удовлетворительными результатами [29]. Важными алгоритмами контролируемой классификации являются: Naïve Bayes, генеративный классификатор, который оценивает предварительные вероятности P(X|Y) и P(Y) из обучающих данных и генерирует апостериорную вероятность P(Y|X) на основе этих предварительных вероятностей; Support Vector Machine (SVM), дискриминативный классификатор, который не делает предварительных предположений на основе обучающих данных и напрямую оценивает P(Y|X); и алгоритм ленивого обучения K-Nearest Neighbors (KNN), который не требует предварительного построения модели классификации. Как в тематической классификации, так и в классификации мнений, Naïve Bayes и SVM являются наиболее распространенными и эффективными алгоритмами контролируемого обучения. Самое большое ограничение, связанное с контролируемым обучением, заключается в том, что оно чувствительно к количеству и качеству обучающих данных и может потерпеть неудачу, если обучающие данные необъективны или недостаточны. Обнаружение мнений на уровне поддокументов создает дополнительные проблемы для подходов, основанных на контролируемом обучении, поскольку для классификатора имеется мало информации.

2) Неконтролируемое обучение при классификации текстов иногда трудно создать маркированные обучающие документы, но легко собрать немаркированные документы. Методы обучения без надзора позволяют преодолеть эти трудности. Традиционные тематические модели, такие как LDA и PLSA, представляют собой неконтролируемые методы извлечения скрытых тем в текстовых документах. Темы - это признаки, а каждый признак (или тема) - это распределение по (признакам) терминам. Ограничение несамостоятельных подходов заключается в том, что для их точного обучения обычно требуется большой объем данных. Полностью несамостоятельные модели часто дают несогласованные темы, поскольку объективные функции моделей тем не всегда совпадают.

3) Полу самостоятельное обучение (SSL) Модели SSL отличаются от супервизорных и не супервизорных методов. В отличие от контролируемого обучения, которое учится только на меченых данных, SSL учится как на меченых, так и на немеченых данных. SSL: относительно новый подход машинного обучения к поиску мнений, мотивированный отсутствием маркированных данных в реальных приложениях. Согласно, основная идея SSL заключается в том, что, хотя немеченые данные не содержат информации о классах, они содержат информацию о совместном распределении признаков классификации. Поэтому, когда в целевой области данных мало меченых данных, использование SSL с немеченными данными может улучшить качество обучения по сравнению с контролируемым обучением. Также SSL не имеет ограничений, присущих подходам неконтролируемого обучения, если мы включаем некоторые формы предварительных знаний в неконтролируемые модели [10]. Согласно обзору SSL, проведенному в [11], Наиболее часто используемые алгоритмы SSL включают самообучение, генеративные модели, совместное обучение, обучение по нескольким курсам и методы, основанные на графах.

4) Подход на основе лексикона подход на основе лексикона основан на поиске лексикона мнений, который используется для анализа текста. В этом подходе есть два метода. Подход, основанный на словарях, предполагает поиск слов-семян мнений, а затем поиск в словаре их синонимов и антонимов. Подход, основанный на корпусе, начинается с начального списка слов мнений, а затем находит другие слова мнений в большом корпусе, чтобы помочь в поиске слов мнений с ориентацией на контекст. Это может

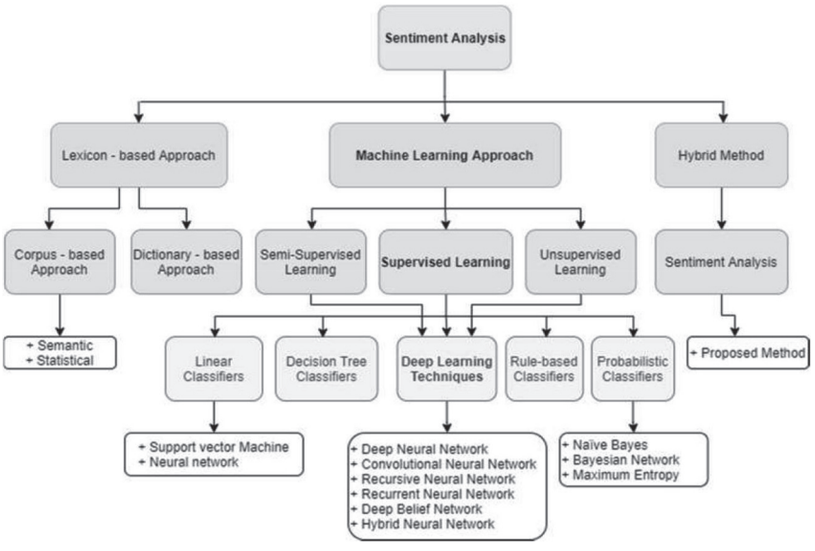

Рис. 1. Подходы и методы анализа настроения

Fig. 1. Approaches and methods of sentiment analysis

быть сделано с помощью статистических или семантических методов.

Методы анализа настроения на основе глубокого обучения используются для анализа настроения текстовых данных. Эти методы могут быть использованы для классификации текста на классы с положительным, отрицательным или нейтральным настроением. Некоторые из наиболее часто используемых методов глубокого обучения для анализа настроений включают рекуррентные нейронные сети (RNNs), сверточные нейронные сети (CNN) и долговременную кратковременную память (LSTM). RNN используются для фиксации временных связей между словами в предложении. CNN используются для определения пространственных отношений между словами в предложении. LSTM используются для исправления долгосрочных зависимостей между словами в предложении, каждый из методов имеет свои преимущества и недостатки. Например, RNN хорошо улавливают временные отношения между словами, но они не способны уловить долгосрочные зависимости между словами. С другой стороны, LSTM лучше улавливают долгосрочные зависимости между словами, но не могут уловить временные зависимости между словами.

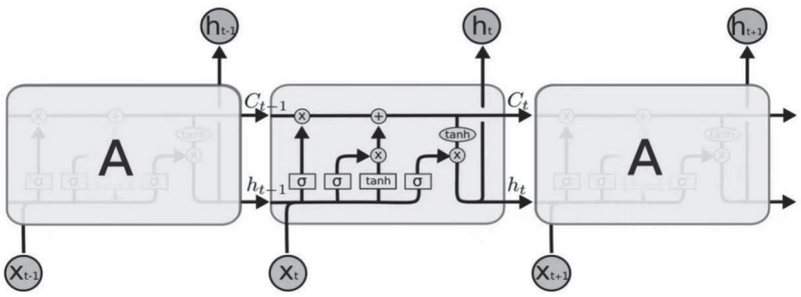

Рис. 2. Основная структура LSTM

Fig. 2. Basic structure of LSTM

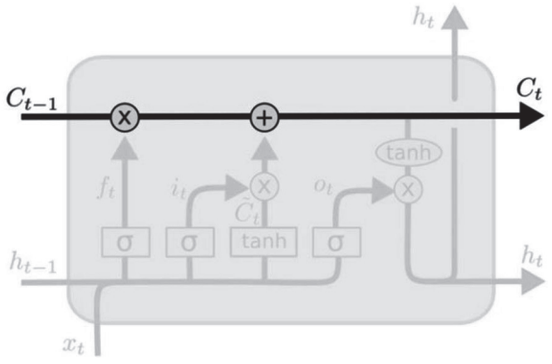

Рис. 3. Основная структура состояния клетки

Fig. 3. The basic structure of the state of the cell

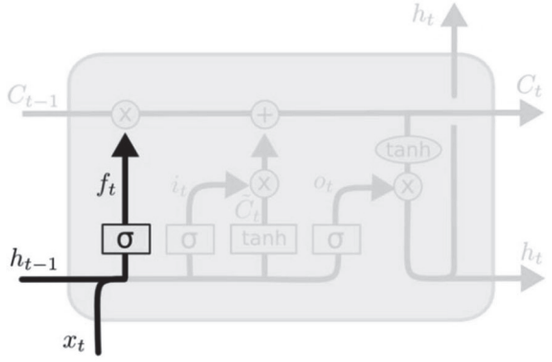

Рис. 4. Основная структура государства-забвения

Fig. 4. The basic structure of the oblivion state

Основной частью LSTM является самый верхний конвейер на рисунке 2. Эту часть принято называть состоянием клетки, которое показано на рисунке 2.22, и которое существует во всей цепной системе LSTM, и мы можем получить для нее формулу (1):

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C_t} \qquad (1)$$

f_t называется воротами забывания, указывающими, какие свойства, $C(t-1)$ используются для расчета C_t. Формула (2) представляет собой вектор, в котором каждый элемент находится в диапазоне формулы. Обычно мы используем сигмоид в качестве функции активации, а выход сигмоида - это значение в диапазоне [0,1]. Однако, если посмотреть на обученный LSTM, можно обнаружить, что значения ворот в основном очень близки к 0 или 1, а остальных очень мало. Среди них, ⊗ является наиболее важным механизмом затвора LSTM и представляет собой отношение умножения единиц между f_t и $C(t-1)$. Математическая формула (2) иллюстрируется следующим образом.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t]b_f) \qquad (2)$$

Как показано на рис. 5, $\tilde{C}_t$ представляет значение обновления состояния блока, которое получается из входных данных x_t и скрытого узла h_{t-1} через слой нейронной сети. Функция активации значения обновления состояния блока обычно использует *tanh*. i_t называется входным гейтом, который, как и f_t, является векто-

ром с элементами в интервале [0,1], и также вычисляется x_t и h_{t-1} через функцию активации сигмоида

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$\tilde{C}_t = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \quad (4)$$

it управляет тем, какая функция из $\tilde{C}_t$ используется для обновления C_t так же, как и f_t

Наконец, чтобы вычислить предсказанное значение $\hat{y}_t$ и сформировать полный вход для следующего временного среза, нам нужно вычислить выход скрытого узла h_t. Соответствующая математическая формула выглядит следующим образом.

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = O_t \times tanh(C_t) \quad (6)$$

Анализ настроений на основе глубокого обучения - это мощный инструмент для понимания естественного языка и извлечения информации из текста. Он может использоваться для обнаружения настроений, определения тем и извлечения значимой информации из неструктурированных данных. Наиболее популярными методами глубокого обучения для анализа настроений являются методы долговременной памяти (Long Short-Term Memory, LSTM) и двунаправленного кодирования представлений из трансформаторов (BERT). Это тип рекуррентной нейронной сети (RNN), которая способна изучать долгосрочные зависимости между словами в предложении. Она может быть использована для классификации настроений в тексте путем обучения модели на маркированных данных. Она хорошо подходит для задач анализа настроений, но ограничена в своей способности улавливать долгосрочные зависимости.

BERT – это мощная языковая модель, которая использует архитектуру трансформатора

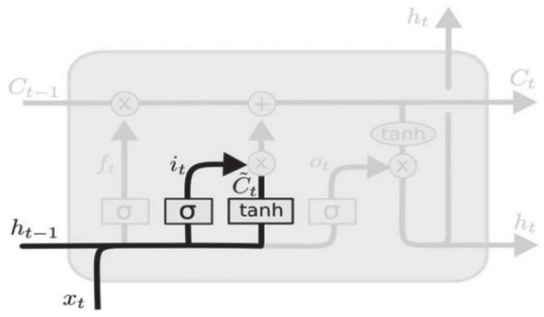

Рис. 5. Основная структура состояния обновления

Fig. 5. Basic structure of the renewal state

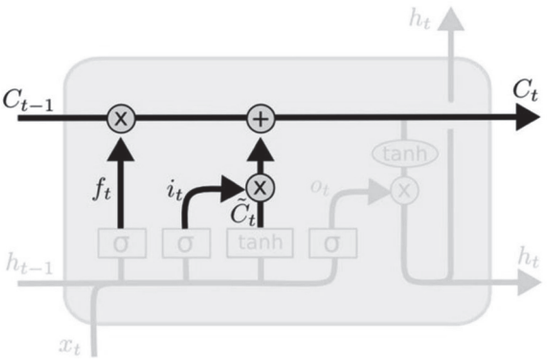

Рис. 6. Основная структура состояния выхода

Fig. 6. Main structure of exit status

для изучения контекстуальных связей между словами в предложении. Она способна понимать контекст предложения, что делает ее хорошо подходящей для задач анализа настроений. BERT также способен улавливать долгосрочные зави-

Таблица 1 (Table 1)

Сравнение методов анализа настроений

Comparison of sentiment analysis methods

Модель	Тип	Преимущества	Недостатки
VADER	Правило-базовая	Быстрая и эффективная, хорошо работает с данными из социальных сетей	Ограниченный словарный запас, не умеет обрабатывать сарказм и иронию
TextBlob	Правило-базовая	Проста и удобна в использовании, хорошо работает для базового анализа настроения	Ограниченная точность и не подходит для сложных задач
Naive Bayes	Вероятностная	Быстрая и эффективная, хорошо работает для коротких текстов	Требует большого и разнообразного набора данных для обучения
SVM	Машинное	Высокая точность и производительность, может обрабатывать большие наборы данных	Требует инженеринга функций и настройки
LSTM	Глубокое обучение	Высокая точность и производительность, может обрабатывать контекст и длинные тексты	Требует больших наборов данных для обучения и значительных вычислительных ресурсов
BERT	Глубокое обучение	Современная производительность, может работать с несколькими языками	Требует значительных вычислительных ресурсов и специализированного оборудования

Таблица 2 (Table 2)

Сравнение моделей для анализа настроений
Comparing models for sentiment analysis

Модель	Тип	Преимущества	Недостатки
VADER	Правило-базовая	Быстрая и эффективная, хорошо работает с данными из социальных сетей	Ограниченный словарный запас, не умеет обрабатывать сарказм и иронию
TextBlob	Правило-базовая	Проста и удобна в использовании, хорошо работает для базового анализа настроения	Ограниченная точность и не подходит для сложных задач
Naive Bayes	Вероятностная	Быстрая и эффективная, хорошо работает для коротких текстов	Требует большого и разнообразного набора данных для обучения
SVM	Машинное	Высокая точность и производительность, может обрабатывать большие наборы данных	Требует инженеринга функций и настройки
LSTM	Глубокое обучение	Высокая точность и производительность, может обрабатывать контекст и длинные тексты	Требует больших наборов данных для обучения и значительных вычислительных ресурсов
BERT	Глубокое обучение	Современная производительность, может работать с несколькими языками	Требует значительных вычислительных ресурсов и специализированного оборудования

симости, что делает его более эффективным, чем LSTM, для задач анализа настроений. В целом, и LSTM, и BERT являются мощными методами глубокого обучения для анализа настроений. Однако BERT более эффективен, чем LSTM, в улавливании долгосрочных зависимостей, что делает его лучшим выбором для сложных задач анализа настроений.

В целом, выбор метода анализа настроений зависит от вопроса исследования, характера данных и желаемого уровня детализации анализа. Методы на основе правил просты в реализации, но могут быть не такими точными, как методы на основе машинного обучения. Гибридные методы объединяют сильные стороны обоих подходов, а аспектный и сравнительный анализ настроений полезны для компаний, которые хотят понять настроение конкретных аспектов или сравнить свои продукты или услуги с продуктами и услугами конкурентов.

Отметим, что выбор модели будет зависеть от конкретных потребностей анализа, таких как размер набора данных, сложность связей между предложениями и доступные вычислительные ресурсы.

Этап эксперимента и результаты обработки данных

Выбор методов и моделей анализ критических ощущений при покупке товара на amazon с использованием моделей машинного обучения:

Мультиномиальный наивный байес с векторизатором подсчета

Multinomial Naive Bayes с векторизатором TF-IDF

SVM с векторизатором счета

SVM с векторизатором TF-IDF

Adaboost с векторизатором счета

Adaboost с векторизатором TF-IDF

Глубокое обучение: CNN и LSTM-модель

Экспериментальные данные

Этот набор данных включает почти 3000 отзывов покупателей Amazon (введенный текст), звезды, дату отзыва, вариант и комментарии различных продуктов Amazon Alexa, таких как Alexa Echo, Echo dots, Alexa Firesticks и т.д., чтобы узнать, как обучить машину анализировать настроения. Вопрос в том, что мы можем сделать с этими данными? Например, мы можем использовать эти данные для анализа продукта Alexa компании Amazon, обнаружить информацию о потребительских отзывах и помочь с моделями машинного обучения, а также обучить машинные модели для анализа настроений и проанализировать отзывы потребителей Сколько положительных отзывов? И сколько отрицательных отзывов?

Экспериментальный процесс

1. Сначала мы обрабатываем набор данных, сбрасывая заголовки 'sentiment' и 'text' для набора данных и отбрасывая бесполезные столбцы, как показано на рисунке 7.

2. Для обработки текста используется загруженный тезаурус английских стоп и извлечение английского корня (например, причастие настоящего времени английского языка станет основой словаря). Регуляризованное выражение также используется для обработки специальных символов в тексте. В процессе очистки данных прописные буквы также преобразуются в строчные для последующей обработки. Обработка одного фрагмента данных показана на рисунке 8.

В этом проекте будут применяться методы обработки естественного языка (NLP) для обнаружения крупномасштабных закономерностей среди письменных отзывов, оставленных покупателями на устройствах Alexa. Цель проекта – предсказать, понравился ли покупателям купленный ими товар, используя информацию, содержащуюся в их отзывах.

3. Мы делим набор данных на обучающее и тестовое множество в соответствии с определенной пропорцией. Слово "английский" является словом. Мы индексируем каждое слово и устанавливаем максимальную длину текста для обучения.

Таким образом, слова с похожими значениями будут иметь схожее векторное представление. Результат представления вектора слов в словах показан на рисунке 8

4. Мы строим модель, используем метод отсева, добавляем слой свертки, добавляем оптимизатор и задаем параметры для обучения данных.

5. Пишем тестовую функцию, вводим тестовый текст, сравниваем экспериментальные результаты с обученной моделью и выбираем оптимальную модель глубокого обучения.

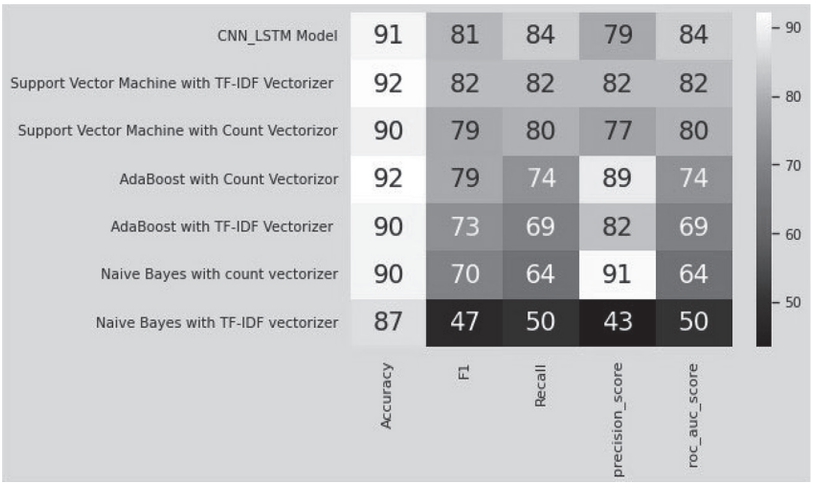

Рис. 7. Сравнение моделей

Fig.7. Comparison of models

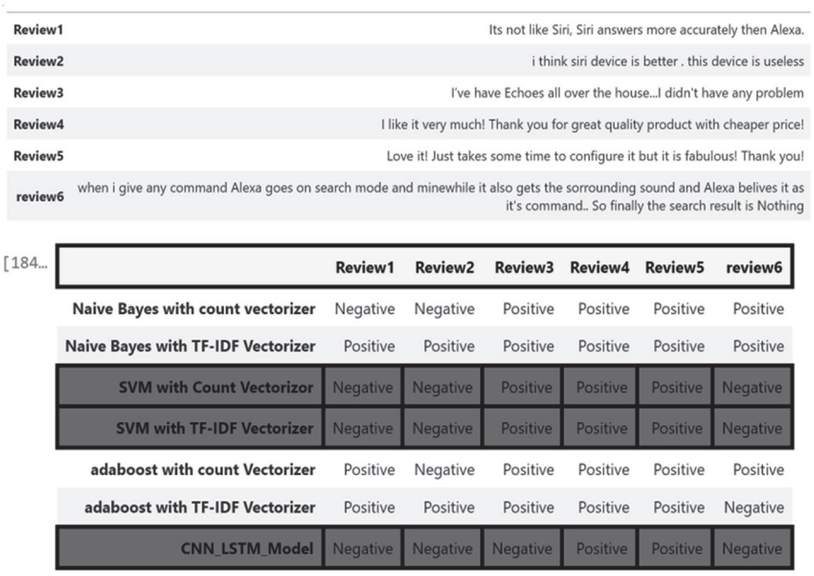

Рис. 8. Сравнение результатов моделей

Fig. 8. Comparison of model results

Результаты

В этой части я проверю результат моделей с помощью пользовательских обзоров и сравню точность моделей, чтобы увидеть, какая модель имеет самую высокую точность.

Здесь мы увидим результаты моделей с невидимыми данными, которые представляют собой случайные обзоры, взятые из Интернета, чтобы увидеть, как модели совершенствуют реальные данные.

В целом, модель CNN-LSTM представляет собой мощную архитектуру глубокого обучения, которая может обрабатывать сложные данные с высокой точностью и эффективностью.

Заключение

За последние годы анализ текста на естественном языке продвинулся довольно далеко вперед, и не исключено, что в обозримом будущем подобные задачи будут полностью решены. Несколько различных моделей в ML и CNN с LSTM-моделью, но SVM с TF-IDF векторизатором оказалась наиболее эффективной для этого несбалансированного набора данных. Наша модель показала ROC AUC 0,82, точность 0,92, F1 0,82, Pecision 0,82 и Recall 0,82. Для поиска более эффективной модели можно провести дополнительные исследования и внедрение модели. В целом, как глубокое обучение, так и методы выбора на основе признаков могут быть использованы для решения некоторых наиболее актуальных проблем. Глубокое обучение полезно, когда наиболее значимые признаки заранее неизвестны, в то время как методы выбора на основе признаков могут помочь повысить точность и эффективность алгоритма классификации. Комбинация обоих подходов также может быть использована для дальнейшего повышения эффективности алгоритма.

Литература

1. Романов А.С., Куртукова А.В., Соболев А.А. Определение возраста автора текста на основе глубоких нейросетевых моделей // Information. 2020. № 11(12). С. 589.

2. Шломо А. Э., Мошер К., Галит А. Классификация текста по стилю: какую газету я читаю? // В сборнике. Из семинара AAAI по категоризации текстов, 1998. С. 1–4.

3. Бай С., Колтер Дж.З, Колтун В. Эмпирическая оценка общих сверточных и рекуррентных сетей для моделирования последовательностей // Препринт arXiv arXiv. 2018. Т. 2. С. 1803–01271.

4. Конно А., Швенк Х., Барро Л. и др. Очень глубокие сверточные сети для классификации текстов // Препринт arXiv arXiv. 2017. Т. 2. С. 1606–01781.

5. Жанг Х., Чжао J., Лекун Ы. Сверточные сети символьного уровня для классификации текста // Препринт arXiv arXiv. 2016. Т. 3. С. 1509–01626.

6. Инь У., К. Каннан К. и др. Сравнительное исследование CNN и RNN для обработки естественного языка // Препринт arXiv arXiv. 2017. Т. 1. С. 1702.

7. Йогатама Д., Дайер Chr., Линг У. и др. Генеративная и дискриминативная классификация текстов с помощью рекуррентных нейронных сетей // Препринт arXiv arXiv. 2017. Т. 2. С. 1703–01898.

8. Балакришнан В., Лок П.Я., Рахим Х.А. Полууправляемый подход к выявлению настроений и эмоций на основе обзоров цифровых платежей // J Supercomput. 2021. Т. 77. С. 3795–3810.

9. Каросия А.Э., Коэльо Г.П., Сильва А.Э. Инвестиционные стратегии, применяемые к бразильскому фондовому рынку: методология, основанная на анализе настроений с использованием глубокого обучения // Приложение Expert Syst. 2021. Т. 184.

10. Цзин Н., Ву З., Ванг Х. Гибридная модель, интегрирующая глубокое обучение с анализом настроений инвесторов для прогнозирования цен на акции // Приложение Expert Syst. 2021. Т. 178.

11. Ядав А., Джа К.К., Шаран А. и др. Анализ настроений в финансовых новостях с использованием неконтролируемого подхода // Proced Comput Sci. 2020. Т. 167. С. 589–598.

12. Чжан Ю., Хан Р., Цзе М. и др. Аналитическая платформа социальных сетей для улучшения операций и управления услугами: исследование розничной аптечной индустрии // Технология прогнозирования изменений в Soc. 2021. Т. 163.

13. Ву Дж.Дж., Чанг С.Т. Изучение настроений потребителей в отношении онлайн-розничных услуг: тематический подход // J Retail Consumer. 2020. Т. 55. С. 102145.

14. Чжан Дж., Чжан А., Лю Д. и др. Извлечение предпочтений потребителей в отношении воздухоочистителей на основе детального анализа настроений онлайн-отзывов // Система, основанная на знаниях. 2021. Т. 228.

15. Сюй Ф., Пан З., Ся Р. Обзор продуктов электронной коммерции и классификация настроений на основе наивной системы непрерывного обучения Байеса // Управление процессами Inf. 2020. Т. 6(57).

16. Тапария А, Багла Т. Анализ настроений: прогнозирование оценок отзывов о товарах с использованием онлайн-отзывов покупателей. 2020. DOI: 10.2139/ssrn.3655308.

17. Колон-Руис С., Сегура-Бедмар И. Сравнение архитектур глубокого обучения для анализа настроений в отзывах о лекарствах // J Biomed Inform. 2020. Т. 110.

18. Ву Ф., Ши З., Донг З. и др. Анализ настроений онлайн-обзоров продуктов на основе SenBERT-CNN // Международная конференция 2020 по машинному обучению и кибернетике (ICMLC). 2020. С. 229–234.

19. Пота М., Вентура М., Кателли Р. и др. Эффективный конвейер на основе BERT для анализа настроений в Twitter: тематическое исследование на итальянском языке // Sensors. 2021. Т. 21(1). С. 133.

20. Шортен К., Хошгофтаар Т. М., Фурхт Б. Расширение текстовых данных для глубокого обучения // Big Data. 2021. Т. 8. С. 101.

21. Крижевский А., Суцкевер И., Хинтон Г.Е. Классификация Imagenet с использованием глубоких сверточных нейронных сетей // Commun ACM. 2017. С. 84–90.

22. Кобаяши С. Контекстуальная аугментация: приращение данных с помощью слов с парадигматическими отношениями // В NAACL HLT. 2018. Т. 2. С. 452–457.

23. Дуонг Х.Т., Нгуен-Тхи Т.А. Обзор: методы предварительной обработки и увеличение объема данных для анализа настроений // Вычислительная сеть. 2021. Т. 8. С. 1.

24. Чжоу С., Чен К., Ван Х. Метод активного глубокого обучения для классификации настроений под контролем пользователя // Нейрокомпьютинг. Т. 120. С. 536–546.

25. Дэн Л., Хинтон Г., Кингсбери Б. Новые типы глубокого обучения нейронных сетей для распознавания речи и связанных с ними приложений: обзор // IEEE Int. Конф. Акустика. Обработка речевого сигнала. 2013. С. 859–860.

26. Бенгио С., Денг Л., Ларошель Х., Салахутдинов Р.И. Введение приглашенных редакторов: специальный раздел по изучению глубоких архитектур // IEEE Trans Pattern Anal Mach Intell. 2013. Т. 35(8). С. 1795–1797.

27. Арнольд Л., Ребекки С., Шевалье С. и др. Введение в глубокое обучение // Esann. 2011. С. 479–488.

28. Го Y., Лю Ю., Эрлеманс А. и др. Глубокое обучение для визуального понимания: обзор // Нейрокомпьютинг.2016. Т. 187. С. 27–48.

29. Гуань З. Ян Дж. Сдержанное самообучение: метод классификации настроений под полуконтролем для китайских микроблогов // Материалы шестой Международной совместной конференции по обработке естественного языка. 2013. С. 455–462.

30. Чен З., Мукерджи А., Лю Б. Извлечение аспектов с автоматизированным изучением предварительных знаний // в трудах ACL. 2014. С. 347–358.

31. Пракаш В. Дж., Нитья Д. Л. Обзор методов обучения с полуконтролем // Международный журнал компьютерных тенденций и технологий. 2014. Т. 8(1). С. 25–29.

32. Руководство по анализу настроений [Электрон. ресурс]. Режим доступа: https://monkeylearn.com/sentiment-analysis/.

33. Основное руководство по анализу настроений [Электрон. ресурс]. Режим доступа: https://www.telusinternational.com/insights/ai-data/article/the-essential-guide-to-sentiment-analysis.

References

1. Romanov A.S., Kurtukova A.V., Sobolev A.A. Determination of the age of the author of the text based on deep neural network models. Information. 2020; 11(12): 589.

2. Shlomo A. E., Mosher K., Galit A. Text classification by style: what newspaper do I read? In the collection. From the AAAI Workshop on Text Categorization; 1998: 1-4.

3. Bay S., Kolter Dzh.Z, Koltun V. Empirical evaluation of general convolutional and recurrent networks for sequence modeling. Preprint arXiv arXiv. 2018; 2: 1803-01271.

4. Konno A., Shvenk KH., Barro L. et. al. Very deep convolutional networks for text classification. Preprint arXiv arXiv. 2017; 2: 1606-01781.

5. Zhang KH., Chzhao J., Lekun Y. Symbol-level convolutional networks for text classification. Preprint arXiv arXiv. 2016; 3: 1509-01626.

6. In' U., K. Kannan K. et. al. Comparative study of CNN and RNN for natural language processing. Preprint arXiv arXiv. 2017; 1: 1702.

7. Yogatama D., Dayyer Chr., Ling U. et. al. Generative and discriminative text classification using recurrent neural networks. Preprint arXiv arXiv. 2017; 2: 1703-01898.

8. Balakrishnan V., Lok P.YA., Rakhim KH.A. A semi-managed approach to detecting sentiment and emotion based on surveys of digital payments. J Supercomput. 2021; 77: 3795-3810.

9. Karosiya A.E., Koel'o G.P., Sil'va A.E. Investment Strategies Applied to the Brazilian Stock Market: A Methodology Based on Sentiment Analysis Using Deep Learning. Expert Syst Application. 2021: 184.

10. TSzin N., Vu Z., Vang KH. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. Expert Syst Application. 2021: 178.

11. Yadav A., Dzha K.K., Sharan A. et. al. Analysis of sentiment in financial news using an unsupervised approach. Proced Comput Sci. 2020; 167: 589-598.

12. Chzhan YU., Khan R., TSze M. et. al. Social media analytics platform for improving operations and service management: A study of the retail pharmacy industry. Change Prediction Technology in Soc. 2021: 163.

13. Vu Dzh.Dzh., Chang S.T. Exploring Consumer Sentiment for Online Retail Services: A Thematic Approach. J Retail Consumer. 2020; 55: 102145.

14. Chzhan Dzh., Chzhan A., Lyu D. et. al. Extracting consumer preference for air purifiers based on detailed sentiment analysis of online reviews. Knowledge Based System. 2021: 228.

15. Syuy F., Pan Z., Sya R. E-commerce Product Review and Sentiment Classification Based on Naive Bayesian Continuous Learning. Process Management Inf. 2020: 6(57).

16. Tapariya A, Bagla T. Sentiment Analysis: Predicting Product Review Scores Using Online Customer Reviews. 2020. DOI: 10.2139/ssrn.3655308.

17. Kolon-Ruis S., Segura-Bedmar I. Comparison of deep learning architectures for sentiment analysis in drug reviews. J Biomed Inform. 2020: 110.

18. Vu F., Shi Z., Dong Z. et. al. SenBERT-CNN Based Online Product Review Sentiment Analysis. International Conference on Machine Learning and Cybernetics (ICMLC). 2020: 229-234.

19. Pota M., Ventura M., Katelli R. et. al. Efficient BERT-based pipeline for Twitter sentiment analysis: a case study in Italian. Sensors. 2021; 21(1): 133.

20. Shorten K., Khoshgoftaar T. M., Furkht B. Text data extension for deep learning. Big Data. 2021; 8: 101.

21. Krizhevskiy A., Sutskever I., Khinton G.Ye Imagenet classification using deep convolutional neural networks. Commun ACM. 2017: 84–90.

22. Kobayashi S. Contextual Augmentation: Incrementing Data with Words with Paradigmatic Relationships. V NAACL HLT. 2018; 2: 452-457.

23. Duong KH.T., Nguyen-Tkhi T.A. Review: preprocessing methods and data augmentation for sentiment analysis. Computational Network. 2021; 8: 1.

24. Chzhou S., Chen K., Van KH. Active deep learning method for user-controlled mood classification. Neurocomputing. 120: 536-546.

25. Den L., Khinton G., Kingsberi B. New Types of Deep Learning Neural Networks for Speech Recognition and Related Applications: A Review. IEEE Int. Conf. Acoustics. Speech signal processing. 2013: 859-860.

26. Bengio S., Deng L., Laroshel' KH., Salakhutdinov R.I. Introduction by Guest Editors: A Special Section on the Study of Deep Architectures. IEEE Trans Pattern Anal Mach Intell. 2013; 35(8): 1795-1797.

27. Arnol'd L., Rebekki S., Sheval'ye S. et. al. Introduction to deep learning. Esann. 2011: 479-488.

28. Go Y., Lyu YU., Erlemans A. et. al. Deep learning for visual understanding: a review. Neurocomputing.2016; 187: 27-48.

29. Guan' Z. Yan Dzh. Restrained self-learning: a semi-supervised sentiment classification method for Chinese microblogging. Proceedings of the 6th International Joint Conference on Natural Language Processing. 2013: 455-462.

30. Chen Z., Mukerdzhi A., Lyu B. Aspect extraction with automated prior knowledge learning. In ACL Proceedings. 2014: 347-358.

31. Prakash V. Dzh., Nit'ya D. L. A review of semi-supervised learning methods. International Journal of Computer Trends and Technologies. 2014; 8(1): 25-29.

32. Guidance on sentiment analysis [Internet]. Available from: https://monkeylearn.com/sentiment-analysis/.

33. Basic guide to sentiment analysis [Internet]. Available from: https://www.telusinternational.com/insights/ai-data/article/the-essential-guide-to-sentiment-analysis.

Сведения об авторах

Жан Макс Тапе Хабиб
Аспирант, Факультет инновационных технологий
Национальный исследовательский Томский государственный университет, Томск, Россия
Эл. почта: Jeanmax.habib@mail.ru

Алексей Андреевич Погуда
К.т.н, доцент, Факультет инновационных технологий,
Национальный исследовательский Томский государственный университет, Томск, Россия
Эл. почта: alexsmail@sibmail.com

Information about the authors

Jean Max T. Habib
Postgraduate student, Faculty of Innovative Technologies
National Research Tomsk State University, Tomsk, Russia
E-mail: Jeanmax.habib@mail.ru

Alexey A. Poguda
Cand. Sci. (Engineering), Associate Professor, Faculty of Innovative Technologies,
National Research Tomsk State University, Tomsk, Russia
E-mail: alexsmail@sibmail.com