СТАТИСТИЧЕСКИЙ АНАЛИЗ ТЕКСТА И ИССЛЕДОВАНИЕ ЕГО ЦЕЛЕВОГО НАЗНАЧЕНИЯ

В. Д. Клюев

Белорусский государственный университет, г. Минск;
vladislav.klyuev.work@gmail.com;
науч. рук. – А. А. Лагуто, ст. преп.

Целью данной работы является построение моделей машинного обучения на основании статистических признаков текста, а также программная реализация методов классификации текста.

Ключевые слова: статистический анализ текста, машинное обучение, анализ тональности текста, обработка естественного языка, классификация данных.

ВВЕДЕНИЕ

Сентимент-анализ или анализ тональности текста является областью компьютерной лингвистики, позволяющей автоматизировать нахождение в текстах эмоционально окрашенной лексики, эмоциональных оценок авторов. Одной из подзадач анализа тональности текста является определение полярности текста: позитивно или негативно окрашен исследуемый набор текстов, т.е. классификация текстов. Данная работа посвящена решению задачи бинарной классификации по 4 эмоциям с помощью классификаторов машинного обучения без нейронных сетей и с их использованием.

НАБОРЫ ДАННЫХ И МЕТОДЫ

Наборов текстов с размеченными эмоциями относительно их контекста мало, и они имеют много минусов. Некоторые из них не являются достаточно надежными, так как основаны на потенциально недобросовестной разметке данных, и у каждого источника есть свой способ классификации. Поэтому сложно установить критерии оценки определенной эмоции.

Таким образом, возникает проблема поиска наборов данных для проведения анализа тональности текста. Для решения данной проблемы был создан свой набор данных на основе коротких сообщений из социальной сети Twitter длиной в 140 символов. Данные использовались на английском языке, так как в дальнейшем для обучения будут использоваться наборы с данными на английском языке.

В контролируемом обучении набор данных должен иметь метку для классификации данных. Поэтому были использованы встроенные элементы, определяющие эмоцию сообщения: хэштеги (например, #злость, #радость) и символы эмоций (англ. emoji, например, □). Это простой и достаточно надежный способ разметки данных, так как сам автор коротко обозначает свои эмоции. Используя распространенные символы эмоций и хэштеги, составляется таблица соотношения эмоций с символами эмоций и хэштегами. Для интерпретации символов эмоций в тексте была использована сторонняя библиотека Emoji.

Пример. Необходимо найти эмоции страха и злости. В качестве набора запросов для этих эмоций использованы следующие хэштеги: #страх и #ужас, а также символы эмоций □ и □, приведенные в таблице 1.

Таблица 1

Примеры символов эмоций, полученных по эмоции злости

😡(:pouting_face:)	24301
□(:face_with_symbols_on_mouth:)	1255
😠(:angry_face:)	671
😂(:face_with_tears_of_joy:)	610
😤(:face_with_steam_from_nose:)	496
😭(:loudly_crying_face:)	455
🙄(:face_with_rolling_eyes:)	355

Первые две строки являются начальными символами эмоций. Остальные символы анализируются как потенциально подходящие под эмоцию в дальнейшем. Аналогичная операция была произведена и с хэштегами: был сформирован необходимый набор данных с размеченными эмоциями.

Для повышения доверия к созданной разметке использовался анализ тональности текста, с помощью которого были сделаны выводы о достоверности выборки. Имели место следующие суждения: если текст содержит метку злость, то он будет иметь отрицательную полярность после предсказания модели.

В качестве полярного набора данных использовалась база данных – Sentiment140. Данные этой базы также были собраны из социальной сети Twitter и содержат тексты, помеченные как отрицательные (0), нейтральные (2) или положительные (4).

Предварительная обработка текстов включала в себя преобразование текста в нижний регистр, удаление специальных символов, преобразование символов эмоций в текстовое представление, удаление повторов, замена сокращенных отрицаний, удаление стоп-слов. Затем было выполнено преобразование текста в векторное представление, основанное на оценке TF-IDF. Далее была определена модель машинного обучения –

GRU, или модель управляемых рекуррентных блоков. Для упрощения выходные данные были представлены в двоичной классификации: 1 – сообщения с положительной меткой (4), 0 – сообщения с отрицательной меткой (0), нейтральных значений в обучении не будет. Т.е. чем более отрицательный текст, тем больше выходное значение будет стремиться к 0. В результате получаем модель с бинарной классификацией.

Для проведения анализа на соответствие выборки по четырем эмоциям, были использованы среднее, максимальное, минимальное значения, а также стандартное отклонение [1]. В таблице 2 представлены результаты, сгруппированные по эмоциям.

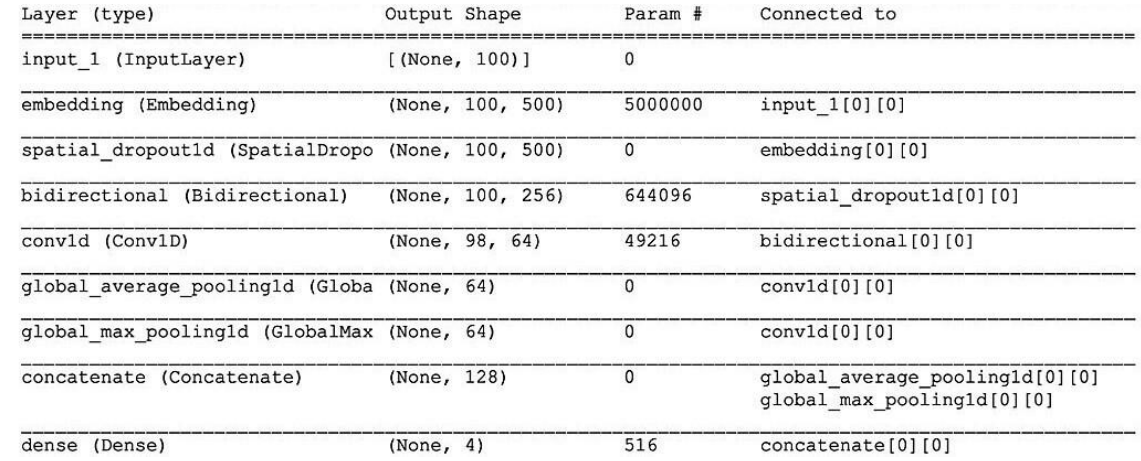
Таблица 2

Значения анализа выборки по 4 эмоциям

	Эмоция	Среднее значение	Максимальное значение	Минимальное значение	Стандартное отклонение
0	Радость (joy)	0.791465	0.995362	0.016203	0.213368
1	Злость (anger)	0.458358	0.992345	0.005212	0.259689
2	Грусть (sadness)	0.244089	0.991080	0.001725	0.226408
3	Страх (fear)	0.498132	0.995031	0.005922	0.255033

Согласно полученным данным, можно сделать вывод, что данные являются скорее корректными, так как для эмоции радость среднее значение сгруппированного результата ближе к 1, чем к 0. Аналогичный, но обратный вывод, может быть сделан об эмоции грусти: среднее значение сгруппированного результата эмоции находится ближе к значению 0. Полученные значения являются ожидаемыми, поскольку эмоцию радости можно классифицировать как положительную, а эмоцию грусти – как отрицательную. Отметим, что сбалансированность выборки по количеству представленных эмоций была соблюдена.

В качестве архитектуры модели машинного обучения для выделения эмоции из подготовленных ранее данных использована двунаправленная модель LSTM со слоем CNN [2], представленная рисунком:



Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 100)]	0	
embedding (Embedding)	(None, 100, 500)	5000000	input_1[0][0]
spatial_dropout1d (SpatialDropo	(None, 100, 500)	0	embedding[0][0]
bidirectional (Bidirectional)	(None, 100, 256)	644096	spatial_dropout1d[0][0]
conv1d (Conv1D)	(None, 98, 64)	49216	bidirectional[0][0]
global_average_pooling1d (Globa	(None, 64)	0	conv1d[0][0]
global_max_pooling1d (GlobalMax	(None, 64)	0	conv1d[0][0]
concatenate (Concatenate)	(None, 128)	0	global_average_pooling1d[0][0] global_max_pooling1d[0][0]
dense (Dense)	(None, 4)	516	concatenate[0][0]

Рис. Конечная архитектура двунаправленной модели LSTM со слоем CNN

Суть такого подхода в том, что модель LSTM собирает информацию о контексте предложения в тексте, а слой CNN извлекает отличительные особенности.

РЕЗУЛЬТАТЫ

В работе также была произведена классификация с помощью алгоритмов, не использующих нейронные сети обучения: «Случайный лес», «Полиномиальная логистическая регрессия», «K ближайших соседей» и «Метод стохастического градиентного спуска».

В таблице 3 представлены усредненные значения каждой из оценочных мер [3] для различных методов классификации, где ДПКО – общая доля правильно классифицированных объектов.

Таблица 3

Оценочные значения различных классификаторов

Классификатор	Точность	Полнота	F-мера	ДПКО
CNN+LSTM	0.84	0.87	0.86	85.53
Логистическая регрессия	0.67	0.67	0.66	66.58
Градиентного стохастического спуска	0.66	0.66	0.66	65.57
Случайный лес	0.64	0.64	0.64	64.02
K ближайших соседей	0.58	0.58	0.57	57.81

Отметим, что подход с использованием нейронных сетей с архитектурой LSTM и CNN для исследуемого набора данных, достигает лучших результатов, чем использование традиционных классификаторов.

Данная работа имеет ряд ограничений, которые нельзя не принимать во внимание:

- эксперимент проводился с 4-мя категориями эмоций;
- случайное разделение на обучающую и тестирующую выборки;
- использование TF-IDF как метода выбора признаков;
- ограниченная длина текстов в наборе данных (140 символов).

Библиографические ссылки

1. *Форсайт Д. А., Понс Ж*. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М., 2004.
2. *Xingyou Wang, Weijie Jiang, Zhiyong Luo*. Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short // Proceedings of COLING 2016. Osaka, 2016. P. 2428–2437.
3. *Шитиков В.К., Мастицкий С.Э*. Классификация, регрессия и другие алгоритмы Data Mining с использованием R. // Атрибуция. – Тольятти, 2017. – 351 с.