DOI: 10.46548/CIT-2021-0034-0015

А.А. Зоткина, А.И. Мартышкин

АНАЛИЗ МЕТОДОВ ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ ТЕКСТОВЫХ ДАННЫХ ПОЛЬЗОВАТЕЛЯ СОЦИАЛЬНЫХ СЕТЕЙ

Россия, г. Пенза, Пензенский государственный технологический университет

This article discusses methods for analyzing the information space using the "Sentiment Analysis" method. It is noted that an increased interest in tone analysis of information data of social network users began in the 2000s and continues to this day. Sentiment analysis has been studied in the works of both domestic and foreign scientists and studies. The article presents two popular methods for determining the sentiment of a text: unsupervised learning and the statistical method. Examples of each of them are given. An example of data parsing code is considered.

Введение. В настоящее время социальные сети стремительно набирают обороты. На сегодняшний день уже более миллиона людей зарегистрированы в той или иной социальной сети, где они делятся информацией о своей жизни, интересах, об окружающем мире со своей аудиторией. Данные, которыми активно делится человек, могут быть использованы для определения интересов как бизнеса и государства, так и отдельной личности. Имея расширенные данные об интересах пользователя, можно определять различные отклонения в психике, определять потенциальных преступников, предотвращать различные нарушения со стороны пользователя социальной сети и прогнозировать конфликты.

Анализ тональности текста. Для анализа информационного пространства пользователя социальной сети, чаще всего используется метод, который заключается в определении эмоциональной окраски текста, размещаемого пользователем. [1]. Данный метод получил название – анализ тональности текста (от анг. «Sentiment analysis»), в котором текст можно рассматривать с точки зрения эмоций, которые выражает пользователь. Анализировать данные можно как вручную, так и при помощи компьютерных средств. В данной работе рассматривается анализ информационного пространства посредством использования автоматизированного способа, при помощи компьютерных средств, т.к. данный способ обладает высокой скоростью чтения и анализа данных, а также у него отсутствует понятие субъективности при измерении эмоций пользователя социальной сети.

Повышенный интерес к тональному анализу информационных данных пользователей социальных сетей начался в 2000-х году и продолжается до сих пор. Анализ тональности текста был изучен в трудах не только отечественных ученых, но и зарубежных исследователей, таких как Д. Усталов (2012г.), И.Четверкин (2012г.), А. Пазельская (2011г.), И. Меньшиков (2012г.), Е. Котельников (2012г.), Н. Лукачевич (2012г.), П. Шарма (2020г.), Н. Пономарева (2012г.) и др.

Под анализом тональности текста обычно понимается задача анализа эмоционально окрашенной лексики, которую использует пользователь в своем тексте. Важно знать, что для анализа тональности информационных данных пользователя используются три типа эмоций, такие как: положительные, негативные и нейтральные.

Методы. Для решения задач, связанных с анализом тональности текста в информационном пространстве пользователя социальной сети, наиболее часто используются метод обучения без учителя и статистический метод. Преимуществом

использования данных методов является использование алгоритмов машинного обучения для анализа и кластеризации немаркированных наборов данных.

Метод обучения без учителя. Как говорилось ранее, преимуществом использования данного метода является то, что метод не требует заранее подготовленных данных для анализа тональности текста в информационном пространстве пользователя. В этом случае нейронная сеть, которая работает при помощи метода обучения без учителя, анализирует данные пользователя, и сама находит факты, которыми они связаны между собой. Следует отметить, что метод обучения без учителя менее точен, чем метод обучения с учителем [2].

Приведем пример алгоритма метода обучения нейронной сети без учителя. Данный метод применим к задаче анализа взаимосвязи данных пользователя и извлечения фактов.

Bootstrapping – метод распространения. В основе данного метода лежит идея многократной генерации выборок на базе уже имеющейся, т.е. при помощи небольшой базы фактов примеров, которые были определены вручную. Далее, только уже автоматически, извлекать похожую текстовую информацию, со временем наращивая множество фактов. Благодаря данному методу можно легко оценивать самые различные факты для сложных моделей.

Самостоятельные части речи могут выступать в качестве фактов.

Таким образом, рассмотрим алгоритм данного метода, который включает многократное отрисовку выборочных данных с заменой из источника данных для оценки параметра генеральной совокупности.

1. Задаем множество $\bigcup_0 = \{s_1, ..., s_k\}, i = 0$, где $\bigcup_0$ – исходное множество терминов;
2. Находим в документе n-граммы $n_1 ... n_m$, близкие терминам из $\bigcup$, где n-граммы – слова или части слов, которые состоят из n - символов;
3. $\bigcup_{i+1} = \bigcup_i + \{n_1, ..., n_m\}$, $i = i + 1$, где $\bigcup_i$ – новое множество терминов, с включенными проанализированными n - граммами;
4. Переходим к пункту 2, в случае, если $i < n$.

Схожесть n - граммы и фактов из множества $\bigcup_i$ определена с помощью $R\log F$ метрики [3]:

$R\log F(ng) = \log freq(ng, U_i) * (freq(ng, U_i) / freg(ng))$, где $freq(ng)$ $freq(ng)$ – частота встречаемости терминов в тексте, который состоит из заданного количества слов.

Статистические методы. При обучении модели статистическим методом, следует учитывать то, что должны быть заготовлены заранее полученные по тональности и эмотивности корпуса текстов. С помощью заранее полученных данных происходит определение тональности текста информационного пространства.

В данном методе используются n-граммы, к которые в свою очередь являются альтернативой традиционному морфологическому разбору и удалению стоп-слов. В тексте выделяются те n – граммы, частота использования которых в корпусе текста, наиболее заметна. Обычно предполагается что в состав n- граммы входит не менее 2 слов. Далее выделенные n – граммы проходят проверку на компактность. N-грамма попадает в список фактов в том случае, если она компактна как минимум в двух предложениях.

Наиболее часто встречающиеся при анализе одного факта и описывающие объекты существительные или словосочетания, в составе которых содержится имя существительное называются терминами [4].

Компактность определяется следующим образом:

1. Предположим, что s – это предложение, а f – это n-грамма из n – слов предложения;

2. N – грамма из n – слов предложения (f) компактна в конкретном предложении в том случае, если расстояние между двумя словами, смежными в f, в s составляет не более трех слов.

Термины, состоящие из одного слова, также проходят статистический тест на чистоту. Отыскиваются все предложения, содержащие термин. Среди найденных предложений подсчитываются предложения, прошедшие тест на компактность n-граммы, в которые входит этот термин. Термин сможет попасть в список фактов в том случае, если число предложений окажется больше, чем те, что имеются условно.

Далее вычисляется значимость для всех найденных n-грамм по следующей формуле [5]:

$$\text{C-value}=\begin{cases}\log(len(tern))* freq(term), если\ |e_{terms}|=0;\\ \log_2(len(term))* freq(term)-1(|e_{terms}|)\sum_{вterms} freq(elder).\end{cases}$$

где: $term$ – соответствующая n-грамма;

e_{terms} - множество всех n-грамм большего порядка, которое содержит в себе $term$;

$|e_{terms}|$ - интенсивность/мощность;

$elder$ – обозначение элемента, входящего в множество;

len - размер исследуемого множества.

Сбор данных и их маркировка. Сбор данных осуществляется из социальной сети «Вконтакте». Социальная сеть снабжена методами для взаимодействия и извлечения информации при помощи VK_API. После извлечения всех данных, информация маркируется.

Чтобы обратиться к методу API ВКонтакте, необходимо выполнить POST или GET запрос следующего вида: requests.get ('https://api.vk.com/method/wall.get'). Данный запрос состоит из нескольких частей, таких как: params – входные параметры, в состав которых входят token – ключ доступа, v – используемая версия API. Ниже представлена часть кода обращения к методу API:

```
token = 'b6e60a65b6e60a65b6e60a65ffb69e88f8bb6e6b6e60a65d603965d2127f9cdc8a7beb8'
version = 5.131
domain = '___'
count = 100
offset = 0
```

Код для парсинга данных из социальной сети написан на Python. Python – язык программирования, считается высокоуровневым, который поддерживает динамическую строгую типизацию, т.е. переменная начинает работать с типом в момент ее присваивания, что обозначает, что одна и та же переменная может принимать различные типы данных. Еще одним преимуществом использования Python является свойство автоматического управления памятью. Язык программирования Python имеет библиотеку NLTK, которая позволяет эффективно работать с различными лингвистическими данными и анализировать их, с помощью текстовых классификаторов.

На основании полученных в статье результатов сделаем следующие выводы.

1. Рассмотрены два популярных метода для определения тональности текстовых данных пользователя социальной сети.

2. Отмечено, что у каждого метода имеются свои особенности, которые нужно учитывать при использовании в решении практических задач.

1. Pang B. & Lee L. Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval, v.2 n.1-2, January, 2008 - pp.1-135.

2. Yarowsky, D. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods // Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, MA, pp. 189-196, 1995.

3. Jingbo Zhu, Huizhen Wang, Muhua Zhu, Benjamin K Tsou, Matthew Ma. Aspect-Based Opinion Polling from Customer Reviews by // IEEE Transactions on affective computing, vol. 2, no. 1, january-march 201

4. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. // Proceedin// Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04) ACM, New York, NY, USA, 168-177.

5. Frantzi, K., Ananiadou, S. and Mima, H. Automatic recognition of multi-word terms // International Journal of Digital Libraries 3(2), pp.117-132.,2000.