Part I
Literature review

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

Начнём с рассмотрения обзорной статьи 3 авторов [7], в которой они изучили все имеющиеся работы, обсудили возможные алгоритмы как по выделению признаков, так и по построению моделей, способных определять тональность. В общей сложности, в работе было рассмотрено около 54 статей разных авторов. Говоря об анализе тональности, авторы постоянно возвращаются к рисунку 1, который описывает существующие и использующиеся сейчас техники по анализу тональности текста.

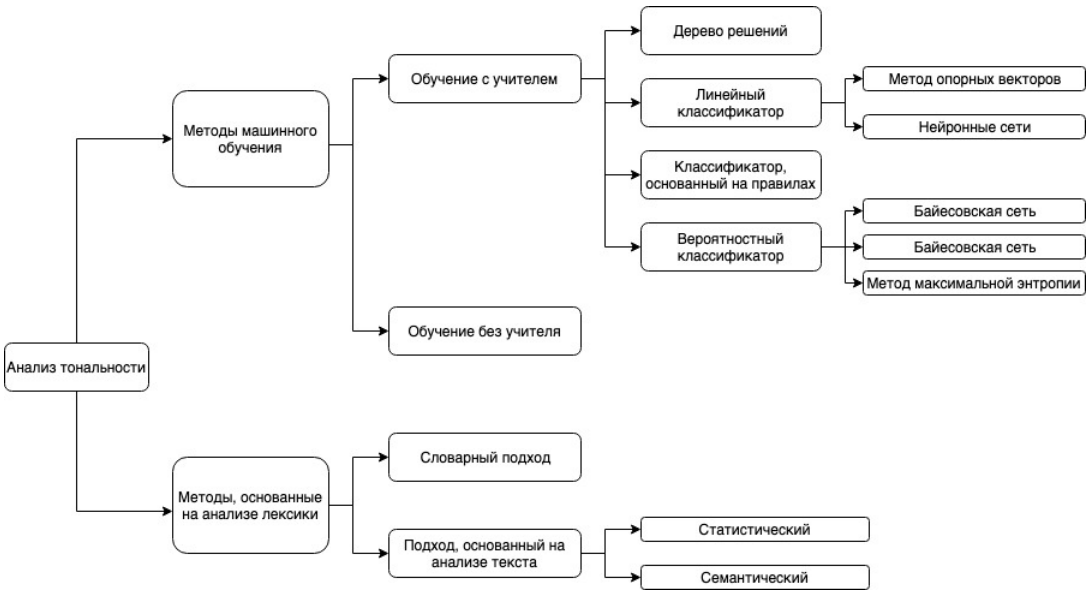

Figure 1: Техники по анализу тональности текста

Также авторы рассмотрели основные методы по извлечению признаков из текста:

1. Методы, в основе которых лежит измерение корреляции между словами и классами:

    - Метод поточечной взаимной информации (от англ. Point-wise Mutual Information (PMI)): метод основан на введении и использовании меры, которая показывает контекстную связь между признаками

и классами. Пример использования: метод использовался в работе Yu и Wu, где авторы использовали данный метод, чтобы расширить первоначальный набор слов, который был собран из небольшого корпуса, посвященного новостям с фондового рынка. Модель превзошла другие разновидности PMI, так как учитывала не только силу совпадения рассматриваемых слов, но и их контекстное распределение;

- Хи-квадрат (от англ. Chi-square (χ^2)): метод также отвечает за поиск величины связи конкретного слова и класса, однако лучше PMI в том, что внутри него используются нормализованные значения, соответственно, эти значения больше подходят к терминам в той же категории. Пример использования: в результате исследования Фаном и Чангом контекстной рекламы, ими было показано, что их метод может выделить рекламные объявления, которые непосредственно коррелируют с интересами блогера. Для выделения признаков в работе использовался метод хи-квадрат, тогда как для модели классификации использовался метод опорных векторов, изображённый на рисунке 1;

2. Методы, суть которых заключается в уменьшении размерности данных:

   - Латентно-семантический анализ (от англ. Latent Semantic Indexing (LSI)): суть метода в том, что

3. Методы, основанные на статистическом подходе:

   - Hidden Markov Model (HMM)
   - Latent Dirichlet Allocation (LDA)

4. Другие методы:

   - Метод Мешка слов (от англ. Bag of Words (BoW)):как метод, так и пример его использования уже были описаны в данной работе;

3 автора в [8] 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 [9] 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 [10] 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.

Ещё 1 работа по анализу тональности отзывов с сайта Amazon является работа [11], где авторы предлагают алгоритм по обработке отзывов, призванный разбить оставленные пользователями отзывы на две категории:

- Отзывы о представленном сервисе;
- Отзывы о характеристиках продукта;

Далее авторами используется собственный алгоритм, который проводит анализ отзывов на предмет того, является ли отзыв прямым или косвенным. Например: “Камера не лучше, чем у iPhone 4s” - косвенный отзыв, фразы из предложений сравниваются с заранее выделенными в ходе ручной работы с отзывами фразами, по итогу чего и происходит классификация каждого отзыва. Также интересна обработка отрицаний в статье авторов, когда алгоритм встречает в отзыве фразу “не хороший товар”, то сначала алгоритм анализирует тональность слова хороший (стоит также отметить, что он ищет следующую связку: частицу, обозначающую отрицание и прилагательное). После чего, сначала, при помощи словаря тональности определяется тональность прилагательного и, так как ещё есть частица, означающая отрицание, то данной фразе проставляется противоположная тональность.

Алгоритм работы с отзывами, после извлечения набора фраза из предложений, может быть представлен следующим образом:

1. Все фразы проверяются на то, являются ли они косвенными или прямыми относительно продукта;
2. Проверяются на наличие с частицей “не”;
3. Если фраза не прошла первые два пункта, то она разбивается на отдельные слова, где при помощи словаря тональности определяется конечная тональность;

Кроме того, в работе был описан метод обработки сырых отзывов, скачанных напрямую с сайта Amazon.

ДОДЕЛАТЬ:

// 4 автора в [13] 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 [12] авторы применяют метод Modified Balanced Winnow к трём наборам данных: Sanders Corpus (Twitter), STS_Gold (Twitter), SentiStrength (MySpace, YouTube, BBC и другие).