Содержание

ВВЕДЕНИЕ

В настоящее время социальные сети и форумы становятся частью повседневной жизни подавляющего количества пользователей сети Интернет. Активный обмен мнениями в сети привел к увеличению интереса к автоматическому извлечению тональности текста как со стороны научного сообщества, так и со стороны многих коммерческих организаций. Автоматизация определения тональности текста — это алгоритмически сложная задача, включающая некоторые не менее сложные подзадачи, о которых пойдет речь в представленной работе.

Например, крупнейшие интернет – магазины, получают многочисленное количество отзывов на свои товары. Выделение общих закономерностей в большом количестве неструктурированных текстов — это рутинная работа, требующая автоматизации. Таким образом, подбор эффективного алгоритма выделения является актуальной задачей экономической и производственной аналитики.

Однако, существует множество подходов для автоматического определения тональности текста, каждый из которых имеет свои особенности. В работе предлагается классификация этих методов и их краткий обзор.

Анализ тональности текста часто рассматривается как часть научной области, называемой *Text Mining* — интеллектуальный анализ неструктурированных данных. Методы анализа неструктурированных данных лежат на стыке нескольких областей: обработка естественных языков, интеллектуальный анализ данных, поиск информации, извлечение информации и управление знаниями.[1]

Направление интеллектуального анализа текста, рассмотренное в данной работе — задача анализа тональности текстов (Sentiment Analysis). Задача близка к классической задаче контент – анализа текстов массовой коммуникации, в которой оценивается общая тональность высказываний в целом.

Целью проводимой работы является классификация существующих методов анализа тональности текста, выделение прикладных задач, которые наиболее подходят для каждого класса.

1. Анализ предметной области

В данном разделе представлена информация об актуальности исследуемой задачи и определены основные понятия, используемые при дальнейшем анализе.

1.1 Основные понятия

Очевидно, что анализ тональности текста тесно связан с понятием естественного языка. Определение естественного языка с точки зрения компьютерной лингвистики дано в работе [2] следующим образом:
Естественный язык — большая открытая многоуровневая система знаков, возникшая для обмена информацией в процессе практической деятельности человека, и постоянно изменяющаяся в связи с этой деятельностью.

Тональностью текста принято называть эмоциональную оценку, выраженную в неструктурированном тексте по отношению к сущности, и определяемую тональностью составляющих ее лексических единиц и правил их сочетания.

Автоматический анализ текста осуществляется на основе двух подходов: инженерно-лингвистический подход и подход на основе векторного анализа.

Инженерно лингвистический подход (подход с использованием паттернов и шаблонов) заключается в генерации правил, на основе которых будет определяться тональность текста. Для реализации таких методов создаются словари оценочной лексики и разрабатываются алгоритмы применения лингвистических правил для учета контекста употребляемых отдельно слов и выражений.

Подходы основанные на векторном анализе принято делить на классы, представленные ниже.

- Машинное обучение с учителем. Отбирается некоторое количество текстов для обучения и на их основе обучается классификатор.

- Машинное обучение без учителя. Подход основан на идее, что наибольший вес в тексте имеют самые часто встречаемые термины.[3] Выделив данные термины, можно сделать вывод о тональности текста в целом.

Важным термином для работы с текстами на естественном языке является корпус текстов. Под корпусом текстов понимают проработанную под определенными правилами совокупность текстов, которая используется как база для исследования. Далее будет указано, что классификаторы, применяемые в методах на основе векторного анализа, в основнм работают с корпусами текстов.

1.2 Некоторые сложности при анализе тональности

Тональность может выражаться с помощью эксплицитных и имплицитных оценок. Использование оценочной лексики — это имплицитный способ выражения эмоциональной оценки. Например, высказывание *«фен хороший»* — это имплицитная оценка, поскольку высказывание содержит оценочное слово. Однако, высказывание *«расческа широкая и не для густых волос»* не содержит оценочных слов, но в то же время дает эмоциональную оценку товару путем указания реального факта, приводящего к оценке.

Таким образом, имплицитное мнение (оценка)[4] — это объективное высказывание, из которого следует оценка, т. е. имплицитное мнение сообщает желательный или нежелательный факт.

Важной проблемой является определение сферы действия модификатора полярности в конкретном предложении. Например, *«Мне не нравится состав актеров в этом фильме, но некоторые актеры были очень запоминающимися.»* Частица «не» модифицирует только слово «нравится», но не модифицирует слово «запоминающимися».

1.3 Вывод

Приведены некоторые определения, которые будут использованы в дальнейшем при обзоре методов анализа тональности текста. Сформулирована актуальность задачи.

2. Определение тональности на основе правил и шаблонов

В разделе описано применение инженерно — лингвистического подхода, приведены примеры словарей оценочной лексики и примеры математических моделей, реализующих данный подход.

2.1 Словари оценочной лексики

При применении метода, основывающееся на правилах и шаблонах, сначала необходимо составить словарь оценочных слов и выражений. В таких словарях обычно каждому слову приписывается оценка тональности. Словари могут быть сформированы следующим образом:

- сопоставление существующего словаря с корпусом анализируемых текстов, выборка предметной лексики;
- автоматизированный анализ корпуса текста и отбор оценочной лексики по заранее известному алгоритму;
- перевод иноязычных существующих словарей оценочной лексики на необходимый язык;
- просмотр корпуса текстов и пополнение словаря вручную.

В работе [5] предложен подход автоматического порождения тонального словаря ProductSentiRus. Извлечение оценочных слов в заданной предметной области основано на нескольких текстовых корпусах: корпус отзывов о продуктах с оценками пользователей, коллекции описаний продуктов и контрастного корпуса (например, новостного). Такие корпусы могут быть автоматически сформированы для разных предметных областей.

Кроме того, в работе предполагается, что некоторые части корпуса мнений можно выделить: например, корпусы, в которых выше концентрация

оценочных слов, большее разнообразие знаков пунктуации (троеточия и восклицательные знаки), большая концентрация коротких предложений (менее семи слов), предложения, содержащие аспект, к которому применена оценочная лексика, без других существительных.

Словарь представлен как список 5 тысяч слов, упорядоченных по мере снижения вычисленной вероятности их оценочности без указания позитивной или негативной тональности. Точность оценочных слов в первой тысяче слова списка составляет более 91%.

Ниже представлена таблица наиболее вероятных оценочных слов по версии словаря ProductSentiRus.

Таблица 2.1 – Примеры слов из словаря ProductSentiRus

бесподобный	0.963
невнятный	0.953
отличнейший	0.935
обалденный	0.933
безумно	0.924
непонятно	0.921
неприятно	0.920
отвратный	0.920
нежный	0.916

Также можно привести в пример словарь РуСентиЛекс. Он представляет собой упорядоченный по алфавиту список слов и выражений, содержащий следующие типы русскоязычных слов, значения которых связаны с тональностью:

- слова литературного русского языка, явно выражающие отношение к аспекту (например, *хороший*);
- слова литературного русского языка, подразумевающие в контексте наличие имплицитной оценки, то есть не передающие сами по себе отношение автора, но имеющие положительную или отрицательную оценку, например, *болезнь, спам, драка*;

- сленговые слова русскоязычных пользователей Твиттера.

Словарь РуСентиЛекс хранится в простом текстовом формате, подобном формату словаря MPQA [6]. Каждой единице словаря приписываются следующие атрибуты:

- часть речи;
- слово или фраза, в которой каждое слово стоит в лемматизированной форме(для существительных — именительный падеж, единственное число, для прилагательных — именительный падеж, единственное число, мужской род, для глаголов, причастий, деепричастий — глагол в инфинитиве (неопределённой форме) несовершенного вида);
- тональность (негативная, позитивная, нейтральная, двойная). Последнее означает, что оценка зависит от контекста;
- источник тональности (эксплицитная или имплицитная оценка).

Словаря Linis-Crowd [7] создавался для анализа тональности текстов социальных сетей. Солварь включает:

- наиболее частотные прилагательные русского языка, употребляемые в текстах социальных сетей;
- образованные от отобранных прилагательных наречия,;
- словарь ProductSentiRus, из которого были выбраны слова, подходящие для анализа сообщений в социальных сетях и др.

Оценка выражается целым числом по шкале от -2 (сильно негативный) до $+2$ (сильно позитивный). Оценки различных разметчиков усреднялись.

2.2 Построение математической модели

Для правильности построения оценки предложения используется ряд правил, которые модифицируют оценку в зависимости от контекста. Список наиболее часто используемых правил приведен ниже.

1. Суммирование оценок слов, входящих в состав одного высказывания относящегося к определенной сущности.

2. Применение правил отрицания.

3. Выставление негативной оценки для словосочетания, в котором употреблено хотя бы одно негативное выражение.[8]

Применения правил отрицания требует более детального рассмотрения. Так, например, в работе [9] представлены правила, использующиеся при обнаружении слов «не», «никогда», и так далее. Правила можно сформулировать следующим образом:

- отрицание отрицательной оценки → положительная оценка;
- отрицание положительной оценки → отрицательная оценка;
- отрицание слова без эмоциональной окраски → отрицательная оценка (например, «не работает»).

С учетом применения правил отрицания определение тональности слова в данной работе имеет следующий вид (листинг 1):

Алгоритм 1 Определение тональности слова

```
Процедура оценкаСлова(слово, предложение, оценка)
    Если Слово выражает отрицание тогда
        применить правила отрицания
        отметить слова в предложение, к которым применено отрицание
    иначе
        оценить слово согласно оценка
    Конец условия
Конец процедуры
```

Учитывая данное правило, проставление оценок происходит следующим образом. Пусть $ast_1 \dots sw_i$ — оценочные термины, для которых уже проведена процедура оценки, a_i — оценка из словаря оценочной лексики. Тогда оценка для каждого аспектного термина вычисляется следующим образом:

$$score(a_i, S) = \sum_{sw_j \in S} \frac{num\,(sw_j)}{dist\,(sw_j, a_i)}, \tag{2.1}$$

где $num\,(sw_j)$ — числовая оценка тональности, $dist\,(sw_j, a_i)$ — расстояние между оценочным словом и аспектом. На оценку каждого оценочного термина влияют все слова в предложении, однако мера в которой они влияют на итоговую оценку определяется расстоянием от слова до оценочного термина.

Работа [10] учитывает фактор нереального контекста — в этой работе описываются правила, в результате применения которых тональность слов, являющихся индикаторами нереального контекста, не принимаются во внимание. В работе определены следующие маркеры нереального контекста:

- индикаторы условного наклонения (например, if);
- модальные глаголы;
- вопросы и слова, заключенные в кавычки.

Работа [11] предлагает более детальный подход — используется шесть правил для составления оценки:

- перевод в противоположную тональность и применение отрицаний;
- приписывание доминирующей тональности идентификатора для синтаксических групп — например, POS(*завораживающий*) + NEG(*хаос*) = POS(*завораживающий хаос*);
- распространение модификатора на слово, стоящее рядом, если используется глагол распространения, такой как *ненавидеть, обожать* или *восхищаться*;
- доминирование полярности глагола над объектом, к которому применяется глагол;
- нейтрализация оценочного выражения предлогом — модификатором, таким как, например, *не смотря на*;
- усиление или ослабление веса тональности при обнаружении таких слов как *очень*.

Таким образом, инженерно — лингвистический подход имеет следующие достоинства:

- результат работы зависит от содержимого словаря оценочной лексики и используемых правил — то есть, результат предсказуем;
- возможен более глубокий анализ тональности на уровне высказывания, если подобрать словарь, оценка слов в котором будет шире шабл.

Однако, можно выделить и следующие недостатки:

- составление словаря оценочной лексики и правил оценки вручную — трудоемкая и дорогостоящая операция;
- при узком диапазоне слов в словаре оценочной лексики метод дает неточные результаты.

2.3 Вывод

Инженерно лингвистический подход показывает высокую точность работы. Однако, результат напрямую зависит от организации и содержания словаря. Данный метод удобно применять в тех ситуациях, когда известна предметная область исследования и сущности, с которыми будет проведена работа.

3. Определение тональности на основе векторного анализа

В данном разделе указан общий алгоритм определения тональности на основе векторного анализа и дан краткий обзор некоторым методам и математическим моделям.

3.1 Общее описание подхода

Как было указано ранее, для реализации подхода, основанного на векторном анализе требуется отобрать некоторое количество текстов на основе которых обучается классификатор. Текстам вручную расставляется отметка тональности, после этого выбирается алгоритм классификации. Алгоритм создания системы анализа тональности текста на основе векторного анализа приведен ниже.

1. Выбор алгоритма предобработки текста.
2. Выбор алгоритма классификации.
3. Выбор признаков представления текста.
4. Выбор способа подсчета веса признака.

Некоторые этапы алгоритма требуют более детального рассмотрения.

3.2 Предобработка текста

Перед обучением классификатора следует осуществить предобработку текстовой информации. Например, в работе [12] использовалась леммизация — все слова преобразовывались к словарной форме (лемме) при помощи морфологического анализатора mystem[13] от компании Яндекс. Из

текстов исключались англоязычные и русскоязычные «стоп-слова» (частицы, предлоги, местоимения), удалялись слова длиной менее трех символов.

3.3 Некоторые алгоритмы классификации

Одними из самых широко используемых алгоритмов классификации являются:

- SVM (*Support Vector Machines*) — алгоритм, создающий линейную гиперплоскость, которая разделяет данные на классы;
- наивный байесовский классификатор — вероятностный классификатор, основанный на применении теоремы Байеса;
- метод на основе ключевых слов — применение лексико — статического анализа и составление списка ключевых слов для каждого класса.

3.4 Признаки представления текста

Существует множество подходов представления текста.

Метод «мешок слов» представляет собой представление текста в виде вектора слов, встречающихся в тексте. В модели [14] нейронная сеть строит векторное пространство слов, над которыми производятся векторные операции, включая определение мер близости или семантически значимые операции вычитания и сложения (Например, *«принц» - мужчина + женщина = «принцесса»*).

В работе [15] описан иной подход на основе кластеризации bag-of-words векторов лексических подстановок, сгенерированных нейросетевыми языковыми моделями. Идея метода состоит в следующем: интересующее слово в целевом корпусе заменяется токеном или последовательностью токенов, и в таком виде подается предсказателю, который предсказывает возможные замены для токена в этом контексте.

Часто при выборе представления текста учитываются части речи и пунктуация(например, количество восклицательных или вопросительных знаков).

3.5 Способ подсчета веса признака

Как и в случае оценочных словарей, существует множество подходов к подсчету веса признака. Некоторые модели ограничиваются «булевой» формой (вес есть — веса нет), а некоторые вводят шкалу взвешивания и вводят частотность.

Наиболее часто используемый подход взвешивания — *tf.idf*[16]. Tf.idf (term frequency-inverse document frequency) является популярной метрикой при решении задач информационного поиска и анализа текста. Tf.idf представляет собой статистическую меру того, насколько термин важен в документе, который является частью корпуса. С использованием Tf.idf важность термина пропорциональна количеству встречаемости термина в документе и обратно пропорциональна количеству встречаемости термина во всей коллекции документов. У данной модели есть улучшения — в работе [17] представлен подход delta TFIDF. Основное отличие в том, что учитывается, насколько неравномерно распределено слово в двух классах тональности (положительный или отрицательный). Формула для определения веса имеет следующий вид:

$$weight_i = tf_i \cdot \log_2 \left(\frac{N_1 \cdot df_{i,2}}{N_2 \cdot df_{i,1}} \right), \tag{3.1}$$

где N_1 — количество документов в положительном классе, N_2 — в отрицательном, $df_{i,1}$ — количество документов положительного класса, в которых встречалось слово, $df_{i,2}$ — соответственно, в отрицательном, tf — частота вхождения, которая вычисляется согласно следующей формуле:

$$tf(t,d) = \frac{n_t}{\sum_k n_k}, \tag{3.2}$$

n_t — число вхождений исследуемого слова в данный документ, а в знаменателе — общее число слов.

Главное преимущество методов, использующих машинное обучение — это высокие показатели эффективности. Как правило, если при обучении классификатора используется достаточно обширная размеченная подборка тек-

стов, то метод показывает большую эффективность, чем шаблонный метод. Однако, метод имеет ряд недостатков.

Во — первых, создание обучающей коллекции — это трудоемкий процесс. Во — вторых, алгоритмы плохо переносятся на другую предметную область, поскольку это требует перенастройки системы [18].

3.6 Вывод

Методы, основанные на векторном анализе, показывают высокую эффективность работы. Однако, построенная система «привязана» к конкретной предметной области. В то же время, для работы данного метода необязательно заранее знать сущности, с которыми будет проведена работа. Соответственно, метод применим для текстов любой тематики и направленности, но ограниченных одной предметной областью.

ЗАКЛЮЧЕНИЕ

На основе исследуемой литературы был проведен анализ методов определения тональности текста на естественном языке. Были выявлены две категории методов: методы на основе шаблонов и правил и методы на основе векторного анализа.

В ходе работы были сделаны некоторые выводы, представленные ниже.

1. Инженерно — лингвистические методы требуют наличия словаря оценочной лексики. Чем обширнее словарь, тем точнее будет определена тональность текста.

2. Методы векторного анализа требуют предварительно подготовленной обучающей коллекции для классификатора. Если подборка достаточно большая, то метод покажет наибольшую эффективность.

3. В отличие от методов на основе векторного анализа, в инженерно — лингвистических методах требуется, чтобы заранее были известны сущности, с которыми работает алгоритм.

4. При использовании методов на основе векторного анализа трудно сменить предметную область для заранее построенной системы.

Инженерно — лингвистические методы предлагается использовать для тех случаев, когда заранее определены сущности оценивания: например, отзывы на определенную вещь в интернет — магазинах.

Методы на основе векторного анализа предлагается использовать для текстов, в которых чаще всего оцениваются сущности одной предметной области, но в разных ее аспектах. К таким текстам относятся личные блоги и новостные тексты.

Список литературы

[1] Барсегян А.А. и др. *Технологии анализа данных: Data Mining, Visual Mining, TextMining, OLAP*. СПб.: БХВ-Петербург, 2008.

[2] Касевич В.Б. *Элементы общей лингвистики.* издательство «НАУКА», 1977.

[3] Turney P. "humbs up or thumbs down? Semanticorientation applied to unsupervised classification ofreviews". в: *Proceedings of ACL-02* (2002).

[4] Zhang L. Liu B. "A survey of opinion mining and sentiment analysis". в: *Mining Text Data. Springer* (2012).

[5] Natalia Loukachevitch Ilia Chetviorkin. "Extraction of Russian Sentiment Lexiconfor Product Meta-Domain". в: *Proceedings of COLING-2012* (2012).

[6] Janyce Wiebe Lingjia Deng. "MPQA 3.0: An Entity/Event-Level Sentiment Corpus". в: *HLT-NAACL* (2015).

[7] Кольцова О. Ю. Алексеева С. В. Кольцов С. Н. "Linis-crowd. org: лексический ресурс для анализа тональности социально-политическихтекстов на русском языке". в: *Компьютерная лингвистика и вычисли-тельные онтологии: сборник научных статей.* (2015).

[8] Киселев С.Л. Ермаков А.Е. "Лингвистическая модель для компьютерного анализа тональности публикаций СМИ". в: *Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог.* (2005).

[9] Bing Liu Philip S. Yu. "A holistic lexicon-based approach to opinion mining". в: *Proceedings of the International Conference on Web Search and Web Data Mining* (2008).

[10] Tofiloski Taboada M. Brooke J. "Lexicon-basedmethods for sentiment analysis". в: *Computational linguistics* (2011).

[11] Mitsuru Ishizuka Alena Neviarouskaya Helmut Prendinger. "Recognition of Affect, Judgment, and Appreciation in Text". в: *Proceedings of the 23rd International Conference on Computational Linguistics* (2010).

[12] Клековкина М. В. Котельников Е. В. “Автоматический анализ тональности текстов на основе методов машинного обучения”. в: *Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог.* (2011).

[13] *MyStem*. URL: `https://yandex.ru/dev/mystem/`. (Дата обр.: 19.01.2022).

[14] В. Мызников П. “Моделирование рассуждений на основе прецедентов в автоматическом анализе новостных текстов”. в: *Вестник Новосибирского государственного университета. Серия: Информационные технологии* (2017).

[15] Yoav Goldberg Asaf Amrami. “Towards better substitution-based word sense induction”. в: *ArXiv* (2019).

[16] Schutze H. Manning C. D. Raghavan P. “Introduction to information retrieval”. в: *Cambridge: Cambridge University Press* (2008).

[17] Tim Finin ustin Martineau. “Delta TFIDF: An Improved Feature Space for Sentiment Analysis”. в: *In Proceedings of the Third AAAI Internatonal Conference on Weblogs and Social Media* (2009).

[18] X. Pan S.J. Ni. “Cross-domain sentiment classification via spectral feature alignment”. в: *Proceedings of the 19thinternational conference on World wide web.* (2010).