УДК 004.89, 004.912

Н.Г. Ярушкина, В.С. Мошкин, А.А. Константинов

ПРИМЕНЕНИЕ ЯЗЫКОВЫХ МОДЕЛЕЙ WORD2VEC И BERT В ЗАДАЧЕ СЕНТИМЕНТ-АНАЛИЗА ТЕКСТОВЫХ СООБЩЕНИЙ СОЦИАЛЬНЫХ СЕТЕЙ[1]

Ярушкина Надежда Глебовна, *доктор технических наук, профессор, окончила Ульяновский политехнический институт по специальности «Электронные вычислительные машины». Заведующая кафедрой «Информационные системы» Ульяновского государственного технического университета. Имеет более 400 научных работ в области мягких вычислений, нечеткой логики, гибридных систем. [e-mail: jng@ulstu.ru].*

Мошкин Вадим Сергеевич, *окончил факультет информационных систем и технологий УлГТУ, доцент кафедры «Информационные системы» УлГТУ. Имеет более 90 статей в области интеллектуальных систем анализа данных. [e-mail: v.moshkin@ulstu.ru].*

Константинов Андрей Алексеевич, *магистрант кафедры «Информационные системы» УлГТУ. Имеет статьи в области интеллектуального анализа текстовых данных. Область научных интересов – автоматизация анализа текстов с использованием машинного обучения. [e-mail: adwaises@mail.ru].*

Аннотация

В работе предложен оригинальный алгоритм формирования обучающей выборки для нейронной сети, обеспечивающей сентимент-анализ текстовых сообщений социальных сетей. Особенностью алгоритма является использование расширенного русскоязычного семантического тезауруса WordNetAffect и экспертного словаря авторских символов выражения эмоций. Помимо этого, в работе описывается применение нейронной сети на базе LSTM-архитектуры для определения эмоциональной окраски текстовых сообщений социальной сети с применением двух алгоритмов векторизации текста «word2vec» и «BERT». В результате проведенных экспериментов был достигнут показатель точности определения эмоциональной окраски сообщений в 87% с использованием лемматизации в качестве алгоритма предобработки текста и алгоритма «BERT» при его преобразовании в векторную форму.

Ключевые слова: сентимент-анализ, BERT, word2vec, нейронная сеть, социальная сеть.

doi: 10.35752/1991-2927-2020-3-61-60-69

WORD2VEC AND BERT LANGUAGE MODELS USED FOR A SENTIMENT ANALYSIS OF TEXT POSTS IN SOCIAL NETWORKS

Nadezhda Glebovna Yarushkina, *Doctor of Sciences in Engineering, Professor; graduated from Ulyanovsk Polytechnic Institute with the specialty in Electronic Computing Machines; Head of the Department of Information Systems at Ulyanovsk State Technical University; an author of more than 400 papers in the field of soft computing, fuzzy logic, and hybrid systems. e-mail: jng@ulstu.ru.*

1 Работа выполнена при финансовой поддержке РФФИ, гранты № 18-47-730035 и 18-47-732007.

Vadim Sergeevich Moshkin, *graduated from the Faculty of Information Systems and Technologies of Ulyanovsk State Technical University; Associate Professor of the Department of Information Systems of UlSTU; an author of more than 90 papers in the field of data analyzing intelligent systems. e-mail: v.moshkin@ulstu.ru.*
Andrei Alekseevich Konstantinov, *a student in the master's degree at the Department of Information Systems of Ulyanovsk State Technical University; an author of articles in the field of text mining; the area of his scientific interests relates to the automation of text analysis using machine learning. e-mail: adwaises@mail.ru.*

Abstract
The paper proposes an original algorithm for the formation of a training sample for a neural network that provides a sentiment analysis of text posts in social networks. A feature of the algorithm is the use of the extended Russian-language semantic thesaurus WordNetAffect and the expert dictionary of author's symbols for expressing emotions. In addition, the paper describes the application of a neural network based on the LSTM architecture to determine the emotional coloring of text messages on a social network using two text vectorization algorithms "word2vec" and "BERT". As a result of the experiments, an indicator of the accuracy of determining the emotional coloring of messages of 87% was achieved using lemmatization as a text preprocessing algorithm and the BERT algorithm when converting it into a vector.

Key words: sentiment analysis, BERT, word2vec, neural network, social network.

Введение

Исследование социальных сетей с каждым годом приобретает все большую актуальность в связи с обостряющейся необходимостью обеспечения безопасности населения и мониторинга общественных настроений. Анализ комментариев и постов может помочь оценить изменения в настроениях многих пользователей и найти применение в политических и социальных исследованиях, в том числе и в исследованиях потребительских предпочтений.

В рамках данного исследования под постом понимается отдельно взятое текстовое сообщение, размещенное пользователем в своем профиле в социальной сети. Комментарий – это пояснение к посту, рассуждение, замечание о чём-нибудь, представленное в текстовой форме. Оба понятия объединяются в рамках терминов «текстовое сообщение» или «сообщение».

Благодаря анализу тональности текстов сообщений пользователей, исследователь может сделать следующие выводы:

- об эмоциональной оценке пользователей различных событий и объектов;
- о предпочтениях отдельных пользователей;
- об отдельных чертах характера пользователей [1].

Определение тональности текста – это задача классификации. В настоящее время при работе с естественным языком лучшие результаты классификации текстов по нескольким критериям показывают алгоритмы, основанные на машинном обучении. Однако при использовании нейросетевых подходов актуальным становится подбор и формирование обучающей выборки.

В данной работе будет рассмотрено применение языковых моделей word2vec и «BERT» для решения задачи анализа тональности сообщений в социальных сетях, а также решение задач предобработки текстовой информации и формирования обучающей выборки.

Применение алгоритмов машинного обучения в сентимент-анализе данных социальных сетей

В настоящее время в работах зарубежных и отечественных исследователей предлагается использование нейронных сетей различных архитектур для определения эмоциональной окраски текстов. В качестве базовых часто используются следующие подходы: алгоритм опорных векторов (SVM) [2], Байесовские модели [3], различного рода регрессии [4], методы Word2Vec [5], Doc2Vec [6], CRF [7], а также сверточные и рекуррентные нейронные сети [8].

Для анализа тональности текстов из социальных сетей, в отличие от анализа формальных и строго структурированных текстов, требуется больше шагов по предобработке ресурсов, в том числе и для формирования обучающей выборки.

В работе [9] описывается прототип определения тональности постов социальной сети Twitter. Для формирования обучающей выборки первоначально был составлен корпус смайлов для разметки текста и отнесения текста к конкретной эмоции. Затем тексты были представлены в векторной форме с помощью подхода «мешок слов».

Для построения классификационной модели были выбраны три классификатора: логистическая регрессия, дерево решений, многослойный персептрон. В результате экспериментов точность определения эмоциональной окраски постов была около 75–76% для каждой модели.

В работе [10] для определения тональности текстовых сообщений выбраны 2 модели нейронных сетей: нейронная сеть с двумя рекуррентными слоями и нейронная сеть с рекуррентным и сверточным слоями.

При обучении нейронной сети использовались два набора размеченных вручную по тональности текстов, а именно коротких сообщений длиной до 140 символов.

Для сети с двумя рекуррентными слоями точность классификации составила 69%. Для сети с рекуррентным и сверточным слоями точность немного выше – 71%.

В работе [11] рассматриваются результаты создания автоматического классификатора русскоязычных интернет-текстов, распределяющего тексты на 8 классов в соответствии с 8 базовыми эмоциями.

В основе работы классификатора лежал алгоритм машинного обучения с использованием метода опорных векторов. На вход классификатору подавались различные лингвистические параметры, например, частотность использования пунктуационных знаков и усилительных наречий. Точность определения эмоциональной окраски эмоций «злость» и «страх» составила 48%, «тоска» – 40%, «отвращение» – 6%, «радость» – 7%.

Как видно из результатов приведенных исследований, задача разработки подхода, позволяющего эффективно оценивать тональность текстов социальных сетей, является актуальной.

Подход к сентимент-анализу данных социальных сетей с использованием моделей «word2vec» и «BERT»

В рамках данного исследования был разработан подход к анализу тональности текстовых данных социальных сетей, состоящий из следующих этапов:

1. Формирование обучающей и тестовой выборки.
2. Векторизация текста с использованием моделей «word2vec» и «BERT».
3. Обучение и классификация с использованием нейросетевого подхода.

Алгоритм формирования обучающей выборки

Текстовые данные, извлекаемые из социальных сетей, имеют ряд особенностей, одной из которых является непосредственное обозначение автором эмоциональной окраски сообщения посредством использования авторских символов выражения эмоций (так называемых «смайлов» и «эмодзи») [12]. В рамках данной работы эта особенность была использована при формировании обучающей выборки для нейронной сети при решении задачи оценки эмоциональной окраски текстовых сообщений.

Формирование обучающей и тестовой выборок предполагает необходимость предобработки текстовой информации, а также разметку эмоциональной окраски отдельных текстовых сообщений.

Формально процесс отбора текстовых сообщений можно представить схемой, показанной на рисунке 1. Каждый этап отбора, представленный на данном рисунке, включает процессы отбора сообщений для каждой конкретной эмоции.

1. На первом этапе выполняется отбор текстовых сообщений на основе экспертных словарей авторских символов выражения эмоций. Если сообщение содержит подобный авторский символ, то оно относится к конкретному классу и добавляется в соответствующий список.

2. На втором этапе выполняется отбор сообщений на основе словарей ключевых фраз. В качестве базового словаря использовался расширенный русскоязычный семантический тезаурус WordNetAffect [13].

Разработанные словари с авторскими символами выражения эмоций и ключевыми фразами состоят из объектов 7 классов:

$$D^E = \left\{ D^E_{joy}, D^E_{sad}, D^E_{surp}, D^E_{anger}, D^E_{disg}, D^E_{cont}, D^E_{fear} \right\},$$

где D^E_{joy} – класс объектов с эмоцией «радость»,

D^E_{sad} – класс объектов с эмоцией «грусть»,

D^E_{surp} – класс объектов с эмоцией «удивление»,

D^E_{anger} – класс объектов с эмоцией «злость»,

D^E_{disg} – класс объектов с эмоцией «отвращение»,

D^E_{cont} – класс объектов с эмоцией «презрение»,

D^E_{fear} – класс объектов с эмоцией «страх».

Помимо этого, на данном этапе выполняется лемматизация каждого слова текстового сообщения. Затем сообщение проверяется на содержание каждого слова из словаря. Если сообщение содержит фразу, значит оно принадлежит к конкретному классу эмоциональной окраски.

3. На этапе предобработки текстовых сообщений происходит исключение всех символов, кроме символов кириллицы и пробелов, также все слова приводятся к нижнему регистру.

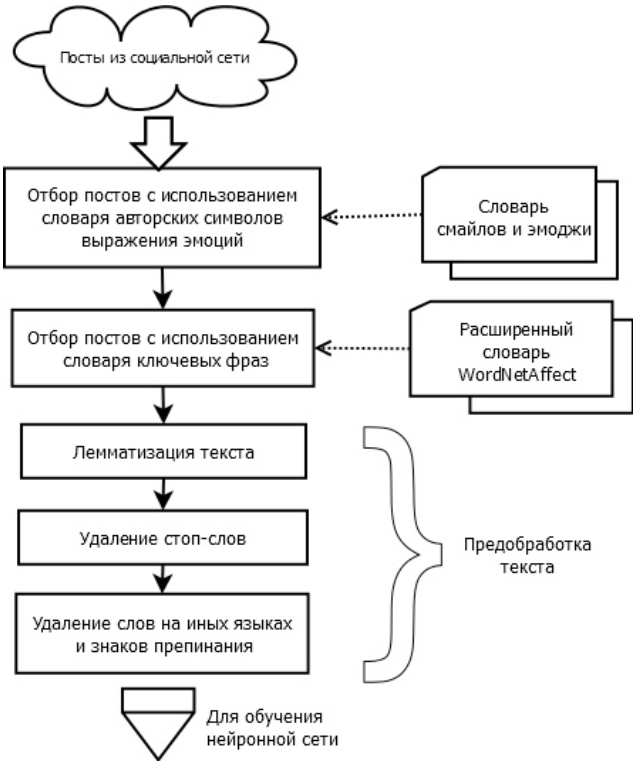

Рис. 1. Схема отбора сообщений

Алгоритмы векторизации текстов

В рамках данного исследования для представления слов в векторном пространстве были использованы два метода: word2vec и «BERT».

Модель алгоритма «BERT» можно представить в виде функции, на вход которой подаётся текст, а на выходе получается вектор. В данном алгоритме каждый слог преобразуется в число. Первоначально загружается обученная для определённого языка модель, по которой происходит разбиение последовательности на слоги. Подробное описание алгоритма приведено в работах [14] и [15].

Математическая модель алгоритма «BERT». Загруженную модель можно представить в виде:

$$\Theta = \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ w_n \end{bmatrix},$$

где Θ – это длинный вектор, который содержит слова w, входящие в словарь слов загруженной модели. Алгоритм преобразует слово в набор слогов или векторов, каждый слог получается из набора общих слов.

Пусть $w_1, w_2, \ldots, w_n$ – множество слов в словаре и $s_{m1}, s_{m2}, \ldots, s_{mn}$ – множество слогов в слове w_n, тогда функция $(s_m) = f(w_{11}, w_{12}, \ldots, w_{1n})$ позволяет получить множество слогов для последовательности слов.

Затем по полученным слогам появляется возможность получать векторное представление последовательности слов.

Также для сравнительной оценки эффективности использования языковой модели «BERT», был апробирован алгоритм «word2vec».

Модель алгоритма «word2vec» можно представить как функцию, которая преобразует слово в вектор. Подробное описание алгоритма приведено в работе [16].

Математическая модель алгоритма «word2vec». Первоначально составляется словарь всех слов, которые входят в корпус. Все векторы слов можно представить следующим образом:

$$\Theta = \begin{bmatrix} v_{w1} \\ v_{w2} \\ \cdot \\ v_{wn} \\ u_{w1} \\ u_{w2} \\ \cdot \\ u_{wn} \end{bmatrix},$$

где Θ – это длинный вектор, который содержит векторы v и u длины d для всех слов.

Алгоритм предсказывает вероятность слова по его контексту. То есть получаются такие векторы слов, чтобы вероятность, присваиваемая моделью слову, была близка к вероятности встретить это слово в этом окружении в реальном тексте.

$$P(w_o | w_c) = \frac{e^{s(w_o, w_c)}}{\sum w_i \in V e^{s(w_o, w_c)}},$$

где w_o – вектор целевого слова, w_c – это некоторый вектор контекста, вычисленный (например, путем усреднения) из векторов других слов, окружающих нужное слово. А $s(w_1, w_2)$ – это функция, которая двум векторам сопоставляет одно число.

В стандартной модели, рассмотренной выше, предсказываются и оптимизируются вероятности слов. Функцией для оптимизации служит дивергенция Кульбака-Лейблера:

$$KL(p \| q) = \int p(x) \log \frac{p(x)}{q(x)} dx,$$

где $p(x)$ – распределение вероятностей слов, которое берется из корпуса, $q(x)$ – распределение, которое порождает модель. Дивергенция – это буквально «расхождение», насколько одно распределение не похоже на другое.

Модель нейронной сети для сентимент-анализа текстовых фрагментов

Модель нейронной сети можно представить в виде слоёв, использованных в её архитектуре. Нейронная сеть состоит из семи слоёв и представлена на рисунке 2.

Математически нейрон представляет собой взвешенный сумматор, единственный выход которого определяется через его входы и матрицу весов следующим образом:

$$y = f(u), \text{ где } u = \sum_{i=1}^{n} w_i x_i + w_0 x_0,$$

где x_i и w_i — соответственно сигналы на входах нейрона и веса входов, функция u называется индуцированным локальным полем, а $f(u)$ – передаточной функцией. Возможные значения сигналов на входах нейрона считают заданными в интервале [0,1]. Дополнительный вход x_0 и соответствующий ему вес w_0 используются для инициализации нейрона. Под инициализацией подразумевается смещение активационной функции нейрона по горизонтальной оси.

Предложенная архитектура нейронной сети имеет следующий набор слоев:

• Слой Embedding – входной слой нейронной сети, состоящий из нейронов:

$$Emb = \{Size(D), Size(S_{vec}), L_{sec}\},$$

где $Size(D)$ – размер словаря в текстовых данных;

$Size(S_{vec})$ – размер векторного пространства, в которое будут вставлены слова; $Size(S_{vec})$ = 32;

L_{sec} – длина входных последовательностей, равная максимальному размеру вектора, сформированного при предобработке слов.

• Слой Conv1D – свёрточный слой, необходим для глубокого обучения. С данным слоем точность классификации текстовых сообщений повышается на 5–7%. Количество фильтров – 32, длина каждого равна 3. Функция активации – «relu».

• Слой MaxPooling1D – слой, отвечающий за уменьшение размерности сформированных карт признаков. Максимальный пул равен 2.

• Слой LSTM – рекуррентный слой нейронной сети. В модели используется два LSTM слоя, один составляет 50 блоков, второй – 20.

• Слой Dropout – нужен, чтобы избежать переобучения нейронной сети. В качестве параметра подаётся значение 0,5, которое означает, что нейронная сеть может исключать до половины неактивных нейронов.

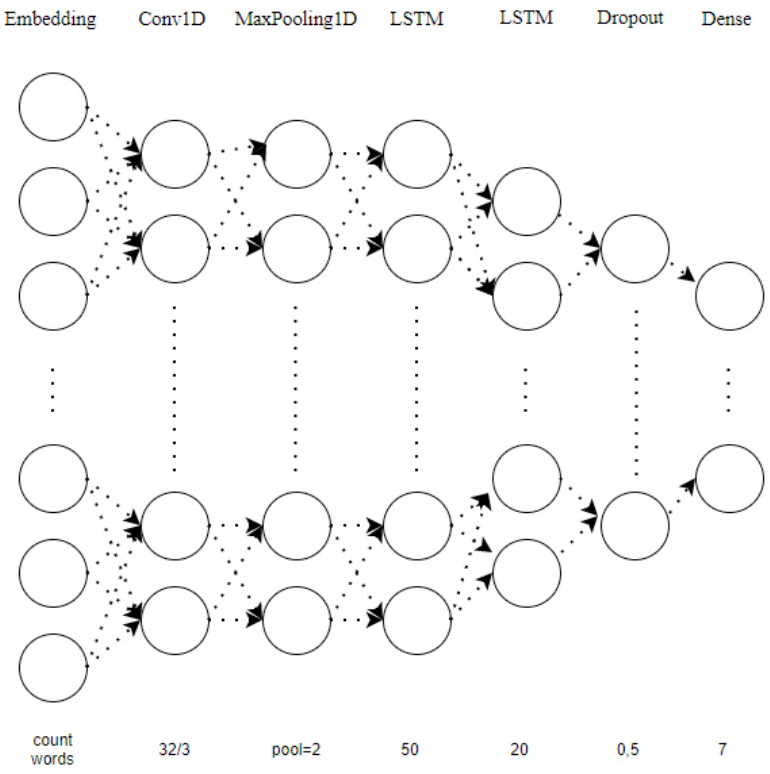

Рис. 2. Архитектура разработанной нейронной сети

• Слой Dense – выходной слой, состоящий из семи нейронов. Каждый нейрон отвечает за конкретную эмоцию.

Последовательность работы нейронной сети предложенной архитектуры предполагает последовательное преобразование данных на каждом из активных слоев и включает в себя следующие этапы:

1. Подача данных, полученных путем векторизации с использованием алгоритмов word2vec или BERT, на слой Embedding. В этом слое вектор преобразуется для обработки в нейронной сети (изменяется структура данных). Слой Embedding является обычным полносвязным слоем, состоящим из нейронов, как в обычном перцептроне.

2. Передача данных из слоя Embedding в свёрточный слой. В свёрточном слое используется ядро свертки небольшого размера, которое перемещается по всей входной матрице, формируя после каждого сдвига сигнал активации для нейрона следующего слоя с аналогичной позицией (рис. 3).

Полученный в результате свёртки слой показывает наличие данного признака в обрабатываемом слое и её координаты, формируя карту признаков. Ядра свёртки формируются путём обучения сети методом обратного распространения ошибки.

3. Передача данных в субдискретизирующий слой.

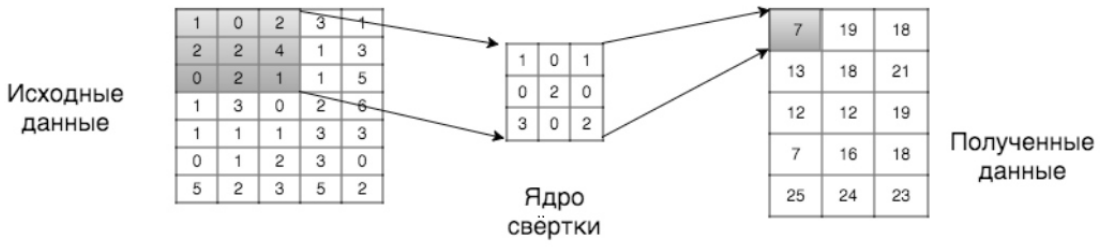

Рис. 3. Пример работы свёрточного слоя

Рис. 4. Пример работы субдискретизирующего слоя нейронной сети

Субдискретизация (MaxPooling1D) выполняет уменьшение размерности полученных карт признаков. Из нескольких соседних нейронов карты признаков выбирается максимальный и принимается за один нейрон уплотнённой карты признаков меньшей размерности (рис. 4).

4. Передача в рекуррентный слой LSTM размером в 50 нейронов и в слой LSTM с 20 нейронами.

В отличие от обычных слоёв из нейронов, рекуррентный слой состоит из блоков. Рекуррентный слой представлен на рисунке 5.

Принцип работы LSTM подробнее рассмотрен в работах [16–18].

5. Передача данных в слой Dropout, который решает, какие нейроны или блоки можно исключить в предыдущих слоях, так как они являются избыточными.

6. Передача данных в слой Dense, где происходит преобразование данных в выходной вектор, включающий 7 значений. Слой Dense также является полносвязным.

Пример работы алгоритма формирования обучающей выборки

Рассмотрим пример, демонстрирующий как работает алгоритм отбора текстовых сообщений для получения обучающей выборки. Возьмём 7 сообщений социальной сети, которые представлены в таблице 1.

Таблица 1

Первый этап отбора текстовых сообщений

Текст	Эмоция
я люблю лето 😀	joy
июль с августом вышли очень продуктивными	-
за год я убила две пары кед 😢	sad
вот почему у нас так грязно	-
фотографии хорошо отражают внутреннее состояние	-
закрытие сезона прошло отлично 😍	joy
мой организм продолжает бунтовать 😠	anger

После первого этапа получается 7 групп сообщений, каждая из которой содержит сообщения, отобранные на основе авторских символов выражения эмоций. В примере группа «радость» содержит 2 сообщения, так как в них встретились авторские символы выражения эмоций из группы «радость». Группы «грусть» и «злость» содержат по 1 сообщению, так как в них встретились символы выражения эмоций из соответствующих групп. Остальные группы не содержат постов, так как сообщения, содержащие авторские символы выражения эмоции для данных групп, не встретились.

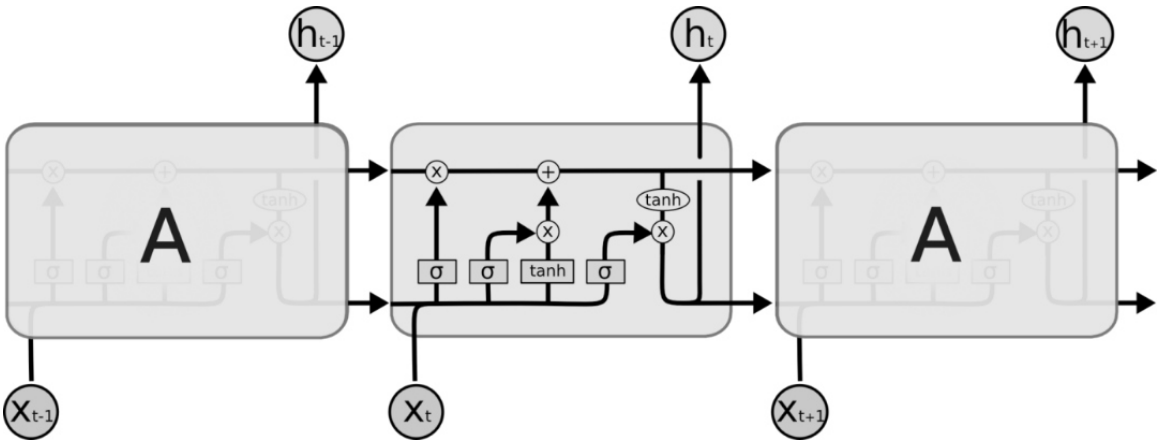

Рис. 5. Рекуррентный слой нейронной сети

Таблица 2

Второй этап отбора текстовых сообщений

Текст	Эмоция
я люблю лето 😀	joy
за год я убила две пары кед 😢	-
закрытие сезона прошло отлично 😍	joy
мой организм продолжает бунтовать 😠	-

На втором этапе происходит отбор на основе ключевых фраз. Каждая эмоциональная группа уточняется на основе словаря ключевых фраз. Результат отбора представлен в таблице 2.

В качестве выходных становятся сообщения, в которых присутствуют ключевые слова из словаря. Для группы «радость» это слова «люблю» и «отлично». Для других сообщений ключевые слова по данной эмоциональной группе не были найдены.

Результаты экспериментов по формированию обучающей выборки и оценке эмоциональной окраски

Для оценки эффективности предложенных моделей и алгоритмов была реализована программная система оценки тональности текстовых сообщений социальной сети.

Нейронная сеть была реализована на языке Python с применением фреймворков TensorFlow и Keras, предназначенных для машинного обучения. В качестве языка программирования был выбран язык Python.

При решении задачи формирования обучающей и тестовой выборок было обработано 2,5 млн. текстовых сообщений из социальной сети «ВКонтакте». Сообщения были получены из открытых групп социальной сети через API «ВКонтакте» и содержали только текстовую информацию.

Количество сообщений после каждого этапа отбора приведено в таблице 3. После каждого этапа отбора количество сообщений в каждой группе сокращалось в среднем в 2–3 раза. Стоит отметить, что сформированная выборка содержала текстовые сообщения различной длины.

Из полученного множества текстовых сообщений были отобраны данные для тестовой и обучающей выборок.

Нейронная сеть, состоящая из семи слоёв, была обучена на разном количестве текстовых сообщений, от 500 до 1000 и выше. Количество эпох обучения равно 100. Выборка делилась на обучающую и тестовую, 90% и 10% соответственно.

Для экспериментов 4–7 были отобраны сообщения, длина которых была в некотором интервале (40–50 слов или 290–310 символов). Иначе из-за множества коротких сообщений, в которых вектор дополняется нулями, нейронная сеть не обучается.

В ходе экспериментов была проверена гипотеза о том, что обучающая выборка, сформированная на основе авторских символов выражения эмоций и ключевых фраз, является более качественной, чем выборка, сформированная только на основе ключевых фраз, или выборка, сформированная только на основе авторских символов. Для экспериментов было сформировано 3 обучающих набора:

- на основе авторских символов выражения эмоций и ключевых фраз (включает первый и второй этапы отбора сообщений);
- на основе только авторских символов выражения эмоций (включает только первый этап отбора постов, исключён второй этап отбора);
- на основе только ключевых фраз (включает только второй этап отбора текстовых сообщений, исключён первый этап отбора).

При применении алгоритма BERT выборка содержала текстовые сообщения длиной 290–310 символов, word2vec – длиной 40–50 слов.

Эксперименты показывают, что наилучшая точность при использовании алгоритма BERT, веса классов заданы, так как выборка несбалансированная – точность 0,87.

Из результатов экспериментов (табл. 4) видно, что выборка, сформированная на основе авторских символов выражения эмоций и ключевых фраз, является более качественной, чем выборка, сформированная только на основе авторских символов выражения эмоций или выборка, сформированная только на основе ключевых фраз.

Эксперимент 8 показывает, что нейронная сеть, обученная на выборке со стоп-словами, имеет более высокую точность, чем обученная на выборке без стоп-слов.

Для тестирования качества классификации по отдельным классам эмоций была взята модель с наилучшей точностью: предобработка текста с помощью алгоритма BERT, выборка на основе авторских символов выражения эмоций и ключевых фраз, веса классов заданы, длина сообщения – 290–310 символов. Тестовая выборка формируется путём разделения входной выборки на обучающую и тестовую и составляет 30% от входной выборки.

Таблица 3

Формирование обучающей выборки

Эмоция	Текстовых сообщений после 1 этапа	Текстовых сообщений после 2 этапа
Радость	237837	74309
Грусть	7274	2629
Удивление	2739	1535
Страх	4640	2436
Злость	1363	512
Презрение	9960	5613
Отвращение	5011	1206

Таблица 4

Результаты экспериментов

№	Алгоритм	Кол-во постов	Обучающая выборка	Выборка сбалансирована	Веса классов	Точность на обучающей выборке	Точность на тестовой выборке
1	word2vec	1042	смайлы и ключевые слова	нет	нет	0,98	0,77
2		1042	смайлы и ключевые слова	нет	да	0,97	0,79
3	BERT	556	смайлы и ключевые слова	нет	нет	0,95	0,86
4		556	смайлы и ключевые слова	нет	да	0,95	0,87
5		726	смайлы	нет	да	0,94	0,8
6		726	смайлы	нет	нет	0,91	0,82
7		2100	ключевые слова	да	нет	0,87	0,83
8		513	смайлы и ключевые слова, без стоп-слов	нет	да	0,95	0,82

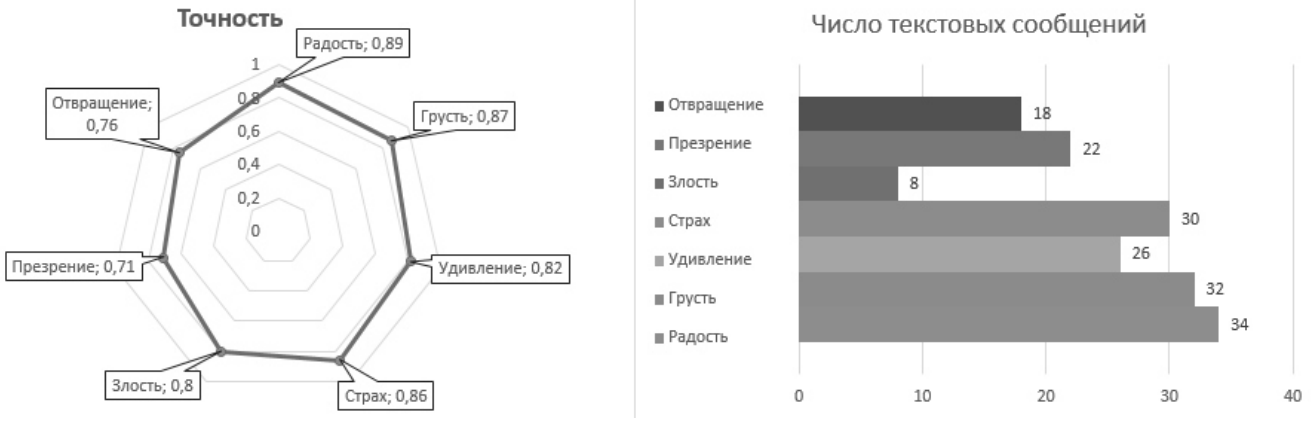

Рис. 6. Точность по каждому классу эмоциональной окраски

Тестовая выборка была разделена по классам эмоциональной окраски. Количество тестовых постов для каждой эмоции различается по тому, что выборка не сбалансирована и посты выбираются по длине в диапазоне 290–310 символов. Тестовая выборка получается небольшой, а в некоторых случаях минимальной из-за того, что посты отбираются по длине в некотором диапазоне. Это нужно, чтобы длины векторов были примерно одинаковыми, иначе нейронная сеть не обучается из-за коротких постов, векторы которых дополняются нулями. Результат эксперимента приведён на рисунке 6.

Как видно из результатов проведенных экспериментов, лучше всего нейронная сеть распознаёт эмоции радости и грусти, а хуже всего – эмоции презрения и отвращения.

Заключение

В результате работы была применена нейронная сеть LSTM архитектуры для определения эмоциональной окраски постов социальной сети. Лучший результат оказался при использовании алгоритма BERT для обработки текста. В ходе исследования был достигнут показатель точности определения эмоциональной окраски постов в 87%.

В будущих исследованиях планируется совершенствовать алгоритм формирования обучающей выборки, в том числе и посредством расширения используемых словарей путем автоматизации процесса их формирования.

СПИСОК ЛИТЕРАТУРЫ

1. Описание информационного образа пользователя социальной сети с учетом его психологической характеристики / Д.А. Власов [и др.] // International Journal of Open Information Technologies. 2018. Т. 6, №. 4. URL: https://cyberleninka.ru/article/n/opisanie-informatsionnogo-obraza-polzovatelya-sotsialnoy-seti-s-uchetom-ego-psihologicheskoy-harakteristiki/viewer (дата обращения: 17.07.2020).

2. Sabuj M.S., Afrin Z., Hasan K.M.A. Opinion Mining Using Vector Machine for Web Based Diverse Data / Pattern Recognition and Machine Intelligence. PReMI 2017. Lecture Notes in Computer Science. Vol. 10597. Springer. pp. 673–678.

3. Dinu L.P., Iuga I. The Best Feature of the Set // Computational Linguistics and Intelligent Text Processing. CICLing 2012. Lecture Notes in Computer Science / Gelbukh A. (eds) Vol. 7181. Springer. pp. 556–567.

4. Chetviorkin I.I., Loukachevitch N.V. Sentiment Analysis Track at ROMIP-2012 // Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2013» : сб. науч. ст. 2013. Т. 2. С. 40–50.

5. Мошкин В.С., Андреев И.А. Сравнение эффективности применения алгоритмов сентимент-анализа неструктурированных ресурсов социальных сетей // Восьмая Междунар. конф. «Системный анализ и информационные технологии» САИТ – 2019 : тр. конф. М. : ФИЦ ИУ РАН, 2019. С. 534–540.

6. Chen, Qufei and Marina Sokolova. “Word2Vec and Doc2Vec in Unsupervised Sentiment Analysis of Clinical Discharge Summaries.” CoRR abs / 1805.00352 (2018).

7. Антонова А., Соловьев А. Использование метода условных случайных полей для обработки текстов на русском языке // Компьютерная лингвистика и интеллектуальные технологии : «Диалог-2013». сб. науч. ст. М. : Изд-во РГГУ, 2013. Вып. 12 (19). С. 27–44.

8. Learning word vectors for sentiment analysis / A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts // The International Language Technologies. 2011, June. Vol. 1. International Association for Computational Linguistics. pp. 142–150.

9. Богданов А.Л., Дуля И.С. Сентимент-анализ коротких русскоязычных текстов в социальных медиа // Вестн. Том. гос. ун-та. Экономика. 2019. № 47. С. 210–249.

10. Смирнова О.С., Шишков В.В. Выбор топологии нейронных сетей и их применение для классификации коротких текстов // International Journal of Open Information Technologies. 2016. № 8. С. 50–54.

11. Колмогорова А.В., Вдовина Л.А. Лексико-грамматические маркеры эмоций в качестве параметров для сентимент-анализа русскоязычных интернет-текстов // Вестник Пермского университета. Российская и зарубежная филология. 2019. № 3. С. 38–46.

12. Moshkin V., Yarushkina N., Andreev I. The Sentiment Analysis of Unstructured Social Network Data Using the Extended Ontology SentiWordNet // IEEE, 12th International Conference on Developments in eSystems Engineering (DeSE). Kazan, Russia, 2019. pp. 576–580.

13. Bert: Pre-training of deep bidirectional transformers for language understanding / J. Devlin et al. // arXiv preprint arXiv:1810.04805. 2018. URL: https://arxiv.org/pdf/1810.04805.pdf (дата обращения: 17.07.2020).

14. WordNetAffect. URL: http://wndomains.fbk.eu/wnaffect.html (дата обращения: 17.07.2020).

15. Rani Horev BERT Explained: State of the art language model for NLP. URL: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 (дата обращения: 17.07.2020).

16. Алгоритм Word2Vec. URL: https://neurohive.io/ru (дата обращения: 17.07.2020).

17. Understanding LSTM Networks. URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs (дата обращения: 17.07.2020).

18. Illustrated Guide to LSTM’s and GRU’s: A step by step explanation. URL: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21 (дата обращения: 17.07.2020).

REFERENCES

1. Vlasov D.A. et al. Opisanie informatsionnogo obraza polzovatelia sotsialnoi seti s uchetom ego psikhologicheskoi kharakteristiki [Description of User’s Informational Image of the Social Network with Considering his Psychological Characteristic]. *International Journal of Open Information Technologies, 2018, vol. 6, no. 4.* Available at: https://cyberleninka.ru/article/n/opisanie-informatsionnogo-obraza-polzovatelya-sotsialnoy-seti-s-uchetom-ego-psihologicheskoy-harakteristiki/viewer.

2. Sabuj M.S., Afrin Z., Hasan K.M.A. Opinion Mining Using Vector Machine for Web Based Diverse Data. *Proc. of Conf. on Pattern Recognition and Machine Intelligence. PReMI-2017. Lecture Notes in Computer Science.* Springer-Verlag, vol. 10597, pp. 673–678.

3. Dinu L.P., Iuga I. The Best Feature of the Set. *Proc. of Conf. on Computational Linguistics and Intelligent Text Processing. CICLing-2012. Gelbukh A. (eds). Lecture Notes in Computer Science.* Springer-Verlag, vol 7181, pp. 556–567.

4. Chetviorkin I.I., Loukachevitch N.V. Sentiment Analysis Track at ROMIP-2012. *Kompiuternaia lingvistika i intellektualnye tekhnologii: “Dialog-2013”. Sb. nauch. st.* [Proc. of Int. Conf. on Computational Linguistic and Intellectual Technologies “Dialog-2013”]. 2013, vol. 2, pp. 40–50.

5. Moshkin V.S., Andreev I.A. Sravnenie effektivnosti primeneniia algoritmov sentiment-analiza nestrukturirovannykh resursov sotsialnykh setei [Comparison of Algorithm Efficiency for Sentiment Analysis of Unstructured Social Network Resources]. *Vosmaia Mezhdunar. konf. “Sistemnyi analiz i informatsionnye tekhnologii”. SAIT-2019. Tr. konf.* [Proc. of the 8th Int. Conf. on Systems Analysis and Information Technologies. SAIT-2019]. Moscow, FITs IU RAN Publ., 2019, pp. 534–540.

6. Chen, Qufei and Marina Sokolova. Word2Vec and Doc2Vec in Unsupervised Sentiment Analysis of Clinical Discharge Summaries. *CoRR abs, 1805.00352*, 2018.

7. Antonova A., Solovev A. Ispolzovanie metoda uslovnykh sluchainykh polei dlia obrabotki tekstov na russkom iazyke [Conditional Random Field Models for the Processing of Texts in Russian]. *Kompiuternaia lingvistika i intellektualnye tekhnologii: “Dialog-2013”. Sb. nauch. st.* [Proc. of Int. Conf. on Computational Linguistic and

Intellectual Technologies “Dialog-2013”]. Moscow, RGGU Publ., 2013, iss. 12 (19), pp. 27–44.

8. Maas A.L., R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts. Learning Word Vectors for Sentiment Analysis. *The International Language Technologies.* International Association for Computational Linguistics, 2011, vol. 1, pp. 142–150.

9. Bogdanov A.L., Dulia I.S. Sentiment-analiz korotkikh russkoiazychnykh tekstov v sotsialnykh media [Sentiment Analysis for Short Russian Texts in Social Media]. *Vestn. Tom. gos. un-ta. Ekonomika* [Tomsk State University Journal of Economics], 2019, no. 47, pp. 210–249.

10. Smirnova O.S., Shishkov V.V. Vybor topologii neironnykh setei i ikh primenenie dlia klassifikatsii korotkikh tekstov [The Choice of the Topology of Neural Networks and Their Use for the Classification of Small Texts]. *International Journal of Open Information Technologies*, 2016, no. 8, pp. 50–54.

11. Kolmogorova A.V., Vdovina L.A. Leksiko-grammaticheskie markery emotsii v kachestve parametrov dlia sentiment-analiza russkoiazychnykh internet-tekstov [Lexical and Grammatical Markers of Emotions as Parameters for Sentiment Analysis of Internet Texts in Russian]. *Vestnik Permskogo universiteta. Rossiiskaia i zarubezhnaia filologiia* [Perm University Herald. Russian and Foreign Philology], 2019, no. 3, pp. 38–46.

12. Moshkin V., Yarushkina N., Andreev I. The Sentiment Analysis of Unstructured Social Network Data Using the Extended Ontology SentiWordNet. *IEEE, 12th International Conference on Developments in eSystems Engineering (DeSE).* Kazan, Russia, 2019, pp. 576–580.

13. Devlin J. et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805. 2018.* Available at: https://arxiv.org/pdf/1810.04805.pdf (accessed 17.07.2020).

14. *WordNetAffect.* Available at: http://wndomains.fbk.eu/wnaffect.html (accessed 17.07.2020).

15. Rani Horev. *BERT Explained: State of the Art Language Model for NLP.* Available at: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 (accessed 17.07.2020).

16. *Algoritm Word2Vec.* Available at: https://neurohive.io/ru (accessed 17.07.2020).

17. *Understanding LSTM Networks.* Available at: http://colah.github.io/posts/2015-08-Understanding-LSTMs (accessed 17.07.2020).

18. *Illustrated Guide to LSTM’s and GRU’s: A Step by Step Explanation.* Available at: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21 (accessed 17.07.2020).