Автоматическое определение тональности контента методами машинного обучения

Введение

Мы живём во время, когда Интернет есть в каждом нашем устройстве, независимо от того, где мы находимся и чем занимаемся. Что касается России, то к началу 2019 года количество интернет-пользователей среди населения 16+ стало равно 90 миллионов человек, что есть 75,4% взрослого населения страны.[[1]](#footnote-2)

Проникая во всё более новые отрасли нашей жизни, интернет стал наполняться огромным количеством информации, генерируемой потребителями продуктов. Стоит заметить, что почти у каждого конкурентноспособного бренда есть как минимум свой сайт или форум, поэтому клиенты имеют возможность расписать свой личный опыт использования того или иного продукта.

Если смотреть на этот процесс со стороны успешного функционирования бизнеса, то одна из его важнейших частей – понимание клиента, который является потребителем, а также его потребностей. Приблизительно 95% потребителей в возрасте от 18 до 34 лет читают отзывы на продукт или услугу перед покупкой, из них 91%[[2]](#footnote-3) доверяют онлайн отзывам так же, как и личным рекомендациям.

~~Поэтому сейчас набирают популярность сервисы, которые следят за реакцией клиента.~~

Обычно перед компанией стоит задача по мониторингу реакции клиента на:

1. Выпуски новых продуктов (запуск новых сервисов);
2. Внесение изменений в существующие продуктах (услуги);
3. Внедрённые или готовящихся к внедрению инициатив, связанных с обслуживанием клиентов;
4. Запуск новых рекламных компаний;

Компаниям важно понимать, какие их действия вызывают положительный отклик у клиентов, а какие отрицательный.

Зачастую, количество отзывов по конкретным продуктам (услугам), может превышать 10000 записей в день. Данная цифра включает в себя также и сообщения, поступающие в службу поддержки, которые необходимо сразу же сортировать и классифицировать, ведь от этого зависит опыт использования продукта клиентом и его мнение о том, насколько дорог он как клиент для компании.

Одной из задач, которая стоит перед компаниями сейчас – является задача по определению тональности текста (в пер. с англ. sentiment analysis).

Сейчас решением данной задачи на Российском рынке занимаются два крупных сервиса: YouScan и Brand Analytics.

С одной стороны, эти системы кажутся абсолютно схожими – они решают одинаковые задачи и дают примерно одинаковый анализ.

Но на самом же деле, существуют различия, о которых нельзя не упомянуть при изучении этих сервисов.

Оба сервиса обладают понятным интерфейсом.

----описание заканчивается из файла

Главной же проблемой данных решений является их цена и ограниченный функционал. Рассмотрим ценообразование, которое используется сервисами YouScan и Brand Analytics по минимальной стоимости:

* Standard[[3]](#footnote-4), YouScan – 35 000 руб / мес
  + за каждые 5 тем в месяц
  + 100 000 упоминаний в каждой теме
* Стартовый плюс[[4]](#footnote-5), Brand Analytics – 35 000 руб / мес
  + сообщений в аккаунте за месяц: 20 000
  + количество тем: 5
  + Автоматическая разметка тематик (тегов на тему): 30

Как мы видим, сервисы по своей сути предлагают одинаковый пул услуг: ограниченный набор тем (вы можете отслеживать только 5 услуг/товаров/запросов) и имеют ограничение на количество упоминаний (другими словами, за эти деньги, больше указанного выше лимита вы увидеть не сможете).

Большим и достаточно развитым бизнесам не составит труда купить подписку подороже и отслеживать свои товары на ежедневной основе, получая за большие деньги достаточно хороший сервис.

Однако здесь оказывается незанятой другая ниша рынка – компании, имеющие достаточно много продуктов (брендов 50), обработка отзывов на которые не играет жизненной роли для компании, но такой глубокий анализ было бы неплохо проводить раз в месяц по всем продуктам компании. Такой анализ был бы полезен для того, чтобы отслеживать реакцию потребителей и смотреть на то, как они воспринимают/потребляют продукт и с какими сложностями они сталкиваются.

К таким компаниям относятся:

* Фармацевтические компании;
* Медицинские клиники;
* Агрегаторы такси;

ЗДЕСЬ ПРОСТО ВСТАВИТЬ ОПИСАНИЕ СТАТЬИ, НУЖЕН ЛИ КАКОЙ-ТО переход?

## Описание методов

Предобработка текста

В анализе тональности исследователями выделяется три уровня классификации[2]:

* Классификация на уровне документа;
* Классификация на уровне предложения;
* Классификация по аспектам сущности, в нашем случае отзыва ~~(направленности отзыва)~~ (выделения точки зрения относительно той или иной стороны продукта/услуги про которую идёт речь) aspect level

Lexicon based sentiment analysis

Lexicon based sentiment analysis is an attractive research area that

uses methods such as WordNet distance to label the sentiment as

fine and dire. [6]

На уровне целого документа мы рассматриваем текст как единицу информации, которую нам необходимо классифицировать. Тогда как на уровне предложения за единицу информации берётся каждое предложение в тексте и определяется только его тональность.

Первым этапом при анализе предложений является определение того, высказывает ли конкретное предложение объективное мнение или субъективное. Если мнение, высказанное в предложении, является субъективным, то мы определяем содержит ли отзыв положительную точку зрения или отрицательную.

### Выбор уровня анализа данных (уровень предложения или уровень всего документа) роли не играет, так как предложение может быть представлено как короткий документ. [3] Однако, классификация текста на таких уровнях не даёт точно представления об объекте (или стороне объекта), о котором высказывается положительная/отрицательная точка зрения. Для анализа предложений на таком уровне, нам необходимо перейти к анализу конкретных точек зрения относительно объекта.

При таком уровне анализа, первым этапом является выделение некой точки зрения о конкретной стороне объекта. Люди могут высказывать диаметрально противоположные точки зрения относительно одного объекта, например, “Данное лекарственное средство хорошо лечит боли в голове, однако вызывает сонливость, что является очень опасным фактором при вождении.”

Также стоит учитывать, что анализ тех или иных отзывов, содержащихся в интернете, обладает некоторыми особенностями.

К таким особенностям обычно относят:[[5]](#footnote-6)

1. Присутствие слова важнее, чем частота с которой данное слово встречается в тексте.

В предыдущих работах было доказано [1], что методы, опирающиеся на само присутствие слова в предложении, с более высокой вероятностью определяют тональность текста

1. Обработка двойных отрицаний

Например, “Неправда, что этот препарат не лечит боли в голове“, будет значить, что потребитель считает, что препарат помогает при болях в голове.

1. Часть речи (POS, Part Of Speech) – определенные классы, в которые распределяются слова в зависимости от их грамматических свойств.

Определение части речи, к которой принадлежит то или иное слово может помочь избавиться от неоднозначности его значения.

Например, слово “Отлично”:

* Принадлежит к наречию, если говорим: “Отлично себя чувствую”
* Принадлежит к имени существительному: “Сдать экзамены на отлично!”

Перед тем как начать изучать соответствующие методы по анализу текста, необходимо описать один из важнейших предшествующих процессов – предобработку текста.

Текст является последовательностью слов, слова же в свою очередь являются имеющими смысл последовательностями символов.

Одним из первоначальных этапов, следующим после сбора необходимой информации, является процесс под названием токенизация, представляющий из себя разделение текста на отдельные слова (токены). В русском языке сделать это просто – разделить предложения на отдельные слова, взяв за разделители пробелы.

Получив последовательность слов из текста, необходимо привести отзывы к определённой единой форме, для чего принято использовать 2 процесса:

* Сте́мминг (на англ. Stemming) – процесс выделения основы у слов, удаление суффиксов.

Проблема: может давать ошибки при преобразовании некоторых слов.

Например: “Ключевой момент в данном продукте заключается в том, что он является хорошим со всех сторон” или “Ключ электронный и иногда может разряжаться”. В обоих предложениях после применения стемминга будет стоять слово ключ, однако, применительно к первому примеру, это будет не совсем корректно.

* Лемматиза́ция (на англ. Lemmatization) – процесс преобразования слов к их лемме – нормальной (словарной) форме.

Проблема: не все формы слов корректно преобразуются.

В большинстве случаев достаточно выбрать один из этих методов и использовать его для преобразования текста. Также методы очень хорошо решают проблему размерности матриц, получаемых при преобразовании текста к числовому методу, что будет рассмотрено ниже.

Остальная работа с текстом заключается в приведении всех букв в тексте к нижнему регистру и, зачастую, удалении знаков препинания. Это делается для того, чтобы все слова воспринимались методами одинаково, а не как разные слова.

# Описание методов первоначальной обработки текстов

Исторически, одним из самых первых подходов, отвечающих за определение тональности того или иного текста, являлся словарный подход. Суть подхода состоит в том, что он определяет тональность предложений за счёт поиска в нём “хороших” и “плохих” слов.

Недостатками метода являются:

1. Не разрешает определять принадлежность сравнений к определённой категории. В некоторых случаях тональность текста определится неверно (прямо противоположно);

Например, “X - отличный продукт, тогда как Y просто ужасен”.

Видим, что тональность данного предложения неоднозначна, отсюда следует, что модель с вероятностью 50% укажет верный/неверный результат.

1. Не позволяет обрабатывать контекст предложения;

Например,

1. Не может различить разные по значению слова (не обрабатывает омонимию и полисемию слов);

Например,

К очевидным плюсам данного метода относится то, что он очень прост в реализации, поэтому он часто используется в первичных прототипах, хотя некоторые компании используют его и в производственных масштабах.

Не так давно самым используемым и простым методом для представления предложений в числовом виде являлась модель под названием Bag of Words (пер. с английского: мешок слов, BoW) [4].

Представляется данный метод матрицей, в которой:

* Строки отвечают за предложения или отдельный текст
* Столбцы же содержат слова (токены), из которого состоит конкретное предложение

Отсюда следует, что ячейки матрицы отвечают за число повторений конкретного слова в тексте. Мотивацией данного метода является то, что мы ищем слова маркеры, такие как: “круто”, “отвратительно”, “превосходно”в каждом предложении, считаем их количество и анализируем частоту с которой данные слова встречаются в предложении.

Пример матрицы, которая создаётся методом:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Текст | достаточно | хороший | препарат | помогает |
| Достаточно хороший препарат | 1 | 1 | 1 |  |
| Препарат помогает |  |  | 1 | 1 |

Проблемы метода:

* Теряется порядок слов в предложении, который может быть полезен для определения конечной тональности;
* Проблема размерности, так как размерность матрицы будет совпадать с количеством исключительных слов в предложении, что может вызывать переобучение модели;

Исследователями отмечается тот факт, что определение порядка слов является очень полезным для повышения точности модели, но при этом достаточно энергоёмким процессом.

Для того, чтобы в некотором смысле учесть порядок слов принято использовать пару (или словосочетания из 3-ёх и более слов) токенов – n-граммы (N-gram) [5].

Проблема:

* Возникает слишком много уникальных элементов, особенно при обработке большого количества текстов.

Для решения такого рода проблем принято избавляться от n-грамм в зависимости от частоты их появления в тексте. Для этого происходит удаление самых частых n-грамм, которые принято называть стоп-словами. Стоп-слова – слова, которые не добавляют особой сути предложению. Популярным также является удаление наименее частых n-грамм, оставляют только те, которые являются средними по частоте появления;

# Обзор методом машинного обучения для определения тональности

## Постановка задачи машинного обучения

Имеем обучающие записи, представленные в виде , где каждая запись является размеченной, или, что тоже самое – ей в соответствие поставлен .[2]

Модель обучается на заданной выборке и после этого становится способной к тому, чтобы определять класс у неразмеченных данных. Такие модели относятся к моделям обучения с учителем, где обязательным условием является существование размеченной раннее выборки данных.

Универсальный алгоритм должен обладать достаточно большим набором параметров, таких как:

1. Если слова, обозначающие тональность в предложении и объединённые союзом “и” с высокой вероятностью будут принадлежать к одной сентиментальной группе

<https://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1913/2215>

1371-1373

1. Если слова, обозначающие тональность в предложении объединены союзом “но”, то с высокой вероятностью эти слова будут относиться к различным сентиментальным группам

ИДЕИ:

1. Можно ли провести своё исследование (среди молодёжи) по поводу того, как часто перед покупкой или перед тем как использовать услугу, они читают отзывы о ней?

1. https://www.computerra.ru/234277/61-rossiyan-polzuyutsya-internetom-na-mobilnyh-ustrojstvah/ [↑](#footnote-ref-2)
2. https://www.brightlocal.com/learn/local-consumer-review-survey/ [↑](#footnote-ref-3)
3. https://youscan.io/price/ [↑](#footnote-ref-4)
4. https://br-analytics.ru/price/ [↑](#footnote-ref-5)
5. https://pdfs.semanticscholar.org/bd93/a70ff4ca27b4a1a0898a37aca30254cee8de.pdf [↑](#footnote-ref-6)