Сравнение применения различных нейронных сетей в задачах сентимент-анализа данных социальных сетей

А.А. Константинов[1], В.С. Мошкин[1], Н.Г. Ярушкина[1]

[1]*Ульяновский государственный технический университет, Северный Венец 32, Ульяновск, Россия, 432027*

Аннотация

В работе описывается результаты экспериментов по сравнению применения различных архитектур нейронных сетей в задачах определения эмоциональной окраски текстовых сообщений социальной сети с применением двух алгоритмов векторизации текста «word2vec» и «BERT». В ходе исследования был достигнут показатель точности определения эмоциональной окраски постов в 87%.

Ключевые слова

Сентимент-анализ, социальная сеть, нейронная сеть, обучение

1. Введение

Исследование социальных сетей с каждым годом приобретает все большую актуальность в связи с обостряющейся необходимостью обеспечения безопасности населения и мониторинга общественных настроений. Анализ сообщений 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.

Данная работа является развитием проекта [1] и затрагивает применение нейронных сетей различных архитектур для решения задачи анализа тональности сообщений в социальных сетях.

2. Применение машинного обучения в сентимент-анализе

В рамках данного проекта использовался алгоритм формирования обучающей и тестовой выборки для обучения нейронных сетей, описанный в [2]. Формально процесс отбора постов можно представить схемой, показанной на рисунке 1.

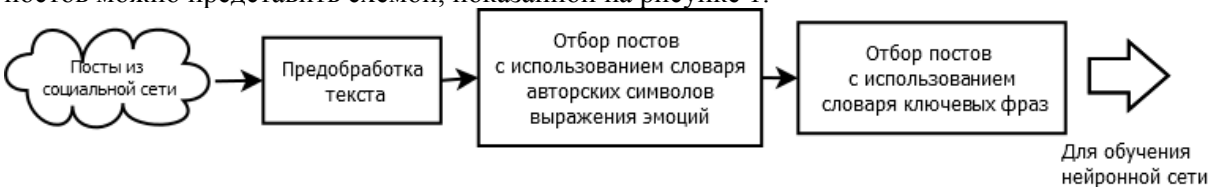

Рисунок 1. Алгоритм отбора постов для обучения нейронных сетей

Для представления слов в векторном пространстве были использованы два метода: word2vec и «BERT»[3][4]. В ходе работы были использованы различные архитектуры нейронных сетей для определения тональности текстов: LSTM, Bidirectional LSTM, CNN, MLP [5] [6].

При решении задачи формирования обучающей и тестовой выборок было автоматически обработано 2,5 млн. текстовых сообщений из открытых групп социальной сети «ВКонтакте». Для экспериментов были взяты наилучшие параметры, представленные в таблице 1.

В результате проведения экспериментов была получена точность для каждой архитектуры нейронной сети на тестовой выборке. Точность классификации на обучающей выборке у всех архитектура равна 1.0. Результаты экспериментов представлены в таблице 2.

Таблица 1
Параметры

Параметр	Значение
Обработка текста	Лемматизация
Длина поста	90-110 символов
Удалять стоп-слова	Не удалять
Метод векторизации	BERT

Таблица 2
Точность на тестовой выборке

Нейронная сеть	Точность
LSTM (рекуррентная)	0,82
Bidirectional LSTM (двунаправленная рекуррентная)	0,84
CNN (свёрточная)	0,86
MLP (многослойный персептрон)	0,87
GRU (рекуррентная)	0,81
Bidirectional GRU (двунаправленная рекуррентная)	0,84
LSTM & CNN (свёрточная и рекуррентная)	0,85

3. Заключение

В результате работы были применены нейронные сети различных архитектур для определения эмоциональной окраски постов социальной сети. Лучший результат оказался при использовании многослойного персептрона для классификации текстов. В ходе исследования был достигнут показатель точности определения эмоциональной окраски постов в 87%.

Стоит отметить, что быстрее всего обучаются MLP и CNN (за 25-30 эпох), дольше всего LSTM и GRU (за 50-60 эпох). Рекуррентные нейронные сети более затратные по памяти и вычислениям. В будущих исследованиях планируется совершенствовать алгоритм формирования обучающей выборки, в том числе и посредством расширения используемых словарей путем автоматизации процесса их формирования.

4. Благодарности

Работа выполнена при финансовой поддержке РФФИ, гранты № 18-47-730035 и 18-47-732007, а также Минобрнауки России в рамках проекта № 075-00233-20-05 от 03.11.2020 «Исследование интеллектуального предиктивного мультимодального анализа больших данных и извлечения знаний из различных источников».

5. Литература

[1] Moshkin, V. Application of the bert language model for sentiment analysis of social network posts / V. Moshkin, A. Konstantinov, N. Yarushkina // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). – 2020. – Vol. 12412 LNAI. – P. 274-283.

[2] Konstantinov, A. An approach to the training dataset formation for assessing the sentiment degree of social network posts using machine learning / A. Konstantinov // CEUR Workshop Proceedings. – 2020. – Vol. 2667. – P. 211-214.

[3] Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding / J. Devlin // arXiv preprint. – 2018 [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/1810.04805.pdf (дата обращения: 06.01.2021).

[4] Алгоритм Word2Vec. [Электронный ресурс]. – Режим доступа: https://neurohive.io/ru (дата обращения: 06.01.2021).
[5] Illustrated Guide to LSTM’s and GRU’s: A step by step explanation [Электронный ресурс]. – Режим доступа: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21 (дата обращения: 06.11.2020).
[6] A Comprehensive Guide to Convolutional Neural Networks [Электронный ресурс]. – Режим доступа: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 (дата обращения: 06.01.2021).