Тема: «Анализ настроений в обзорах фильмов»

**Содержание**

1. Введение

* Обзор предметной области
* Подходы к классификации
* Практическое применение

1. Постановка задачи

* Описание данных
* Выбор метрики качества

1. Приближения модели
2. Построение математической модели

* Bag of Words + TF-IDF
* Word2Vec

1. Возможные методы решения задачи
2. Сравнение описанных моделей
3. Заключение
4. Литература

**Введение**

**Обзор предметной области**

Обработка естественного языка (NLP) — общее направление искусственного интеллекта и математической лингвистики. NLP изучает проблемы компьютерного анализа и синтеза естественных языков и представляет собой огромный спектр задач разного уровня. Одной из главных задач NLP является анализ тональности текста.

Анализ тональности текста – это область компьютерной лингвистики, занимающаяся выделением из текстов эмоционально окрашенной лексики или эмоциональной оценки автора.

Основной целью анализа тональности является нахождение мнений в тексте и выявление их свойств. Какие именно свойства будут исследоваться, зависит от поставленной задачи. К примеру, целью анализа может быть автор, то есть лицо, которому принадлежит мнение.

Основной же задачей анализа тональности текста является его классификация по тональной оценке. Под тональной оценкой или тональностью понимается эмоционально окрашенная лексика и эмоциональная позиция, выраженная автором относительно чего-либо. Тональность может быть позитивной, негативной или нейтральной. Тональность всего текста в целом можно определить как функцию (в простейшем случае сумму) лексических тональностей составляющих его единиц (предложений, слов, словосочетаний) и правил их сочетания.

В основном используется бинарная классификация текста, то есть применяется два типа тональных оценок: позитивная и негативная.

**Подходы к классификации**

Существует несколько подходов к классификации тональности:

1. Подход, основанный на правилах.

Цель данного подхода — поверхностный синтаксический анализ на основе какого-либо правила, которое будет определять тональность текста. Для этого текст разбивается на слова или последовательности слов. Затем полученные данные используются для выделения часто использующихся шаблонов, которым присваивается позитивная или негативная оценка. Такой подход показывает хорошую точность при большом количестве правил.

1. Подход, основанный на заранее составленных тональных словарях с применением лингвистического анализа.

Словарь представляет собой список слов с приписанной им тональностью. Для его составления сначала отрабатывает отдельный лингвистический модуль, автоматически производящий морфологический анализ текста. Затем все слова размечаются по заранее подготовленным спискам тональной лексики. Каждому слову приписывается два атрибута, указывающие на тональность и/или силу тональности. Если слово не нашлось в списках тональной лексики, то оно считается нейтральным. После этого запускается первичный синтаксический анализ: слова и словосочетания объединяются в тональные цепочки, в предложении выделяются субъект, предикат и объект. Затем подсчитывается общая тональность текста.

1. Машинное обучение без учителя.

TF или частота слова - это отношение количества вхождения конкретного термина к суммарному набору слов в исследуемом тексте (документе). Этот показать отражает важность (весомость) слова в рамках определенной статьи/публикации.

IDF или обратная (инвертированная) частота документа - это инверсия частотности, с которой определенное слово фигурирует в коллекции текстов (документов). Благодаря данному показателю можно снизить весомость наиболее широко используемых слов (предлогов, союзов, общих терминов и понятий).

На отношении этих показателей (TF/IDF) и основан данный подход, то есть наибольший вес в тексте имеют термины, которые чаще встречаются в этом тексте, и в то же время присутствуют в небольшом количестве текстов всей коллекции (показатель TF/IDF). Выделив эти термины и определив их тональность, можно сделать вывод о тональности всего текста целиком.

1. Машинное обучение с учителем.

Суть данного подхода заключается в построении статистического или вероятностного машинного классификатора (например, байесовского) на заранее размеченных текстах, а затем в его использовании при анализе новых текстов. Краткий алгоритмданного подхода:

1. Собирается коллекция текстов, на основе которых обучается машинный классификатор.
2. Каждый текст раскладывается в виде вектора признаков, по которым в дальнейшем он будет исследоваться.
3. Указывается правильный тип тональности для каждого текста.
4. Производится выбор алгоритма классификации и метод для обучения классификатора.
5. Полученная модель используется для определения тональности текстов новой коллекции.
6. Комбинации перечисленных подходов.

Данный подход сочетает все или несколько подходов, рассмотренных выше, и заключается в применении классификаторов на их основе в определенной последовательности.

Примеры таких комбинаций:

* TF-IDF + logistic regression
* TF-IDF + naive bayes
* TF-IDF + gradient boosting
* Word2Vec + logistic regression

Word2Vec – набирающая в последнее время технология от *Google*, использующаяся для статистического анализа больших массивов текстовой информации. Она собирает статистику по совместному появлению слов в фразах, после этого с помощью нейронных сетей решает задачу уменьшения размерности и в итоге выдает компактные векторные представления слов, достаточно полно отражающие отношения этих слов в обрабатываемых текстах.

**Практическое применение**

Анализ тональности текста имеет важное практическое применение и находит его во множестве областей. К примеру, в бизнес сегменте, социальных и политических исследованиях:

* определения уровня лояльности потребителя к продукту. На основе данных мониторинга социальных систем делаются выводы о популярности того или иного продукта, нахождение текущих трендов среди покупателей;
* анализ данных о политических позициях пользователей, прогнозирование результатов выборов;
* на основе анализа тональности текстов новостных лент, обзоров финансовых аналитиков, отчетов трейдеров, а также общего настроения пользователей социальных сетей определяется корреляция этих данных с трендами фондовых рынков и строятся прогнозы изменения цен финансовых активов.

Так же анализ тональности тестов используется для борьбы с киберпреступностью. Например, используется автоматический способ идентификации людей, склонных к педофилии, на основе анализа текстовых сообщений в социальных сетях.

Так же приведем примеры готовых программных продуктов, использующих анализ тональности текстов:

1. Tweet Sentiment Visualization App - позволяет анализировать информацию о продукте, который упоминают пользователи, при помощи данных из веб-сервиса   
   Twitter. Пользователь получает в ответ на свой запрос подборку позитивных, негативных или нейтральных микросообщений. Сервис визуализирует соответствующий результат при помощи инфографики. Для анализа тональности данный веб-сервис использует тональный словарь.
2. Text Analytics API - это набор веб-служб для анализа текста, созданных с использованием алгоритмов машинного обучения Microsoft. API можно использовать для таких задач, как анализ настроений, извлечение ключевых фраз и определение языка.

**Постановка задачи**

В качестве предметной области, представляющей входные данные, была выбрана тема обзоров фильмов. Необходимо по заданному отзыву, определить к какому классу относится фильм, к классу «хороший» или «плохой».

**Описание данных**

Входные данные представляют собой три файла:

* labeleledTrainData.tsv – маркированный обучающий набор. Файл разделен табуляцией, имеет строку заголовка, за которой следует 25 000 строк, содержащих идентификатор, тональность и текст для каждого отзыва.
* unlabeledTrainData.tsv - дополнительный обучающий набор без меток. Файл с разделителями табуляции, имеет строку заголовка, за которой следуют 50000 строк, содержащих идентификатор и текст для каждого отзыва.
* testData.cvs – тестовый набор. Файл также табулирован и имеет заголовок, за которым следует 25 000 строк, содержащих идентификатор и текст каждого отзыва.

**Выбор метрики качества**

Так мы рассматриваем задачу бинарной классификации и наши данные являются сбалансированными, то для оценки качества работы алгоритма выберем площадь под ROC-кривой (AUC-ROC). Так же данный показатель часто используется для сравнительного анализа нескольких моделей классификации.

ROC-кривая – графическая характеристика качества бинарного классификатора, которая характеризует зависимость доли верных положительных классификаций от доли ложных положительных классификаций.

Площадь под ROC-кривой является объединенной характеристикой качества классификации, не зависящей от соотношения цен ошибок. Чем больше значение площади под ROC-кривой, тем «лучше» модель классификации.

**Приближения модели**

В качестве приближений модели приведем описание предварительной обработки данных - процесса очистки и подготовки текста к классификации.

Предварительная обработка текста включает в себя:

* удаление HTML тегов;
* удаление всех не буквенных символов (знаков препинания, чисел и т.д.) и замена их на пробелы;
* конвертация слов в нижний регистр;
* удаление "стоп-слов ", к примеру, таких как "a", "and", "is" и т.д. Так же удалим слова "movie" и "film", так как скорее всего они будут довольно часто встречаться в отзывах, но интереса для модели не представляют.

Кроме всего выше перечисленного к приближениям модели так же относится и то, что все отзывы написаны на английском языке.

**Построение математической модели**

После того, как мы провели предварительную обработку текстов отзывов необходимо преобразовать их в какое-либо числовое представление для дальнейшей работы.

**Bag of Words + TF-IDF**

Модель Bag of Words («мешок слов») – самая популярная и простая модель представления текста как набора слов без учета их взаимного расположения и взаимных связей. Модель сводит текст к вектору, где каждая позиция вектора представляет слово, а значение этой позиции представляет число раз, которое это слово используется в тексте. В данных IMDB содержится большое количество рецензий, следовательно, необходимо ограничить размер векторов объектов, то есть выбрать максимальный размер словаря.

Добавим дополнительное преобразование TF-IDF. Данное преобразование используется для корректировки значений вектора в соответствии с числом рецензий, использующих слово. Слова, встречающиеся во многих отзывах, могут быть менее дискриминационными, чем слова, встречающиеся реже. Преобразование TF-IDF уменьшает значение данного слова пропорционально количеству документов, в которых оно появляется.

**Word2Vec**

Word2vec - реализация нейронной сети, c помощью которой технология выдает компактные векторные представления слов. Для любой пары таких векторов мы можем найти меру схожести. В простейшем случае этой мерой будет косинусная мера сходства (скалярное произведение векторов):

где – векторные представления слов, *α* – угол между ними.

Предобработка данных для Word2Vec немного отличается от предыдущей модели. Во первых, нет необходимости удалять "стоп-слова" (лучше не удалять), поскольку алгоритм использует более широкий контекст предложения для получения высококачественных векторов слов.

Во вторых, для преобразования текста Word2Vec ожидает отдельные предложения, каждое из которых представляет собой список слов, следовательно, необходимо разбить абзацы на предложения.