Е. Н. КАРУНА, П. В. СОКОЛОВ
Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург

СРАВНЕНИЕ МЕТОДОВ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ РУССКОЯЗЫЧНЫХ ТЕКСТОВ

В данной работе выполняется сравнительный анализ методов автоматической классификации текстовой информации с помощью различных алгоритмов машинного обучения, основанных на искусственных нейронных сетях. В работе рассмотрены результаты классификации текстовых данных для различных способов формирования векторного пространства, такие как bag of words, n-граммы, векторное представление слов, и различные архитектуры нейронных сетей, исследования проводились применительно к корпусу текстов на русском языке.

Введение. Большое количество текстовых данных, хранящихся как в открытых источниках, так и в закрытых базах данных предприятий, требуют качественных инструментов для анализа этих данных, в частности, для автоматической категоризации этих данных. Одной из проблем задач машинной классификации текстов является то, что текстовые данные не всегда можно однозначно определить к одному из конкретных классов по количественному составу всех слов, входящих в документ, решением этих проблем является разработка системы классификации, учитывающей взаимное расположение слов внутри текста. На текущий момент доступно достаточно большое количество материалов по исследованию систем классификации текстовой информации, но при этом, практически отсутствуют исследования, посвященные системам автоматической классификации данных, состоящих из русскоязычного набора текстов, хотя особенности грамматики русского языка могут значительно влиять на результаты работы подобных систем. В предлагаемом докладе выполняется сравнительный анализ методов автоматической классификации, и даются выводы о результатах работы исследуемых систем для русскоязычного корпуса данных.

Решение задачи классификации можно разделить на 2 основных этапа: предварительная обработка текстовых данных и алгоритм машинного обучения. Первый этап необходим для преобразования исходного текста в набор признаков, который представляет собой числовой вектор или матрицу. На втором этапе уже происходит реализация одного из алгоритмов машинного обучения, который должен быть способен по набору признаков текста в виде вектора или матрицы, определить его тематическую принадлежность.

Предварительная обработка текстовых данных. На данном этапе необходимо выполнить ряд задач: удаление небуквенных символов из текста, разбиение текста на набор токенов, удаление стоп-слов, выполнение операции приведения к основе слова. Удаление небуквенных символов и стоп-слов является стандартной процедурой и позволяет очистить текст от лишних элементов, которые слабо влияют на общую тематику текста. Для приведения к основе слова применяется один из двух способов: стемминг и лемматизация, в первом случае выполняется грубое отсечение окончания слова по определённому алгоритму, во втором случае происходит приведение слова к своей начальной форме, согласно языковой грамматике.

Определение тематики текстов требует наличия заранее промаркированного корпуса данных, который будет использоваться для обучения алгоритма. Корпус текстов был взят из библиотеки corus [1]. В данной работе выбран набор более чем из 102 тысяч новостных статей интернет-издания «Lenta.ru» на русском языке, разбитых на 10 категорий. Каждая новостная статья, помимо основного текста и заголовка имеет метку темы, каждая тема является отдельной категорией. Вся коллекция текстов была разбита на 2 группы: обучающая и тестовая, где размер обучающей группы составлял 80 % от общего числа текстов. Были проведены исследования с некоторыми распространёнными моделями нейронных сетей [2]. Далее, будут отдельно рассмотрены некоторые из них.

Сеть прямого распространения. Для сетей прямого распространения вход подаётся одиночный числовой вектор, и каждый вектор отвечает за один текст из корпуса данных. Наиболее распространенный способ составления входного вектора текста это BOW (Bag of words – ме-

шок слов). В данной работе были использованы 2 вида статистической меры, для вычисления значений векторов TF и TF-IDF, формулы 1 и 2 соответственно:

$$tf_i = \frac{n_{id}}{\sum_{j=1}^{i_{max}} n_{jd}} \quad (1)$$

$$tfidf_i = tf_i \times \ln\left(\frac{|D|}{|\{d \mid i \in d\}|}\right) \quad (2)$$

где tf_i – частота употребления слова i в документе d, n_{id} – количество слов i в документе d, $imax$ – размер словаря, $\sum_{j=1}^{i_{max}} n_{jd}$ – количество слов в документе d. D – набор всех документов в корпусе, $|D|$ – количество документов в коллекции, $|\{d \mid i \in d\}|$ – число документов, в которых встречается слово i.

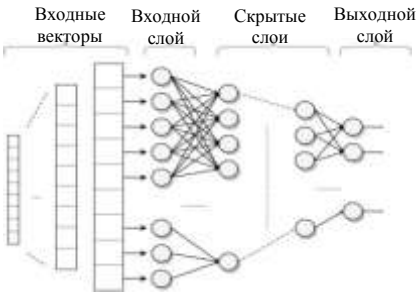

Рис. 1. Нейронная сеть прямого распространения

Размерность векторов ограничена и определяется пользователем. Были проведены исследования точности классификации при использовании различных статистических мер составления векторов, различных размерах входного вектора, различных способах нахождения основы слова. Также получены результаты классификации при составлении словаря как из отдельных слов (униграмм), так и из сочетаний по 2 или 3 слова (биграммы и триграммы, соответственно).

Для следующих экспериментов был использован подход, известный как векторное представление слов, при котором каждое слово представляет собой отдельный вектор [3]. Преимуществом этого подхода является то, что он позволяет применять алгоритмы машинного обучения, которые учитывают не только количественный состав всех слов в тексте, но, также, взаимное расположение слов внутри всего текста. В данной работе будет использоваться предобученная модель векторного представления слов.

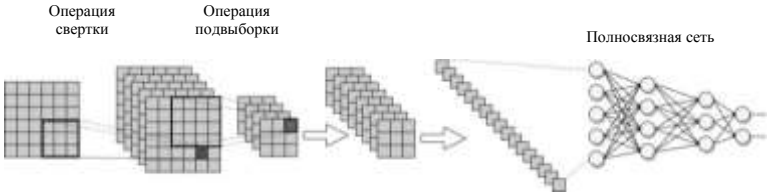

Рис. 2. Сверточная нейронная сеть

Свёрточная нейронная сеть. Модель CNN (convolutional neural network – сверточная нейронная сеть) активно используется при решении задач распознавания образов. При решении таких задач входные данные представляют собой матрицу фиксированного размера. Построить матрицу можно путём склеивания векторов, полученных при использовании алгоритма векторного представления слов. Суть алгоритма в попеременном применении на числовой матрице операций свёртки, для активаций нейронов и операции подвыборки для уменьшения размерности матрицы. В итоге матрица преобразуется в вектор, который поступает на вход сети прямого распространения.

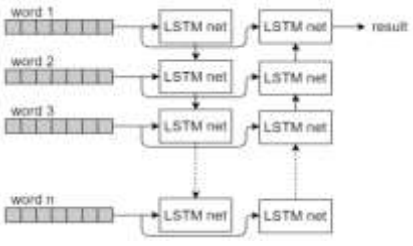

Рис. 3. Рекуррентная нейронная сеть с BiLSTM слоем

Рекуррентная нейронная сеть. Существует большое разнообразие алгоритмов на основе рекуррентных нейронных сетей, на данный момент одним из самых распространенных является LSTM-сеть (Long short-term memory) и её разновидность BiLSTM (Bidirectional Long short-term memory) – нейронная сеть с рекуррентным слоем, состоящим из LSTM-блоков, и проходящих входной массив данных в прямом и обратном направлении [4]. Особенностью рекуррентных нейронных сетей является связь между

нейронами, передающаяся во времени. Так, при отправке в сеть нескольких наборов данных, выходное состояние сети будет определяться не только текущим входом, но и теми данными, которые были отправлены ранее. Преимуществом LSTM-сетей является то, что они позволяют в течение неопределённого времени хранить накопленную информацию. Все результаты исследований были сведены в таблицу.

Таблица

Сравнение моделей алгоритмов классификации

Модель	Точность	Время обучения, сек
FFNN – BOW – TF-IDF – 1-gramm	0,872	1807
FFNN – BOW – TF-IDF – 2-gramm	0,891	1932
FFNN – BOW – TF-IDF – 3-gramm	0,868	1865
CNN	0,894	39563
BiLSTM	0,905	15892
LSTM	0,901	12685

По таблице видно, что лучшую точность классификации обеспечивает алгоритм нейронной сети с двунаправленным LSTM слоем, при этом разница между этим алгоритмом и другим является несущественной. Этот алгоритм выполняет обучение за время, значительно большее времени обучения модели любой конфигурации алгоритма с сетью прямого распространения. Помимо высокой скорости, модели на основе алгоритма мешка слов обладают достаточно высокой точностью и простым способом реализации, не требуя каких-либо предобученных моделей. Нейронная сеть на основе сверточной нейронной сети имеет достаточно высокую точность, сопоставимую с рекуррентными сетями, но требуют очень большого времени обучения.

Заключение. В данной работы были продемонстрированы некоторые распространенные алгоритмы автоматической классификации текстовых данных на новостном корпусе русскоязычных текстов. По полученным результатам видно, что рекуррентные нейронные сети с LSTM слоем показывают очень хорошие показатели эффективности в классификации текстов русскоязычного корпуса данных. Сверточные нейронные сети также демонстрируют достаточно высокий уровень точности классификации. Отличие этих моделей от моделей на основе мешка слов заключается в том, что они способны учитывать семантическую связь между словами. Нечто подобное способны продемонстрировать также модели с использованием биграмм и триграмм, но их результаты оказались хуже, чем в других моделях. Данное исследование показывает разницу между некоторыми распространенными моделями анализа текстовых данных и демонстрирует их показатели эффективности на основе решения задачи классификации текстов в русскоязычном корпусе данных. Дальнейшие исследования могут быть направлены на улучшение существующих алгоритмов анализа текстовых данных на основе полученных результатов для решения различных задач обработки естественного языка.

ЛИТЕРАТУРА

1. natasha/corus: Links to Russian corpora, python functions for loading and parsing. https://github.com/natasha/corus
2. Kowsari K., Jafari Meimandi K., Heidarysafa M., Mendu S., Barnes L., Brown D. Text Classification Algorithms: A Survey. Information 2019, 10(4), 150
3. Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, C.-C. Jay Kuo. Evaluating Word Embedding Models: Methods and Experimental Results. APSIPA Transactions on Signal and Information Processing 8 (2019)
4. Tao Chen, Ruifeng Xu, Yulan He, Xuan Wang. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Systems with Applications. Vol. 72. 2017. P. 221-230

E.N. Karuna, P.V. Sokolov (Saint Petersburg Electrotechnical University “LETI”, St. Petersburg)

Comparison of Methods for Automatic Classification of Russian-language Texts

In this paper, we carry out a comparative analysis of methods for automatic classification of text information using various machine learning algorithms based on artificial neural networks. The paper considers the results of the classification of text data for various ways of forming a vector space, such as bag of words, n-grams, word embedding, and various architectures of neural networks, the research was carried out in relation to the corpus of texts in Russian language.