Автоматическое определение тональности отзывов о продукте

Н. Р. Икрамов[1], Р. Е. Спиридонов[2]
СПбГЭТУ «ЛЭТИ»
[1]ikramovv.nikita@yandex.ru, [2]Rom@nSpiridonov.ru

Аннотация. **Из-за широкого развития сети Интернет все большее количество людей совершают покупки в онлайн магазинах и оставляют там свои отзывы о товарах. Огромный поток таких сообщений в данный момент приходится обрабатывать вручную. Для эффективной работы компаний требуется предоставлять тщательный анализ отзывов, чтобы исключать ненастоящие отзывы и рекомендовать интересные пользователю товары. Автоматизация этой работы возможна с помощью применения методов автоматического анализа тональности текста. В данной статье мы рассматриваем подходящие для решения обозначенных задач методы обработки данных.**

Ключевые слова: тональность текста; стемминг; тональные словари

I. Введение

После появления онлайн платформ для выражения своего мнения о товарах(сайты по продаже товаров, специальные сайты с отзывами) информации стало слишком много. Теперь недостаточно просто иметь огромную базу отзывов. Сейчас для максимального удобства пользователя необходимо исследовать динамику мнений о товаре и выделять ключевые недостатки/плюсы продуктов на платформе. В данной статье описывается возможный подход к решению задачи по анализу данных, извлеченных из отзывов о товарах в Amazon с целью анализа эмоциональной окраски сообщений.

II. Постановка задачи

Конечная цель анализа тональности текста — выявлять эмоциональный окрас сообщения о товаре. Чтобы выявить тренды развития конкретных товаров и их недостатки необходимо собрать информацию о существующих продуктах на рынке и отношения покупателей к ним. Именно эту задача все сложнее делать вручную из-за увеличения объема входящих данных.

Сложность задачи анализа тональности текста в том, что приходится работать с разговорным естественным языком. Поэтому в тексте могут встречаться ошибки, опечатки и сленговые выражения. Также правильный анализ эмоциональной окраски осложняет наличие в тексте иронии. Сарказм изменяет эмоциональное значение текста на противоположное. В последнее время появляются ненастоящие отзывы с целью поднятия рейтинга товара. Со всеми этими проблемами не всегда люди полностью справляются, поэтому автоматизировать этот процесс непросто.

III. Подходы к анализу текста с использованием правил и словарей

На данный момент подходы к определению тональности текста можно разделить на две группы: основанные на правилах и методы с использованием алгоритмов машинного обучения.

A. Метод, основанный на правилах

Метод, основанный на правилах. Метод заключается в том, что существует определенный набор правил для предметной области. Текст разделяется на слова и последовательности n-грамм. В последствии эти данные используются для поиска определенных шаблонов и присвоении им определенной тональности(положительной или отрицательной). Далее шаблоны, которые были выделены ранее, используются с целью создания правил, которые выглядят следующим образом: если «Условие», то «заключение». Если перед последовательностью слов используется отрицание, то тональность всего высказывания меняется на противоположное. В таком случае, чтобы получить итоговую оценку тональности следует выполнить подсчёт суммы всех весов слов в предложении. В этот момент суммарная тональность всех частей присутствующих в тексте может отличаться от общей эмоциональной окраски целого текста

B. Метод, базирующийся на использовании словарей оценочной лексики

Для того чтобы реализовать данный метод необходимо создать специальные словари, которые содержат слова, или их комбинации и соответствующие им веса в зависимости от тональности. Как способ разработки таких словарей – разработка правил, которые используются для извлечения новых оценочных слов их текстов. Например, если какие-то два прилагательных, которые соединены союзом «и» и при этом, одно из них входит в состав словаря, то второе также следует включить и отметить идентичным весом. В результате такой список слов будет способен пополнить словарь.

Например, в предложении «This notebook is powerful and light» («Этот ноутбук мощный и лёгкий») прилагательное «powerful» имеет положительную

тонально, следовательно, прилагательное «light» также будет иметь положительную тональность, потому что пользователь высказывает своё мнение относительно одного предмета, которому даёт оценку.

Данный алгоритм требует иметь большой объём отзывов.

Плюс этого алгоритма в его простоте.

Главный недостаток – это то, что он не является универсальным, так как для каждой новой предметной области существует необходимость составления своего словаря оценочных слов.

IV. Подходы к анализу текста с использованием машинного обучения

В области машинного обучения есть два типа: обучение с контролем учителя и обучение без контроля учителя. Последний тип имеет важное преимущество – он уменьшает необходимость в предварительной подготовке данных предназначенных для обучения. Тем не менее, самым распространённым методом является – это машинное обучение с учителем.

A. Машинное обучение с учителем

Для данного подхода существует потребность в разметке набора текстов. Каждый размеченный фрагмент текста представляет собой пару векторов признаков текста, которая является словом или конструкцией из слов, которой присвоен некоторый вес, и определённой тональности. Далее строится классификатор, который будет использован для определения тональности нового корпуса. Достоинством данного, описанного выше, метода является высокая точность, с которой определяется тональность, а также то, что при помощи обучающей выборки классификатор способен сам выделить те признаки, которые влияют на тональность.

При составлении тональных словарей оценочных слов и при этом дальнейшее использование на других корпусах текстов той же области устраняет частично зависимость от предметной области.

Но у данного метода, конечно, есть недостатки:

- Необходима размеченная обучающая выборка.
- Результаты будут сильно зависимы от выбранного алгоритма, а также его обучающей выборки или параметров обработки.

B. Машинное обучение без учителя

Обучение без учителя представляет собой ещё один раздел машинного обучения. Особенность его состоит в том, что для тренировки алгоритма необходимо использовать обучающую выборку, которая состоит из документов, классы которых заранее неизвестны.

Достоинство этого метода перед обучением с учителем в том, что для обучения не требуется размеченная выборка.

Минус метода в том, что он значительно проигрывает методу обучения с учителем по точности определения эмоциональной тональности текста.

V. Эксперимент по анализу тональности текста на английском языке

Для проведения экспериментов был выбран корпус данных, полученный с веб-ресурса Amazon, в нем насчитывается – 2700 отзывов.

Отзывы представляются следующим образом:

- URL-адрес отзыва,
- заголовок,
- оценка товара в пятибалльной шкале (звездность),
- флаг приобретения,
- количество пользователей, которым помог отзыв,
- дата
- текст отзыва

Алгоритм, который описывает решение поставленной задачи:

- Предварительная обработка данных.
- Разметка данных.
- Отбор слов в словарь на основе разметки. Заполнение тональных словарей. Определение веса слова.
- Выделение конструкций, определяющих свойства продукта.
- Оценка работы алгоритма.

Для проведения экспериментов был выбран язык программирования Python. Из дополнительных пакетов в ходе исследования были применены nltk, numpy, sklearn, pandas. Исследования проводились в интерактивной оболочке Jupyter Notebook.

VI. Составление тональных словарей

Словари составляются на основе, проведенной на этапе предобработки, разметки. Т.е. если в графе «sentimental» значение «1», то слово записывается в словарь позитивных слов, иначе – отрицательных.

Для того чтобы заполнить тональные словари изначально необходимо определить количество уникальных слов, которые входят в отзывы. Это и будет длинной словаря.

Всего уникальных слов в корпусе – 6595.

Слова, которые встречаются более одного раза, оцениваются исходя из разметки отзыва, и имеют счетчики определенной тональности. В этих счетчиках отображается информация о количестве вхождения слова в отзывы с отрицательной и положительной оценкой.

Так как слово в словаре может быть и полярным, т.е. только одной тональности, и многозначным, которое содержится и в отрицательном, и положительном отзыве, необходимо этому слову придать некоторый вес.

Вес слова будет рассчитываться по данной формуле:

$$W=(K\_pos-K\_neg)/(K\_pos+K\_neg)$$

W – вес слова;

K_pos – встречаемость в положительном отзыве;

K_neg – встречаемость в отрицательном отзыве.

При расчете веса никогда не будет ситуации, в которой сумма K_pos+K_neg=0, которая ведет к делению на 0. Это объясняется тем, что в словаре присутствуют все слова, которые даже единственный раз упоминаются в тренировочной выборке, случайно попавших слов нет.

Слова, которые имеют вес близкий к «1» – будем считать положительными, к «-1» – отрицательными. Те слова, которые имеют вес близкий к «0» являются нейтральными, это объясняется тем, что количество вхождений слова в отрицательный и положительный отзыв практически равно.

В тот момент, когда модель будет предсказывать тональность отзыва, она сложит все веса. В случае если слово отсутствует в словаре, модель просто пропускает его.

Таким образом, если сумма всех слов, которые входят в отзыв больше или равно 0, то будем считать, что отзыв имеет положительную тональность, в противном случае – негативную.

VII. Оценка качества определения тональности

Для того чтобы оценить качество классификации проведем следующие эксперименты:

- классификация без предобработки данных;
- очистка от стоп-слов;
- токенизация, лемматизация, стемминг;
- токенизация, лемматизация, стемминг+очистка от стоп-слов;
- токенизация, лемматизация, стемминг+очистка от стоп-слов+ составление словаря на основе всех отзывов.

Результаты экспериментов отображены в таблице:

ТАБЛИЦА I Результаты классификации

Метод	*Точность*	*Полнота*	*F-мера*	*Правильность*
Без предобработки данных	0.71	0.60	0.51	0.600
Очистка от стоп-слов	0.74	0.71	0.67	0.707
Токенизация, лемматизация, стемминг	0.75	0.67	0.63	0.674
Токенизация, лемматизация, стемминг+очистка от стоп-слов	0.81	0.79	0.79	0.790

Результаты классификации в зависимости от предобработки данных

Лучший результат показал эксперимент, в котором были применены токенизация, лемматизация, стемминг, очистка от стоп-слов и показывает точность 79%. Этот результат превосходит очистку от стоп-слов, которая показывает результат хуже на 8.3%. Токенизация, лемматизация, стемминг также не показывают хороших результатов без предварительной очистки и хуже на 11,6%. А вот применение необработанного текста для классификации, дает результаты гораздо хуже, чем все из предложенных методов на 19%, и показывает точность 60%, что является плохим показателем классификации.

VIII. Заключение

Метод определения тональности с использованием тональных словарей позволяет осуществить более детальную настройку на конкретную предметную область, а также выявлять достоинства и недостатки товаров и услуг, описываемых в отзывах. Однако, для использования данного метода на другой предметной области необходимо создание новых тональных словарей, что является сложной и трудоёмкой работой. В случаях, когда необходимо производить анализ конкретной предметной области, использование данного метода имеет место быть, однако, если необходима гибкость и возможность анализировать другие предметные области, стоит отдать предпочтение методам, с использованием алгоритмов машинного обучения.

Список литературы

[1] Клековкина М. В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики/ М. В. Клековкина, Е.В. Котельников // RCDL-2012, Переславль-Залесский, Россия: конференция. 2012

[2] Автоматическая обработка текстов на естественном языке и анализ данных / Е. И. Большакова, К. В. Воронцов, Н. Э. Ефремова и др. – Изд-во НИУ ВШЭ Москва, 2017. 269 с.

[3] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient Estimation of Word Representations in Vector Space,” In Proc. of Workshop at ICLR, 2013.

[4] Ричард Риз. Обработка естественного языка на Java. Москва: ДМК Пресс, 2016. 263 с.

[5] Rudy Prabowo, Mike Thelwall. Sentiment Analysis: A Combined Approach // Journal of Informetrics, 127-129. 2009.