Научная статья

УДК 004.912 + 004.8
DOI 10.25205/1818-7900-2024-22-3-49-61

Сравнение методов машинного обучения для решения задачи анализа тональности

Милана Владимировна Швенк[1], Елена Павловна Бручес[1, 2]
Анна Яковлевна Леман[1]

[1]Новосибирский государственный университет
Новосибирск, Россия

[2]Институт систем информатики им. А. П. Ершова СО РАН
Новосибирск, Россия

m.shvenk@g.nsu.ru
bruches@bk.ru
a.leman@g.nsu.ru

Аннотация

Сеть «Интернет» позволяет делиться своим мнением со всем миром, и эти данные используются для анализа отношения общества к тому или иному субъекту. С течением времени эмоционально окрашенных текстов становится все больше. Современные технологии позволяют обрабатывать огромные массивы данных в кратчайшие сроки, что по-настоящему важно при реализации анализа тональности, который является одним из актуальных направлений автоматической обработки естественного языка. Нами был собран и размечен корпус текстов отзывов на медицинские услуги. Также были апробированы три способа решения задачи анализа тональности, относящиеся к методам традиционного или глубокого машинного обучения. Проведен сравнительный анализ полученных результатов. Размеченный нами корпус выложен в открытый доступ и может быть использован для других исследований.

Ключевые слова

обработка естественного языка, анализ тональности, машинное обучение, логистическая регрессия, сверточная нейронная сеть, LSTM

Для цитирования

Швенк М. В., Бручес Е. П., Леман А. Я. Сравнение методов машинного обучения для решения задачи анализа тональности // Вестник НГУ. Серия: Информационные технологии. 2024. Т. 22, № 3. С. 49–61. DOI 10.25205/1818-7900-2024-22-3-49-61

Comparison of Machine Learning Methods for Sentiment Analysis

Milana V. Shvenk[1], Elena P. Bruches[1, 2]
Anna Y. Leman[1]

[1]Novosibirsk State University,
Novosibirsk, Russian Federation

[2]A. P. Ershov Institute of Informatics Systems SB RAS,
Novosibirsk, Russian Federation

m.shvenk@g.nsu.ru
bruches@bk.ru
a.leman@g.nsu.ru

Abstract

Every day the amount of text data containing the subjective evaluation of the author is increasing thanks to the Internet. This information is used, for example, by numerous companies to assess the loyalty of their target audience. Due to the incredibly fast growth of the volume of such texts, their manual processing becomes impractical. It is in such situations that automated sentiment analysis is used, which is an actively developing area of natural language processing. We collected a corpus of medical service reviews, on the basis of which three classifiers were trained. We also performed a comparative analysis of the obtained results of the models, which belong to traditional or deep machine learning. Our corpus of texts is public and can be useful for other researchers.

Keywords

natural language processing (NLP), sentiment analysis, machine learning, logistic regression, convolutional neural network (CNN), LSTM

For citation

Shvenk M. V., Bruches E. P., Leman A. Y. Comparison of machine learning methods for sentiment analysis. *Vestnik NSU. Series: Information Technologies*, 2024, vol. 22, no. 3, pp. 49–61 (in Russ.) DOI 10.25205/1818-7900-2024-22-3-49-61

Введение

Большинство людей всегда интересуются точкой зрения своих знакомых для принятия решений, укрепления собственной позиции, сравнения взглядов и многого другого. Развитие сети «Интернет», а также ее возросшая доступность для обычного человека позволяют узнавать мнения и опыт людей, которые не являются ни нашими знакомыми, ни профессиональными критиками, ни даже жителями одной с нами страны. Соответственно, все больше людей делятся собственным мнением, увеличивая объем подобной информации, находящейся в открытом доступе.

Эти данные используются компаниями для определения лояльности потребителей к своему бренду/продукции/услугам; должностными лицами органов государственной власти [1]; инвесторами для выбора стратегии инвестирования [2] и т. д.

Анализ тональности текста – одно из актуальных направлений автоматической обработки естественного языка.

Человек способен прочитать несколько текстов и определить их тональность, в то время как программа за это же время обработает тысячи текстов, хоть и с относительно меньшей точностью.

В данной статье рассмотрены три способа решения задачи анализа тональности, принадлежащие традиционному и глубокому машинному обучению. Также проведено сравнение полученных результатов. Обучение и тестирование наших классификаторов проводилось на собранном и размеченном нами корпусе, который доступен по ссылке: https://github.com/m-sh-v/Sentiment_analysis/tree/main/corpus.

Обзор подходов к реализации анализа тональности

Основными методами решения задачи анализа тональности являются следующие три подхода.

Лингвистический. Используя словари эмоционально окрашенной лексики [3], подсчитывается количество пересечений их данных со словами обрабатываемого текста. После чего тексту присваивается тональность в зависимости от количества позитивной и негативной лексики. Однако данный метод имеет весомые недостатки [4]: качественное составление словаря требует огромных временных затрат; также автор-составитель должен разбираться в предметной области; один и тот же словарь нельзя использовать на разных предметных областях без потерь качества классификации.

Использование машинного обучения. Предобученная на заранее подготовленных данных модель используется для классификации новых текстов [5]. При выборе данного метода качество подготовки обучающих данных напрямую влияет на эффективность модели. Машинное обучение подразумевает под собой множество алгоритмов и способов решения задач и хорошо себя зарекомендовало при осуществлении анализа тональности [6; 7].

Гибридный. Совмещение двух вышеописанных методов, являясь наиболее трудоемким и ресурсозатратным, позволяет увеличить качество работы классификатора [8].

Нами были апробированы подходы на основе машинного обучения (сверточная нейронная сеть[1], LSTM) и гибридный метод (логистическая регрессия).

Составление и разметка собственного корпуса

Тематикой текстов нашего корпуса стали отзывы на медицинские услуги. При выборе направленности отзывов мы исходили из факта частого использования ярко эмоционально окрашенной лексики, а также мы старались избежать проблемы мультилингвальности [8], которая проявляется особенно сильно при обработке данных, взятых из социальных сетей. Например, анализ данных могут осложнять хештеги, написанные на языке, отличном от языка основного текста, а также хештег на одном языке может использоваться в постах носителей разных языков.

Сбор материала проводился с использованием открытого веб-ресурса «infodoctor»[2], который содержит обширную базу отзывов, каждому из которых автором комментария присвоена субъективная оценка (от 1 до 5). В нашей работе мы отбирали комментарии с оценками «1» и «5», которые содержат большее количество эмоционально окрашенной лексики. Выбранная нами предметная область позволяет относить даже условно нейтральные отзывы к классу позитивных, например, комментарий «Отправили на анализы. Сделали УЗИ», которому была присвоена оценка «5». Таким образом, для решения нашей задачи, т. е. определения тональности отзывов на медицинские услуги, мы можем ограничиться бинарной классификацией (негативный/позитивный).

Мы стремились сделать разметку нашего корпуса максимально полной, предвосхищая возможность его использования для решения задач из других сфер. При этом стоит отметить, что при обучении наших моделей мы использовали не все атрибуты разметки (см. табл. 1), исходя из специфики задачи анализа тональности.

Описание атрибутов корпуса: «text» – текст комментария; «doc» – отзыв на доктора (1 – да, 0 – нет); «service» – отзыв на услугу (1 – да, 0 – нет); «stars» – субъективная оценка автора комментария (от 1 до 5); «pos» – «позитивный» (1 – да, 0 – нет); «neu» – «нейтральный» (1 – да, 0 – нет); «neg» – «негативный» (1 – да, 0 – нет); «score» – объединение трех предыдущих

[1] Convolutional neural network (CNN).

[2] Сервис подбора врача и онлайн-записи на прием «ИнфоДоктор» основан в сентябре 2011 г., впервые анонсирован в 2012 г. (https://infodoctor.ru/).

атрибутов (1/0/–1); «area» – город расположения медицинского учреждения («spb» – Санкт-Петербург, «msk» – Москва, «ekb» – Екатеринбург, «nsk» – Новосибирск)

Таблица 1

Пример разметки корпуса на основе текста одного комментария

Table 1

An example of corpus markup based on the text of a comment

texts	doc	service	stars	pos	neu	neg	score	area
Ужасный косметолог, делала тредлифтинг и что в итоге, испортила мне лицо, кормит завтраками что скоро пройдет, как работать, как выходить на улицу, с мужем конфликт.	1	0	1	0	0	1	-1	spb

Всего было собрано 11779 текстов отзывов, 5899 из которых являются негативными, 5880 – позитивными. Все данные корпуса для обучения классификаторов были представлены в формате xml.

Впоследствии общее количество текстов было разделено на обучающую и тестовую выборки (см. табл. 2) методом удерживания[3], который подходит для большого объема данных [9].

Таблица 2

Распределение текстов корпуса

Table 2

Dataset samples distribution

	Обучающая выборка	Тестовая выборка
Негативные отзывы	5230	669
Позитивные отзывы	5186	694
Общее количество текстов	10416	1363

Качество разметки проверялось одной из наших моделей, т. е. мы использовали ее для предсказания классов тех же данных, на которых она была обучена: если присвоенная нами тональность текста не совпадала с тональностью, которую предсказала наш классификатор, то атрибуты данного текста отзыва были перепроверены вручную. Таким образом, было выявлено около 30 отзывов, которые подверглись исправлению. Ошибки при разметке были совершены из-за неправильно выставленной оценки авторов своим отзывам, т. е. человеческого фактора. Соответственно, после данной проверки качество разметки нашего корпуса возросло.

[3] Holdout. Подход состоит в разделении исходного набора данных случайным образом на два непересекающихся множества.

Традиционное машинное обучение и глубокое машинное обучение

Схематичное представление традиционного машинного обучения изображено на рис. 1.

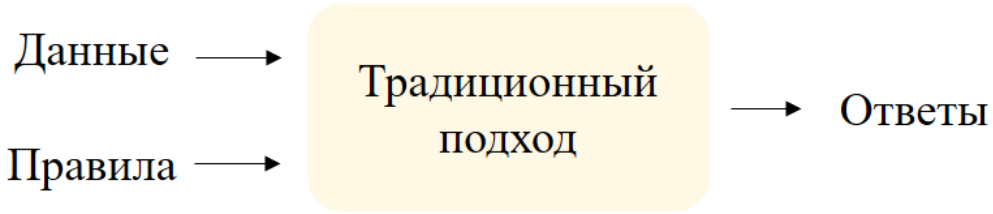

Рис. 1. Схема традиционного машинного обучения
Fig. 1. Scheme of traditional machine learning

Не все возникающие задачи возможно решить традиционными методами, которые основываются на правилах. Тогда в дело вступает глубокое машинное обучение [10].

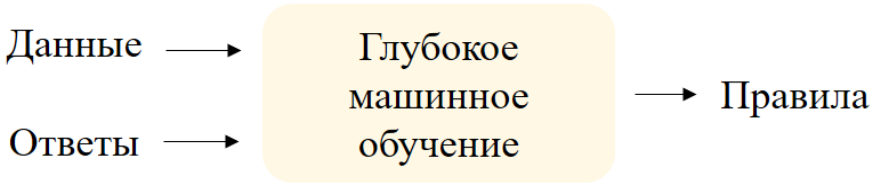

Рис. 2. Схема глубокого машинного обучения
Fig. 2. Scheme of deep machine learning

Пусть схема глубокого машинного обучения (см. рис. 2) и похожа на схему традиционного, ключевым ее отличием является выделение признаков (правил) без участия человека.

Этап предобработки текстов

Компьютер считывает информацию не так, как это делает человек, поэтому необходимо произвести предварительную подготовку данных, т. е. препроцессинг [11], который в нашем случае содержит следующие этапы обработки данных: очистка от числовых и специальных символов, которые не несут значения для классификации; удаление стоп-слов[4] [12]; понижение регистра [13]; токенизация[5] [14] (для нас токеном является слово); лемматизация[6], которая позволяет снизить нагрузку на классификатор [15]; векторизация[7] [14].

Также мы выбрали одну из ведущих библиотек нейронных сетей – Keras[8]. Данная библиотека позволяет проводить эксперименты с моделями нейронных сетей, обладая возможностью

[4] Стоп-слова – слова, не несущие важного семантического смысла для модели, например, служебные части речи, артикли, междометия, союзы. В русском языке к таким словам относятся: «и», «вы», «ах», «за» и т .п.

[5] Токенизация – разбиение строки на слова, фразы или символы.

[6] Лемматизация – приведение слов к их начальной форме.

[7] Векторизация – преобразование текстовых данных в числовые векторы.

[8] https://keras.io.

сократить время и усилия при прототипировании, так как наличие обширной базы уже созданных слоев, оптимизаторов и функций потерь дает возможность не тратить время на кодирование всех настроек вручную [16].

Логистическая регрессия

Наилучшие результаты среди методов традиционного машинного обучения показывают линейные модели, к которым относятся логистическая регрессия и метод опорных векторов [4]. В качестве представителя традиционного машинного обучения нами была выбрана логистическая регрессия, так как в разных исследованиях она показывает более высокие результаты [14; 17; 18].

Для выделения признаков принадлежности текста к определенному классу [19] мы использовали два метода: 1) подсчитывание количества эмоционально окрашенной лексики с помощью готового словаря RuSentiLex (РуСентиЛекс), записи в котором содержат ссылки на понятия РуТез[9]; 2) оценивание значимости слова в документе, т. е. метод TF-IDF[10], из которого следует, что токен, наиболее распространенный в одном тексте, но менее – в остальных, получает больший вес [20].

При использовании первого метода мы придерживались правила: если количество негативной лексики преобладало – отзыв негативный, во всех других случаях – отзыв позитивный.

После применения второго метода было выделено 20 % самых значимых токенов по результатам критерия хи-квадрата Пирсона.

На основе полученных признаков мы обучили классификатор и провели оценку качества его работы (табл. 3).

Таблица 3

Результаты оценки качества работы классификатора, основанного на логистической регрессии

Table 3

Logistic regression based classifier evaluation results

	precision	**recall**	**f1-score11**
Negative	0,9273	0,9925	0,9588
Positive	0,9923	0,9250	0,9575
		Weighted Average[12]	0,9582

Результаты показывают, что:

1) 93 % текстов, которые наша модель определила как негативные, действительно являются таковыми. Модель распознала 99 % всех негативных отзывов;

2) модель распознала 93 % всех позитивных текстов. Из них 99 % действительно оказались позитивными;

3) общая оценка эффективности классификатора составляет 96 %.

[9] Лингвистическая онтология РуТез (англ. RuThes) – тезаурус русского языка, представляющий собой иерархическую сеть понятий.

[10] Term frequency-inverse document frequency.

[11] F1-score – гармоническое среднее.

[12] Weighted average – тип среднего значения, придающий различную важность значениям в наборе данных.

Сверточная нейронная сеть

Наиболее зарекомендовавшими себя в исследуемой нами области методами, относящимися к глубокому машинному обучению, являются сверточные и рекуррентные[13] нейронные сети [4]. В нашей статье мы рассматриваем оба типа.

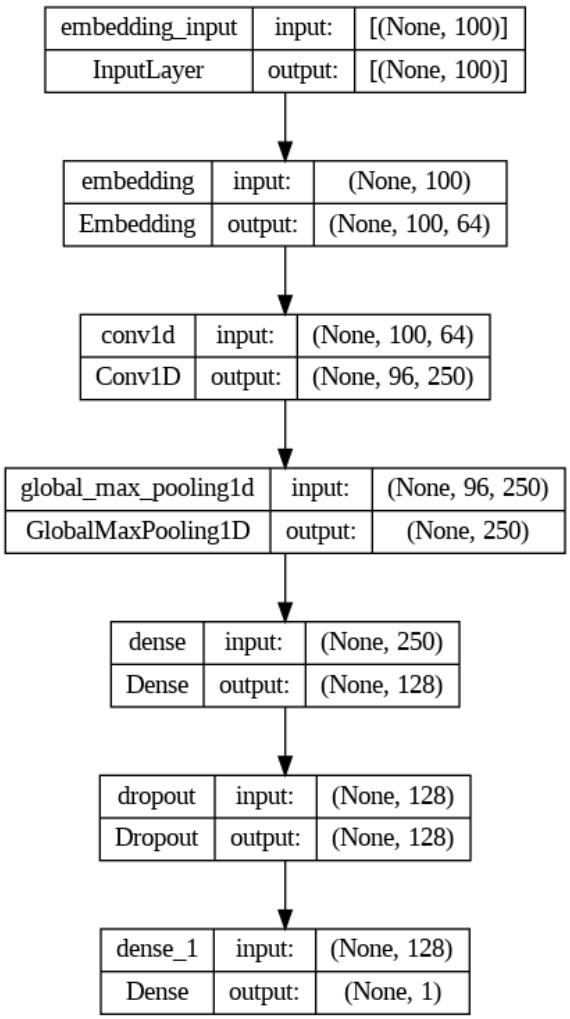

Рис. 3. Архитектура сверточной нейронной сети
Fig. 3. Convolutional neural network architecture

Архитектура нашей сверточной нейронной сети представлена на рис. 3. Наша модель состоит из шести слоев: embedding-слой, сверточный слой, субдискретизирующий слой, полносвязный слой, dropout-слой и еще один полносвязный слой.

При обучении нашего классификатора лучшие результаты были показаны на третьей эпохе обучения (табл. 4), поэтому мы сохранили модель с ее лучшими показателями и использовали именно ее в дальнейшем.

[13] Recurrent neural network (RNN).

Таблица 4

Показатели качества на каждой эпохе обучения классификатора на основе сверточной нейронной сети

Table 4

Quality scores at each epoch of CNN-based classifier training

Эпоха	**train_accuracy**[14]	**val_accuracy**[15]
1	0,8032	0,9155
2	0,9713	0,9702
3	0,9941	0,9731
4	0,9994	0,9568

После проверки классификатора на тестовой выборке мы получили результаты, приведенные в табл. 5.

Таблица 5

Результаты оценки качества работы классификатора, основанного на сверточной нейронной сети

Table 5

CNN-based classifier evaluation results

	precision	**recall**	**f1-score**
Negative	0,9719	0,9821	0,9770
Positive	0,9825	0,9726	0,9775
		Weighted Average	**0,9773**

Данные результаты показывают, что:

1) 97 % текстов, которые наша модель назвала негативными, действительно являются негативными; модель распознала 98 % от всего количества негативных отзывов;

2) модель распознала 97 % от общего количества позитивных текстов; 98 % комментариев, которые модель определила позитивными, действительно такими являются;

3) общая оценка эффективности классификатора равна 98 %.

LSTM

Также в нашей статье мы рассматриваем еще один тип нейронной сети – рекуррентный, который по многочисленным исследованиям является одним из лучших решений в вопросе обработки естественного языка [21].

Подобные нейронные сети способны использовать свою внутреннюю память для обработки серии событий и последовательностей произвольной длины [22]. Наличие обратных связей особенно важно для анализа тональности, поскольку текст необходимо рассматривать именно как последовательность [21].

[14] Показатели accuracy на обучающем наборе данных в процессе обучения.

[15] Показатели accuracy на валидационном наборе данных в процессе обучения.

Как бы хорошо все ни звучало, рекуррентные сети имеют весомый недостаток, названный проблемой исчезающего градиента[16] [23].

Именно для борьбы с данным явлением была создана особая архитектура рекуррентной нейронной сети – LSTM[17] [24].

Наша модель имеет четыре слоя (рис. 4): embedding-слой, два lstm-слоя и полносвязный слой.

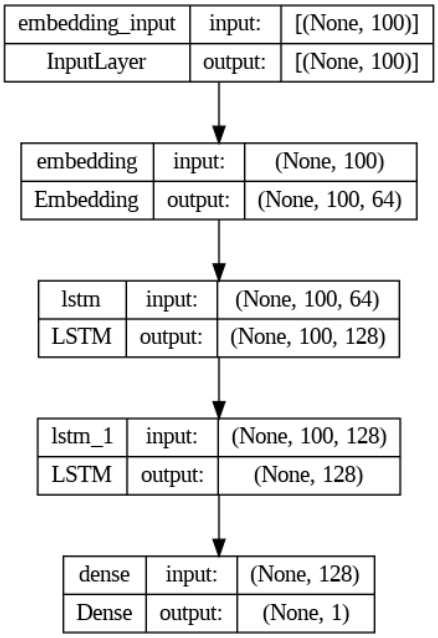

Рис. 4. Архитектура модели LSTM
Fig. 4. LSTM model architecture

Сохранив модель с ее лучшими показателями (табл. 6) в процессе обучения, мы провели оценку ее работы на тестовом наборе данных.

Таблица 6

Показатели качества на каждой эпохе обучения классификатора на основе LSTM

Table 6

Quality scores at each epoch of LSTM-based classifier training

Эпоха	train_accuracy	val_accuracy
1	0,8462	0,9280
2	0,9743	0,9530
3	**0,9931**	**0,9655**
4	0,9984	0,9530
5	0,9998	0,9616

После проверки модели на тестовой выборке мы получили результаты, приведенные в табл. 7.

[16] Vanishing gradients problem.

[17] Long Short Term Memory – «Долгая краткосрочная память».

Таблица 7

Результат оценки качества модели LSTM

Table 7

LSTM-based classifier evaluation results

	precision	**recall**	**f1-score**
Negative	0,9836	0,9821	0,9828
Positive	0,9827	0,9841	0,9834
		Weighted Average	**0,9831**

Данные результаты показывают:

1) 98 % текстов, которые наша модель назвала негативными, действительно являются негативными; модель распознала 98 % от всего количества негативных отзывов;

2) модель распознала 98 % от общего количества позитивных текстов; 98 % комментариев, которые модель определила позитивными, действительно такими являются;

3) общая оценка эффективности классификатора равна 98 %.

Анализ результатов

По результатам апробации трех вариантов решения задачи анализа тональности, основанных на применении традиционных и глубоких методов машинного обучения, были получены следующие результаты.

Представитель традиционного машинного обучения, логистическая регрессия, на которой основана наша первая модель, по оценке f1-score показала эффективность классификатора, равную 96 %.

Методы глубокого машинного обучения показали следующие результаты: классификатор, основанный на сверточной нейронной сети – 98 %; классификатор, построенный с использованием lstm-слоев – 98 %.

Стоит учесть, что оба классификатора, основанных на методах глубокого машинного обучения, показывают результат, равный 98 %, но при этом сверточная нейронная сеть подходит к данному показателю снизу (0,9773), в то время как модель LSTM – сверху (0,9831).

Заключение

В ходе работы мы собрали и разметили корпус, состоящий из текстов отзывов на медицинские услуги на русском языке. Мы описали используемые нами подходы к решению задачи анализа тональности, отметив их преимущества и отличия. Апробировали три способа реализации данного анализа и оценили качество работы классификаторов, каждый из которых можно использовать на практике для быстрой оценки большого объема данных, а конкретно для проведения анализа тональности отзывов на медицинские услуги, что позволит медицинским учреждениям отслеживать отношение пациентов к качеству предоставляемых услуг.

Таким образом, наша гипотеза подтвердилась: глубокое машинное обучение превосходит традиционный подход при решении задачи анализа тональности, а LSTM оказалась более эффективной моделью, чем другой представитель глубокого машинного обучения – сверточная нейронная сеть.

Список литературы

1. **Андросов А. Ю., Бородащенко А. Ю., Леонидова К. С.** Алгоритм определения тональности публикаций СМИ к должностным лицам государственных органов // Известия ТулГУ. Технические науки. 2020. № 2. С. 47–53.
2. **Ксенофонтов Г. С.** Тональный анализ новостей экономики // Скиф. Вопросы студенческой науки. 2022. № 5 (69). С. 584–589.
3. **Пазельская А. Г., Соловьев А. Н.** Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конференции «Диалог'2014». М.: Наука, 2014. С. 574–586.
4. **Самигулин Т. Р., Джурабаев А. Э. У.** Анализ тональности текста методами машинного обучения // Научный результат. Информационные технологии. 2021. Т. 6, № 1. С. 55–62. DOI: 10.18413/2518-1092-2021-6-1-0-7
5. **Пескишева Т. А.** Методы анализа тональности текстов на естественном языке // Общество. Наука. Инновации (НПК-2017). 2017. С. 1730–1742.
6. **Астапов Р. Л., Дубатов Р. С.** Классификация текстов с помощью сверточных нейронных сетей // Вестник науки. 2020. № 8 (29). С. 53–56.
7. **Воробьев Н. В., Пучков Е. В.** Классификация текстов с помощью сверточных нейронных сетей // Молодой исследователь Дона. 2017. № 6 (9). С. 2–7.
8. **Бодрунова С. С.** Кросс-культурный тональный анализ пользовательских текстов в Твиттере // Вестник Моск. ун-та. Серия 10. Журналистика. 2018. № 6. С. 191–212. DOI: 10.30547/vestnik.journ.6.2018.191212
9. **Шунина Ю. С.** Влияние способа формирования обучающей и тестовой выборок на качество классификации // Вестник УлГТУ. 2015. № 2 (70). С. 43–46.
10. **Скрипачев В. О., Гуйда М. В., Гуйда Н. В., Жуков А. О.** Особенности работы сверточных нейронных сетей // International Journal of Open Information Technologies. 2022. № 12. С. 5361.
11. **Майоров Д. В.** Применение глубокого обучения в предиктивном вводе // Научный журнал. 2023. № 2 (67). С. 19–25.
12. **Muhamediyeva D. K., Abdurakhmanova N. N., Mirzayeva N. S.** Using conventional neural networks for the problem of text classification // Central Asian Academic Journal of Scientific Research. 2021. No. 1. p. 213–220.
13. **Abinash Tripathy.** Sentiment Analysis Using Machine Learning Techniques: dissertation // Department of Computer Science and Engineering National Institute of Technology Rourkela. 2017. p. 131.
14. **Бородин А. И., Вейнберг Р. Р., Литвишко О. В.** Методы обработки текста при создании чат-ботов // Хуманитарни Балкански изследвания. 2019. № 3 (5). С. 108–111. DOI: 10.34671/SCH.HBR.2019.0303.0026
15. **Сафаров Л. С.** Использование технологии text mining при автоматической обработке текста // Экономика и социум. 2023. № 1−2 (104). С. 639–642.
16. **Бербасов В. Д.** Сравнительный обзор библиотек нейронных сетей Keras и Pytorch // Экономика и социум. 2023. № 8 (111). С. 423–426.
17. **Васильченко А. М.** Решение задач анализа данных на основе машинного обучения // Universum: технические науки. 2023. № 9−1 (114). С. 50–54. DOI: 10.32743/UniTech.2023.114.9.15959
18. **Шунина Ю. С., Алексеева В. А., Клячкин В. Н.** Критерии качества работы классификаторов // Вестник УлГТУ. 2015. № 2 (70). С. 67–70.
19. **Bing Liu.** Sentiment Analysis and Opinion Mining // Morgan & Claypool Publishers. 2012. P. 168.

20. **Ткаченко А. Л.** Решение задачи классификации документов вуза на основе методов интеллектуального анализа // Вестник кибернетики. 2021. № 1 (41). С. 12–19. DOI: 10.34822/1999-7604-2021-1-12-19
21. **Гальченко Ю. В., Нестеров С. А.** Классификация текстов по тональности методами машинного обучения // SAEC. 2023. № 3. С. 369–378. DOI: 10.18720/SPBPU/2/id23-501
22. **Юркина А. В., Крутиков А. К.** Разработка специализированного программного модуля формирования обучающей выборки для нейрогороскопа // Universum: технические науки. 2023. № 10-1 (115). С. 61–65.
23. **Hochreiter S., Bengio Y., Frasconi P. Schmidhuber J.** Gradient flow in recurrent nets: the difficulty of learning long-term dependencies // IEEE Press. 2001. no. 1. p. 15.
24. **Пустынный Я. Н.** Решение проблемы исчезающего градиента с помощью нейронных сетей долгой краткосрочной памяти // Инновации и инвестиции. 2020. № 2. С. 130–132.

References

1. **Androsov A. Y., Borodaschenko A. Y., Leonidova K. S.** Algorithm for determining the tone of media publications to public officials. *Izvestia TulSU. Engineering Sciences*, 2020, no. 2, pp. 4753. (in Russ.)
2. **Ksenofontov G. S.** Tonal analysis of economic news. *Skif. Student Science Issues,* 2022, no. 5 (69), pp. 584–589. (in Russ.)
3. **Pazelskaya A. G., Solovyov A. N.** The method of determining emotions in texts in Russian. *Computer Linguistics and Intelligent Technologies: Proceedings of the International Conference Dialogue'2014*. Moscow, Science publ., 2014, pp. 574–586. (in Russ.)
4. **Samigulin T. R. Djurabaev A. E. U.** Sentiment analysis of text by machine learning methods. *Research Result. Information Technologies*, 2021, vol. 6, no. 1, pp. 55–62. (in Russ.) DOI: 10.18413/2518-1092-2021-6-1-0-7
5. **Peskisheva T. A.** Methods for sentiment analysis of natural language texts. *Society. Science. Innovations (Scientific and practical conference 2017)*, 2017, pp. 1730–1742. (in Russ.)
6. **Astapov R. L., Dubatov R. S.** Classification of texts using convolutional neural networks. *Vestnik of Science*, 2020, no. 8 (29), pp. 53–56. (in Russ.)
7. **Vorobev N. V., Puchkov E. V.** Classification of texts using convolutional neural networks. *Molodoj issledovatel' Dona*, 2017, no. 6 (9), pp. 2–7. (in Russ.)
8. **Bodrunova S. S.** Cross-cultural sentiment analysis of user texts on twitter. *Vestnik of Moscow University. Series 10. Journalism*, 2018, no. 6, pp. 191–212. (in Russ.) DOI: 10.30547/vestnik.journ.6.2018.191212
9. **Shunina U. S.** Dependence of classification quality on the methods to form a training and test samples. *Vestnik UlSTU*, 2015, no. 2 (70), pp. 43–46. (in Russ.)
10. **Skripachyov V. O., Guyda M. V., Guyda N. V., Zhukov A. O.** Features of convolutional neural networks. *International Journal of Open Information Technologies*, 2022, no. 12, pp. 53–61. (in Russ.)
11. **Mayorov D. V.** Applying deep learning in predictive input. *Scientific Journal*, 2023, no. 2 (67), pp. 19–25. (in Russ.)
12. **Muhamediyeva D. K., Abdurakhmanova N. N., Mirzayeva N. S.** Using conventional neural networks for the problem of text classification. *Central Asian Academic Journal of Scientific Research*, 2021, no. 1, pp. 213–220.
13. **Abinash Tripathy.** Sentiment Analysis Using Machine Learning Techniques: dissertation. *Department of Computer Science and Engineering National Institute of Technology Rourkela*, 2017, p. 131.

14. **Borodin A. I., Veynberg R. R., Litvishko O. V.** Methods of text processing when creating chatbots. *Humanitarian Balkan Studies*, 2019, no. 3 (5), pp. 108–111. (in Russ.) DOI: 10.34671/SCH.HBR.2019.0303.0026
15. **Saifarov L. S.** Using text mining technology in automatic text processing. *Economics and Society*, 2023, no. 1−2 (104), pp. 639–642. (in Russ.)
16. **Berbasov V. D.** Comparative review of keras and pytorch neural network libraries. Economics and Society, 2023, no. 8 (111), pp. 423–426. (in Russ.)
17. **Vasilchenko A. M.** Solving data analysis problems based on machine learning. *Universum: Engineering Sciences*, 2023, no. 9−1 (114), pp. 50–54. (in Russ.) DOI: 10.32743/UniTech.2023.114.9.15959
18. **Shunina U. S., Alekseeva V. A., Klyachkin V. N.** Criteria of quality of qualifiers work. *Vestnik UlSTU*, 2015,no. 2 (70), pp. 67–70. (in Russ.)
19. **Bing Liu.** Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012, p. 168.
20. **Tkachenko A. L.** Solving the problem of university documents classification based on intellectual analysis methods. *Vestnik of Cybernetics*, 2021, no. 1 (41), pp. 12–19. DOI: 10.34822/1999-7604-2021-1-12-19 (in Russ.)
21. **Galchenko Y. V., Nesterov S. A.** Sentiment analysis with machine learning methods. SAEC, 2023, no. 3, pp. 369–378. DOI: 10.18720/SPBPU/2/id23-501 (in Russ.)
22. **Yurkina A. V., Krutikov A. K.** Development of a specialized software module for the formation of a training sample for a neurogoroscope. *Universum: Engineering Sciences*, 2023, no. 10−1 (115), pp. 61–65. (in Russ.)
23. **Hochreiter S., Bengio Y., Frasconi P., Schmidhuber J.** Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. *IEEE Press*, 2001, no. 1, p. 15.
24. **Pustynnyj I. N.** Solving the problem of vanishing gradient using long short term memory neural networks. *Innovations and Investments*, 2020, no. 2, pp. 130–132. (in Russ.)

Сведения об авторах

Швенк Милана Владимировна, бакалавр

Бручес Елена Павловна, младший научный сотрудник, старший преподаватель

Леман Анна Яковлевна, старший преподаватель

Information about the Authors

Milana V. Shvenk, Bachelor

Elena P. Bruches, Junior Researcher; Senior Lecturer

Anna Y. Leman, Senior Lecturer

Статья поступила в редакцию 16.07.2024;
одобрена после рецензирования 08.10.2024; принята к публикации 08.10.2024

The article was submitted 1.07.2024;
approved after reviewing 08.10.2024; accepted for publication 08.10.2024