УДК 004.852, 004.62
doi:10.18720/SPBPU/2/id23-501

Гальченко Юлия Вадимовна [1],
студент магистратуры;
Нестеров Сергей Александрович [2],
доцент, канд. техн. наук, доцент

КЛАССИФИКАЦИЯ ТЕКСТОВ ПО ТОНАЛЬНОСТИ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

[1, 2] Россия, Санкт-Петербург, Санкт-Петербургский политехнический университет Петра Великого,
[1] artyuhova.yuv@edu.spbstu.ru, [2] nesterov@spbstu.ru

Аннотация. В данной статье рассмотрены существующие методы классификации текстов по тональности, создана модель нейронной сети, которая успешно решает поставленную задачу классификации. Нейросеть была обучена на наборах данных, которые включают отзывы о различных услугах и местах, а также рецензии на фильмы. Полученная модель нейросети показала высокий результат в задаче классификации текстов по тональности на тестовых данных.

Ключевые слова: классификация, машинное обучение, нейронные сети, сеть LSTM, методы классификации текстов, тональность текста, интеллектуальный анализ данных.

Yuliia V. Galchenko [1],
Master's Student;
Sergey A. Nesterov [2],
Associate Professor, PhD in Technical Sciences

SENTIMENT ANALYSIS WITH MACHINE LEARNING METHODS

[1, 2] Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russia,
[1] artyuhova.yuv@edu.spbstu.ru, [2] nesterov@spbstu.ru

Abstract. In this article, the existing methods for sentiment analysis were considered, a neural network model that successfully solves the task of classification was created. The neural network was trained on datasets that include reviews of various services and

places, as well as movie reviews. The resulting neural network model showed a high result in the task of sentiment analysis on test data.

Keywords: classification, machine learning, neural networks, LSTM network, text classification methods, text sentiment, data mining.

Введение

В современном мире задача классификации текстов получила большое распространение вследствие увеличения объема текстовой информации в мировом информационном пространстве. В связи с этим возникает потребность в системах обработки текстовой информации, её хранения и анализа.

Предположим, имеется некоторый текст, который может представлять собой сообщение, отзыв или комментарий. Задачей является определить, какую эмоцию несет в себе этот текст: симпатия, разочарование, восторг, недовольство, сомнение и т. п. В общем случае эмоции можно классифицировать на *позитивные* и *негативные* (положительные и отрицательные), поэтому в работе рассматривается именно задача бинарной классификации.

В статье кратко рассмотрены существующие методы классификации текстов по тональности, разработана функционирующая модель для классификации текстов по тональности, а также проведены экспериментальные исследования на тестовых данных.

1. Постановка задачи и обзор методов классификации текстов

Анализ тональности текста является подзадачей обработки естественного языка (Natural Language Processing, NLP), цель которой — классификация текста в соответствии с эмоциональной окраской, которую он в себе несет [1].

Существуют две основные группы подходов к анализу тональности текстов: подходы на основе правил (лингвистические) и подходы на основе машинного обучения. Подходы на основе машинного обучения (с учителем) более универсальны и не требуют создания словарей и правил для конкретной предметной области, поэтому в данной статье рассматриваются именно такие методы.

Задача классификации текстов по тональности формализуется следующим образом: необходимо построить модель для классификации F, которая после обучения на выборке D определяет текст T_i к одному из классов множества y — то есть к отрицательному или положительному классу текстов:

$$F\{D,T_i\} \to \{y_1, y_2\}. \quad (1)$$

Поставленная задача классификации может решаться с помощью различных методов машинного обучения, а также с помощью нейронных

сетей [2]. В последние годы все чаще используются методы глубокого обучения (Deep Learning), к которым относятся нейронные сети. Такие методы могут значительно превосходить классические методы в задаче анализа тональности текстов [3]. Среди различных видов нейронных сетей класс рекуррентных нейросетей зачастую превосходит другие классы в рассматриваемой задаче [1].

Рекуррентные сети имеют обратные связи, и для вычисления текущего состояния они используют предыдущие состояния [4]. Это важно для анализа тональности текста, ведь при определении эмоциональной окраски текста важно анализировать текст именно как последовательность. Рекуррентные нейронные сети имеют весомый недостаток: при каждой итерации информация в памяти смешивается с новой информацией, а после нескольких итераций полностью перезаписывается. Такая проблема называется проблемой исчезающего градиента (vanishing gradient problem) [5].

Архитектура рекуррентной нейронной сети, которая позволяет уменьшить проблему исчезающего градиента — «Долгая краткосрочная память» (от англ. Long Short Term Memory — LSTM) [6]. LSTM-сети нередко применяются в задаче классификации текстов и, в частности, классификации текстов по тональности. Такая архитектура зачастую показывает достаточно высокие результаты по сравнению как с другими методами машинного обучения, так и с другими архитектурами нейронных сетей [7, 8], поэтому основу созданной модели составляет LSTM-сеть.

2. Сеть LSTM — «Долгая краткосрочная память»

В сетях LSTM элементом сети является набор слоёв, взаимодействующих друг с другом по определённым правилам. Подобные наборы называются ячейками. Структура LSTM сети, развернутой во времени, представлена на рисунке 1.

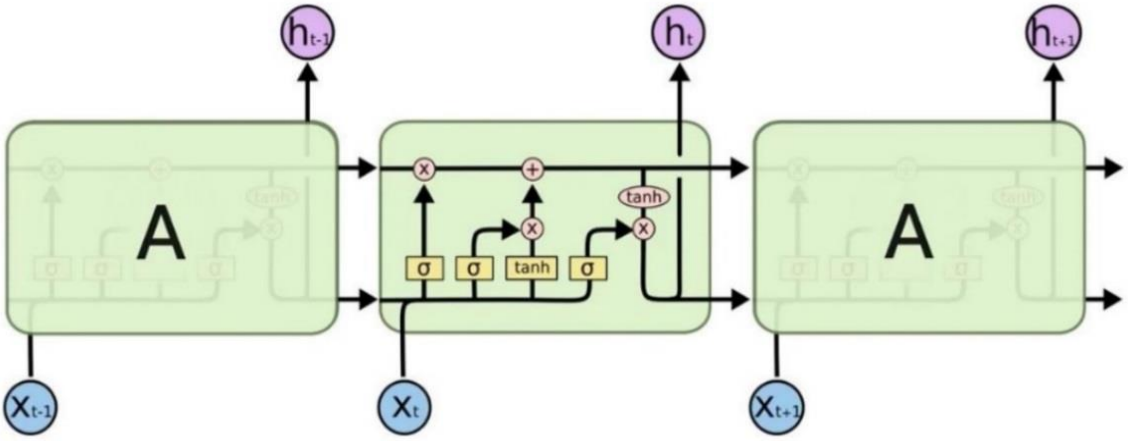

Рис. 1. Сеть LSTM, развернутая во времени

На вход сети в разные моменты времени поступают элементы последовательности x_{t-1}, x_t, x_{t+1}, и в каждый момент времени сеть выдает значения h_{t-1}, h_t, h_{t+1}. Такая сеть передает два значения на вход своей копии в следующий момент времени. Работа ячейки описывается набором формул (2):

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f h_{t-1}); \\ i_t &= \sigma(W_i x_t + U_i h_{t-1}); \\ \tilde{C}_t &= \tanh(W_c x_t + U_c h_{t-1}); \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t; \\ o_t &= \sigma(W_0 x_t + U_0 h_{t-1}); \\ h_t &= o_t * \tanh(C_t); \end{aligned} \qquad (2)$$

где f_t — выходной вектор вентиля забвения, i_t — выходной вектор входного вентиля, $\tilde{C}_t$ — вектор новых значений-кандидатов, которые можно добавить в состояние ячейки, C_t — новое состояние ячейки, o_t — вектор выходного вентиля. Более подробно работа данной архитектуры описана в книге [9].

3. Создание и обучение нейросети

3.1. Используемые средства и наборы данных

Для решения поставленной задачи использовался язык программирования Python, а также библиотеки TensorFlow и Keras.

В качестве наборов данных для обучения и тестирования были использованы YELP [10] и Large Movie Review Dataset (IMDB) [11]. Набор данных Yelp reviews Polarity содержит отзывы о различных услугах, сервисах и местах на английском языке. Набор IMDB содержит рецензии на различные фильмы на английском языке.

Говоря об анализе эмоциональной окраски текстов, важно учитывать, что тексты в интернете могут иметь разную специфику: например, нейронная сеть, обученная на наборе данных с рецензиями на книги, может плохо справляться с определением тональности комментариев в социальной сети. Это связано с тем, что такие тексты имеют ряд различий: у них может значительно отличаться длина (количество слов), лексика (более формальная, либо же более свободная с использованием сленга и аббревиатур), и т. д. Поэтому с целью создания наиболее универсальной модели нейронной сети для ее обучения было принято решение использовать тренировочный набор, составленный из отзывов набора YELP и набора IMDB в совокупности. После слияния тренировочных наборов было получено 604000 отзывов для обучения нейронной сети, структура полученного набора представлена на рисунке 2.

Рис. 2. Структура тренировочного набора

3.2. Предварительная обработка данных

К этапу предобработки данных на естественном языке относятся: очистка данных (удаление из исходного текста особых знаков, символов, пунктуации), предварительная обработка данных (например, перевод всех символов текста в нижний регистр), а также удаление стоп-слов (это часто используемые слова, которые не влияют на смысл текста, такие как артикли, предлоги, союзы, частицы, местоимения).

Однако для задачи анализа тональности текста нельзя удалять все стоп-слова, так как это может отразиться на эмоциональной окраске отзыва. Например, текст “The movie was not good at all” (фильм был не совсем хорош) после удаления стоп-слов превратится в “movie good” (фильм хорош). Как можно заметить, тональность текста при удалении стоп-слов изменилась на противоположную. Кроме того, слова, усиливающие тональность, также стоит оставить в текстах. Таким образом, из списка стоп-слов были удалены слова “no”, “not”, обозначающие отрицание, и “very”, усиливающее тональность последующего слова.

Воспользуемся методом Word2Vec для получения векторных представлений слов и обучим модель на корпусе отзывов тренировочного набора. Word2Vec — одна из наиболее эффективных и широко используемых моделей для формирования векторных представлений слов. Метод основывается на том, что слова, которые похожи по значению, должны иметь схожие значения векторов [12]. В данной работе учитываются 6 соседних слов из контекста, в модель сохраняются слова, которые встречаются более 1 раза.

Далее воспользуемся классом Tokenizer и обучим его на отзывах. В процессе обучения строится словарь соответствия каждого слова и его числового представления. Токенизация производится, опираясь на то, как часто каждое слово встречается в тексте. Затем преобразуем текст в числовое представление (последовательность) на обученном токенайзере.

Ограничим максимальную длину отзыва средним количеством слов в рассматриваемых текстах — числом 100. Если в отзыве больше заданного количества слов, он обрезается, если же меньше — он дополняется нулями в начале числовой последовательности до заданного размера.

3.3. Создание и обучение модели

Для создания нейронной сети был использован последовательный тип модели Sequential, где можно последовательно добавлять слои. Первый слой — Embedding («слой встраивания») который содержит матрицу встраивания, полученную после обучения модели Word2Vec. Параметр trainable равен False, так как эта модель уже обучена. Длина последовательностей, подаваемых на вход, равна 100, как и длина ограниченных ранее отзывов. Второй слой — LSTM со 128 ячейками. Выходной слой выдает 1 нейрон, функция активации — сигмоидальная. Схема модели приведена на рисунке 3.

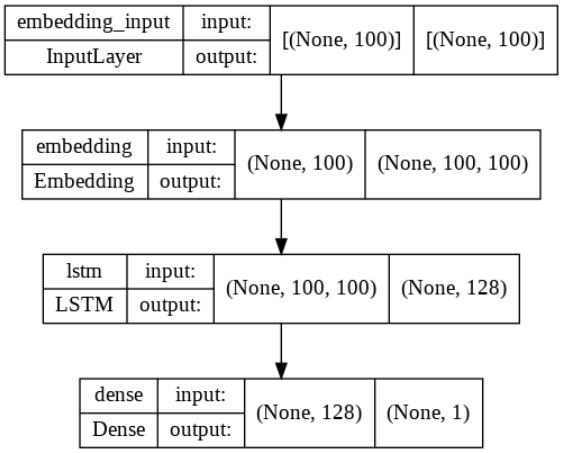

Рис. 3. Схема модели нейронной сети

Для компиляции модели в качестве оптимизатора используем Adam, в качестве функции потери применялась бинарная кросс-энтропия, а в качестве метрик для оценки модели — accuracy (доля правильных ответов), precision (точность) и recall (полнота). Подробно метрики рассмотрены в работе [13].

Для того, чтобы обойти проблему переобучения, можно использовать callback (обратный вызов): таким образом, модель будет сохраняться на каждой эпохе обучения, а лучшая копия (по параметру точности на проверочном наборе данных) будет сохранена в файл.

При обучении нейросети зададим следующие параметры: эпох — 12, размер минимальной выборки — 128, а также для проверки на проверочном наборе данных будет использоваться 10 % набора.

График долей правильных ответов классификатора приведен на рисунке 4.

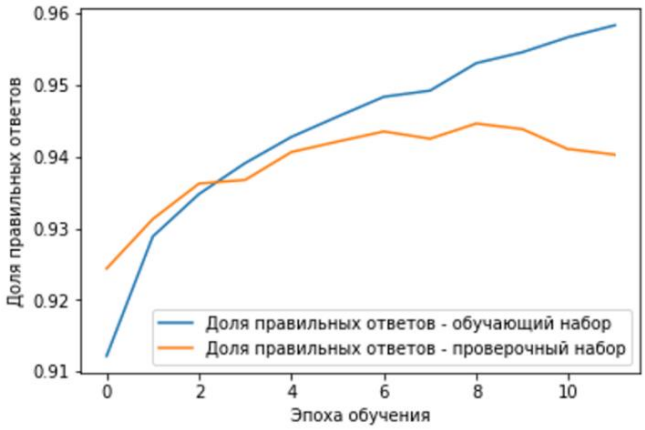

Рис. 4. График долей правильных ответов классификатора на обучающем и проверочном наборах

Наибольшая доля правильных ответов на проверочном наборе данных была достигнута на эпохе 9 — это 0.94463 (или 94,463 %). Таблица всех используемых метрик на обучающей и проверочной выборке для лучшей модели сети приведена ниже (табл. 1).

Таблица 1

Значения метрик качества для обучающей и проверочной выборки тренировочного набора

Тип выборки	Доля верных ответов	Точность	Полнота
Обучающая выборка	0.9531	0.9527	0.9535
Проверочная выборка	0.9446	0.9437	0.9453

4. Оценка полученных результатов и варианты внедрения

4.1. Оценка работы модели на тестовых наборах

Для тестирования используем тестовые выборки из датасетов YELP и IMDB, на которых обучалась модель. В таблице 2 представлены метрики для обоих наборов.

Таблица 2

Метрики качества работы нейросети на тестовых наборах

Тестовый набор	Доля верных ответов	Точность	Полнота
Тест. набор YELP	0,9498	0.9572	0.9418
Тест. набор IMDB	0,8920	0.8932	0,8892

Таким образом, были получены достаточно высокие доли верных ответов на обоих тестовых наборах — почти 95 % на YELP и чуть больше 89 % на IMDB. Точность выше на первом наборе, так как он намного больше набора с рецензиями, и для обучения нейронной сети доминирующую часть обучающего набора составляли именно отзывы из YELP. Соответственно, полученная модель с большой точностью классифицирует подобные отзывы о товарах, услугах и сервисах, однако с рецензиями на фильмы она также справляется на достаточно высоком уровне.

Для того, чтобы сравнить эффективность работы различных методов машинного обучения с полученной моделью нейронной сети, воспользуемся библиотекой Sklearn. Применим наивный байесовский классификатор, логистическую регрессию и стохастический градиентный спуск к тем же наборам данных, с которыми работали ранее: для обучения используем объединенный корпус из отзывов наборов YELP и рецензий IMDB, а для тестирования — тестовые наборы этих датасетов. Также проведем процесс предварительной обработки текстов с обоими наборами, затем применим векторизацию. В таблице 3 представлено сравнение результатов разработанной модели с другими методами машинного обучения.

Таблица 3

Сравнение методов машинного обучения в задаче классификации текста по тональности

Метод машинного обучения	Доля верных ответов	
	YELP	IMDB
Наивный байесовский классификатор	0.8119	0.8003
Логистическая регрессия	0.9279	0.8688
Стохастический градиентный спуск	0.9293	0.8793
Разработанная модель (Word2Vec + LSTM)	**0.9498**	**0.8920**

Как можно заметить, полученная модель нейронной сети показывает наибольшую точность по сравнению с другими методами на обоих тестовых наборах данных.

4.2. Пример практического использования модели

Полученная модель достаточно универсальна и может применяться в различных областях, где тональность текстов имеет значение. Чтобы проиллюстрировать практическое применение, был выбран анализ комментариев под видео на сайте YouTube. Не так давно видеохостинг отключил демонстрацию количества отрицательных отметок к видео (дизлайков), поэтому нельзя понять мнение пользователей о видео по соотношению отметок (лайков/дизлайков), как это было ранее. Для этого можно воспользоваться анализом тональности отзывов и понять соотношение положительных и отрицательных мнений.

Сначала необходимо объединить комментарии в набор данных, затем применить к нему функцию предварительной обработки текста и удалить пустые строки. Далее нужно загрузить и применить предобученный токенайзер, а также модель нейронной сети. Можно проиллюстрировать полученные результаты с помощью круговой диаграммы (рис. 5).

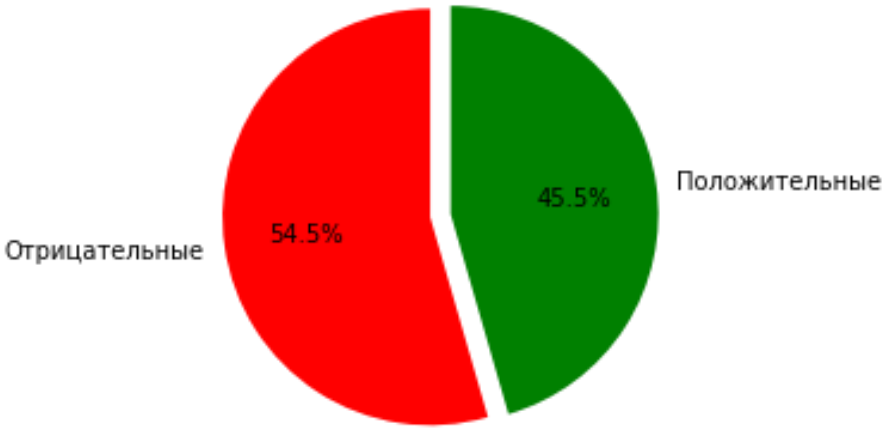

Рис. 5. Пример диаграммы положительных и отрицательных отзывов о видео

Таким образом, применяя полученную в этой работе модель, можно вывести соотношение положительных и отрицательных мнений о видео, представив их наглядно в виде диаграмм. Нельзя исключать тот факт, что какой-то процент отзывов скорее бы относился к «нейтральным» при рассмотрении трех классов текстов (положительные, отрицательные и нейтральные), однако рассмотренная бинарная классификация предоставляет более «грубую» статистику, что также, безусловно, важно для различных целей.

Заключение

В работе были рассмотрены различные методы машинного обучения для задачи классификации текстов. Кроме того, были изучены научные статьи на тему нейронных сетей в задачах анализа данных на естественном языке. Была создана нейронная сеть на языке Python с использованием библиотек TensorFlow и Keras. Нейросеть была обучена на наборах данных YELP и IMDB, которые включают отзывы на различные услуги и места, а также рецензии на фильмы.

Полученная модель нейросети показала высокий результат в задаче классификации текстов по тональности: доля верных ответов на тестовых наборах данных составила почти 95 % на наборе YELP и чуть больше 89 % на наборе IMDB. Также был проиллюстрирован пример практического использования результатов проведенного исследования.

Таким образом, полученная модель нейронной сети для классификации текстов по тональности имеет высокую точность и может применяться для различных прикладных задач, таких, как, например, анализ мнений о различных товарах и услугах или оценка реакции людей разного рода события или изменения.

Список литературы

1. Самигулин Т.Р., Джурабаев А.Э.У. Анализ тональности текста методами машинного обучения // Научный результат. Информационные технологии. – 2021. – Т. 6, № 1. – С. 55–62.

2. Волкова В. Н. Моделирование систем: учеб.пособие / В.Н. Волкова, В.Н. Козлов, Ю.И. Лыпарь [и др .] – СПб. : Изд-во Политехн. ун-та, 2012. – 440 с.

3. Tang D., Qin B., Liu T. Deep learning for sentiment analysis: Successful approaches and future challenges // Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery. – 2015 – Vol.5, No 6, – Pp. 292–303.

4. A simple overview of RNN, LSTM and attention mechanism // Medium. – URL: https://medium.com/swlh/a-simple-overview-of-rnn-lstm-and-attention-mechanism-9e844763d07b (дата обращения: 13.11.2022).

5. Hochreiter S., Bengio Y., Frasconi P. Schmidhuber J. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies – 2001.

6. Пустынный Я. Н. Решение проблемы исчезающего градиента с помощью нейронных сетей долгой краткосрочной памяти // Инновации и инвестиции. – 2020. – № 2. – С. 130–132.

7. Lindén J. Evaluating combinations of classification algorithms and paragraph vectors for news article classification // Federated Conference on Computer Science and Information Systems – 2018 – P. 7.

8. Amajd M., Kaimuldenov Z., Voronkov I. Text classification with deep neural networks // Conference: International Conference on Actual Problems of System and Software Engineering (APSSE) – 2017 – Pp. 363–369.

9. Гафаров Ф.М., Галимянов А.Ф. Искусственные нейронные сети и приложения: учебное пособие. – Казань: Издательство Казанского университета, 2018. – 121 с.

10. Yelp Review Polarity // Kaggle. – URL: https://www.kaggle.com/datasets/irustandi/yelp-review-polarity (дата обращения: 13.11.2022).

11. IMDB Dataset of 50K Movie Reviews // Kaggle. – URL: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews (дата обращения: 13.11.2022).

12. Обзор четырёх популярных NLP-моделей // Proglib. – URL: https://proglib.io/p/obzor-chetyreh-populyarnyh-nlp-modeley-2020-04-21 (дата обращения: 13.11.2022).

13. Метрики в задачах машинного обучения // Habr. – URL: https://habr.com/ru/company/ods/blog/328372/ (дата обращения: 13.11.2022).