Part I
Literature review

В этой части работы будут рассмотрены статьи других авторов, напрямую связанные с простым анализом тональности текста, анализом текста, содержащегося в отзывах о покупках на различных сайтах и анализом тональности текста, содержащегося в социальных сетях, таких так Twitter.

3 автора в [7] провели исследование по анализу тональности наборов данных, в состав которых входили: Cornell Movie Dataset, Twitter Dataset, Amazon Products Dataset и IMDB Dataset, используя различные методы машинного обучения.

В работе рассматривались следующие методы: Naive Bayes (NB), стохастический градиентный спуск (Stochastic gradient descent - SGD), метод опорных векторов (support vector machine - SVM), пассивный агрессивный метод (Passive Aggressive - PA), Максимальной энтропии (Maximum Entropy), Adaptive Boosting (AdaBoost), Multinomial Naïve Bayes (MNB), Bernoulli Naïve Bayes (BNB), Ridge Regression (RR) и Logistic Regression (LR). В результате использования данных методов вместе с методами по извлечению признаков из наборов данных: TF-IDF и N-Грамм (юниграмма, биграмма и триграмма), было получено, что использование алгоритмов PA и RR по сравнению с остальными методами дают одни из самых высоких результатов классификации: от 87% до 99.96%, полученный при использовании метода PA. Также было выделено, что метод логистической регрессии и метод опорных векторов также давали приемлемые результаты: 87.56% и 85.76% соответственно.

2 автора в [8] проводят анализ тональности тектста (положительный/отрицательный), используя набор данных, состоящий из отзывов, оставленных пользователями на сайте Amazon об электрических товарах (Kindle, DVD и других) в период с Февраля 2012 года по Июль 2017 года. Для извлечения признаков из данных использовался метод TF-IDF, а для обучения модели были рассмотрены методы машинного обучения: 3 разновидности Naive Bayes классификатора: обычный метод Naive Bayes, Multinomial Naive Bayes и Bernoulli Naive Bayes, а также логистическая регрессия. В итоге были получены следующие результаты: Multinomial Naive Bayes - 92.87%, Bernoulli Naive Bayes - 92.35% и Логистиская регрессия - 93.34%.

Те же самые авторы в [9] изучали применение методов по анализу тональности текста, используя библиотеки языка программирования R по машинному обучению применительно к отзывам покупателей о приобретённой еде на сайте Amazon. В работе описан алгоритм при помощи которого проводилась классификация отзывов, а также приведены облаки слов, которые имели больший вес в модели при определении тональности.

ДОДЕЛАТЬ:

4 автора в [11] провели анализ тональности отзывов об отелях, расположенных в Нью-Йорке. В основе анализа лежал подход, основанный на использовании словаря тональности (Lexicon-based), в работе происходила классификация

отзывов на три класса: положительные, нейтральные и отрицательные.

Перейдём к рассмотрению работ, связанных с анализом текста в социальных сетях.

В работе [10] авторы применяют метод Modified Balanced Winnow к трём наборам данных: Sanders Corpus (Twitter), STS_Gold (Twitter), SentiStrength (MySpace, YouTube, BBC и другие).