СРАВНЕНИЕ РАЗЛИЧНЫХ ПОДХОДОВ К АНАЛИЗУ ТЕКСТА НА ПРИМЕРЕ ЗАДАЧИ ПРЕДСКАЗАНИЯ ОЦЕНКИ РЕСТОРАНА ПО ОТЗЫВУ ПОСЕТИТЕЛЯ

А.А. Шлеменков
Студент БГУИР

Белорусский государственный университет инфоматики и радиоэлектроники, Республика Беларусь
E-mail: alex.shlemenkov@gmail.com

Abstract. . In this research paper two common methods of texts analysis had been applied to the problem of restaurant review mark prediction. The results of the methods applied to the task had been analyzed and had been listed in table for comparison; reasons of different performance for specific pair of data and problem had been proposed.

Отслеживая тренды современного мира, можно заметить глубокий интерес к области искусственного интеллекта. Одной из областей в «машинном интеллекте» является естественная обработка языков (Natural Language Processing или NLP). Важно заметить, что область NLP является полезной не только для людей, которые тесно связаны с лингвистикой и языками, но и для бизнеса. Например, решив задачу поиска отрицательных отзывов на продукт, можно более оперативно реагировать на изменения.

В данной работе была рассмотрена задача, которая состоит в том, чтобы по тексту отзыва, который оставил посетитель, предсказать оставленную им оценку. Данные представляют собой текст отзыва и оценку в диапазоне от 1 до 5 (5 – лучший отзыв, 1 – худший) [1].

В ходе исследования были проанализированы два популярных метода анализа текстов. Один из них является «классическим» и основан на технике TF-IDF, другой был предложен относительно недавно и называется Word2Vec [2].

TF-IDF – статистическая мера, которая используется для оценки важности слова или сочетания из нескольких подряд идущих слов, которые, по сути, объединяются в одно уникальное. Выделение таких сочетаний часто очень полезно, так как позволяет «уловить» смысл отрицаний или устойчивых сочетаний, используемых в языке. Например, сочетание «не нравится» и слово «нравится» будут иметь абсолютно разный смысл в контексте документа, но если не учитывать такие фразы, то качество классификатора может сильно упасть. Мера TF-IDF каждого слова прямо пропорциональна количеству появлений в документе и обратно пропорциональна частоте появления слова во всех документах коллекции. Таким образом, чем чаще слово появляется в документе, тем выше его TF-IDF. И наоборот, если слово часто встречается во всех документах коллекции, например, общеупотребительная лексика, то даже с большим количеством появлений в документе значение TF-IDF этого слова будет мало. Данный подход позволяет отфильтровать часто встречающиеся элементы общеупотребительной лексики и, с другой стороны, выделить слова, которые встречаются редко во всем наборе, но часто в отдельных типах документов.

Для дополнительного сравнения был использован лемматизированный текст на входе. Общий смысл лемматизации заключается в приведении слов к начальной форме.

Способ анализа текстов под названием Word2Vec рассматривает проблему с другой стороны. В нем делается предположение о том, что слова, которые встречаются в схожих контекстах, имеют схожий смысл. Word2Vec каждому слову в коллекции ставит в соответствие вектор некоторой размерности (обычно это 100, 300, но размер вектора зависит от объема базы текстов, на которой был обучен Word2Vec). Этот вектор имеет смысл некоторой координаты в пространстве слов. Важное замечание состоит в том, что при изначальном предположении о контекстуальной близости слов получается, что вектора со схожим «смыслом» располагаются «рядом», а также существует возможность производить над такими векторами различные операции. Классическим примером является следующий: «король» - «мужчина» + «женщина» ~

«королева».

Оба метода были применены к данным. Для вычисления векторов слов был обучен Word2Vec на всех тренировочной базе отзывов, а после были усреднены вектора слов и обучена двуслойная нейронная сеть. В результате применения «классического» подхода использовался линейный классификатор с L1 регуляризацией. Результаты данного эксперимента в виде количества правильно классифицированных отзывов и матрицы ошибок представлены в таблице 1 и таблице 2:

Таблица 1. Матрицы ошибок

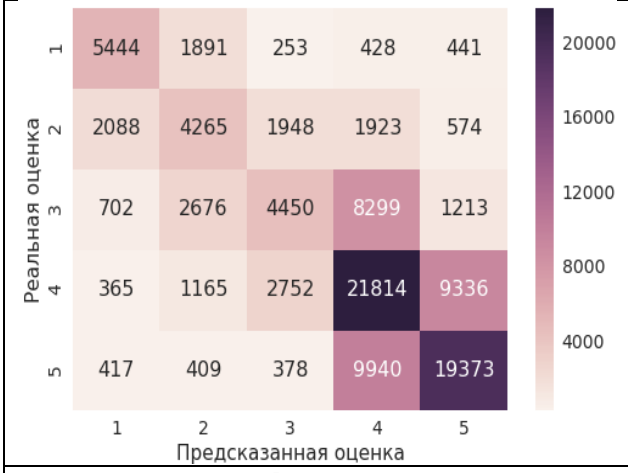	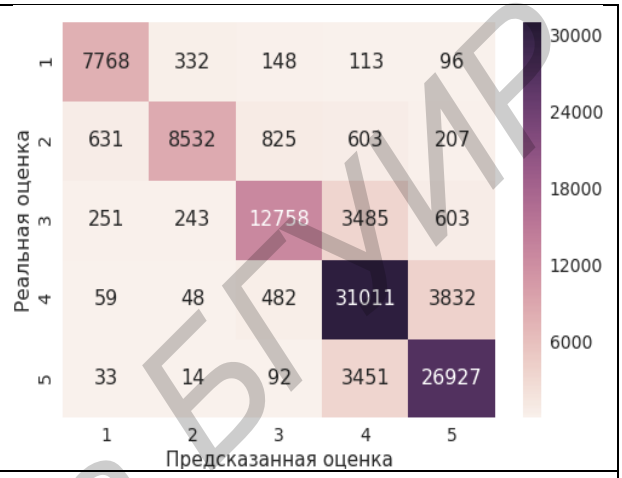
Word2Vec + NN	TF-IDF + LR

Таблица 2. Доля верно классифицированных отзывов

Используемый метод	Доля верно классифицированных отзывов, %
TF-IDF + LR	61,005
Lemmatization + TF-IDF + LR	58,23
Word2Vec + NN	54,852
Lemmatization + Word2Vec + NN	53,726

В таблице 1 в каждой из матриц ошибок на пересечении строки i (реальной оценки) и столбца j (предсказанной оценки) находится число, отражающее количество отзывов, которые были предсказаны классификатором как относящиеся к классу j, хотя на самом деле отзывы принадлежат классу i. Просуммировав все элементы на диагонали и поделив на сумму элементов в матрице можно получить долю правильно предсказанных оценок на тестовой выборке. Стоит отметить, что матрицы ошибок хорошо подходят для визуализации работы классификатора именно потому, что могут показать, где именно ошибается алгоритм.

Несмотря на новизну подхода, основанного на Word2Vec, как видно по результату, он работает не всегда хорошо. Объяснений этому может быть несколько: при усреднении векторов слов смысл их «размывается» слишком сильно. Это оставляет очень мало информации алгоритму для выделения каких-либо связей отзыва с оценкой. Также стоит отметить, что для каждого слова вычисляется его вектор, что не позволяет распознать «смысл» таких фраз, как, например, отрицания. Получается, что алгоритм не только не учитывает нужный «смысл» слов, но и учитывает его с обратным знаком. Подход, основанный на TF-IDF, сработал лучше из-за нескольких причин, самая важная которых: он учитывает сочетания слов. Следовательно, такие конструкции как «не нравится» распознаются как нечто отрицательное.

Заметим, что лемматизация несколько ухудшила результат. Это можно объяснить тем,

что при приведении слов к начальной форме теряется часть потенциально полезной информации.

Литература

[1]. Determine restaurant review sentiment [Электронный ресурс] – Режим доступа: https://in-class.kaggle.com/c/sentiment-analysis2

[2]. Distributed Representations of Words and Phrases and their Compositionality [Электронный ресурс] – Режим доступа: https://arxiv.org/pdf/1310.4546.pdf