Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Новгородский государственный университет имени Ярослава Мудрого»
Институт электронных и информационных систем
Кафедра Информационных технологий и систем

Отчёт по проектно-технологической практике и научно-исследовательской работе на тему «Исследование клиентской обратной связи онлайн сервисов».

Проверил
_____Макаров В.А.
«____» _____ 2022 г.

Выполнил студент группы 2096
_____Скородумов С.С.
«____» _____ 2022 г.

Великий Новгород
2022 г.

Актуальность работы.

Спрос на знания, которые можно получить на основе результатов анализа тональности текстов, весьма велик. С развитием и популяризацией социальных сетей, люди стали выкладывать свои мысли, личную информацию и прочее в открытый доступ. Помимо социальных сетей существует огромное множество сервисов, чаще всего узконаправленных, где люди делятся мнениями и обсуждают отдельные продукты, события или определенные сферы жизни, и некоторые из таких сайтов крайне популярны. Это означает, что теперь любой может собрать и проанализировать мнения интернет-пользователей по поводу интересующей его сферы или продукта, чем на данный момент активно пользуются большинство компаний. В последнее время использование онлайн сервисов для просмотра продуктов кинематографа значительно выросло в связи с чем становиться актуальным разработка способа понять и проанализировать общественное восприятие различных идей и концепций или недавно запущенного продукта. Одним из таких способов является определение тональности текста в комментариях, оставленных после просмотра кинопродукта.

Объект исследования.

Объектом исследования является задача анализа тональности текста в комментариях.

Цель исследования.

Целью данной работы является выявление лучшего метода распознавания тональности текста на основе методов машинного обучения на выбранном наборе данных.

Исследование аналогов.

Sentiment Analysis — это общий классификатор анализа настроений для текстов на английском языке (положительная, нейтральная, отрицательная оценка).

EUREKA ENGINE — это высокоскоростная система лингвистического анализа текстов модульного типа, позволяющая извлекать новые знания и

факты из неструктурированных данных огромных объемов. В том числе данная система может обработать не только правильный «книжный» язык (СМИ, внешний документооборот), но и сообщений социальных сетей, форумов, блогов. Одной из функций системы является снижение уровня конфликтов и повышение качества обслуживания путем раннего выявления смены тональности переписки с контрагентами и клиентами.

Возможные наборы данных для обучения.

1. Корпус коротких текстов Рубцовой Ю., предварительно разделенный на негативные и позитивные предложения, собранные на площадке Твиттер.
2. Набор данных обзора IMDB. Он имеет 50 000 отзывов и соответствующие им мнения, отмеченные как «Положительные» и «Отрицательные»

Обзор объекта исследования.

Анализ тональности (сентимент-анализ) — инструмент компьютерной лингвистики, оценивающий такую субъективную составляющую текста, как отношение пишущего.

При классификации полярности текста пользуются определенной шкалой — набором эмоций, по которым определяется эмоциональная окраска каждого текста. В зависимости от используемой шкалы меняется и задача сентимент-анализа. Так, шкала может иметь набор множества разных эмоций, например, «злой, добрый, грустный, веселый и т.д.». Шкалы подобного вида по-разному нагружены эмоционально, и, как следствие, возникает проблема однозначности классификации текста по данной шкале, то есть один текст может быть оценён несколькими людьми по-разному. По этой причине использование подобных шкал при анализе полярности текста практикуется довольно редко. Для простоты множество возможных значений тональности обычно сводится к шкале «позитивный-нейтральный-негативный». Однако зачастую из множества возможных классов убирают

«нейтральный», то есть тональность определяется по шкале «положительный-отрицательный». Подобная бинарная шкала является самой распространённой, так как в большинстве задач заказчика интересует именно мнение большинства, то есть как народ относится к выпускаемой им продукции/услуге.

Все подходы к анализу тональности можно разделить на три группы. Первая — подходы на основе правил. Чаще всего в них используются вручную заданные правила классификации и эмоционально размеченные словари. Эти правила обычно на основе эмоциональных ключевых слов и их совместного использования с другими ключевыми словами рассчитывают класс текста. Несмотря на прекрасную эффективность в текстах из какой-то определенной тематики, методы на основе правил плохо способны обобщать. Кроме того, они крайне трудоёмки в создании, особенно когда нет доступа к подходящему словарю настроений.

Вторая группа — подходы на основе машинного обучения. Они используют автоматическое извлечение признаков из текста и применение алгоритмов машинного обучения. Классическими алгоритмами классификации полярности являются наивный байесовский классификатор (Naive Bayes Classifier), дерево решений (Decision Tree), логистическая регрессия (Logistic Regression) и метод опорных векторов (Support Vector Machine). В последние годы внимание привлекают методы глубокого обучения, которые значительно превосходят традиционные методы в анализе тональности (свёрточные (CNN) и рекуррентные (RNN) нейросети, а также методы переноса обучения (transfer learning)). Одна из главных особенностей систем на основе машинного обучения — автоматическое извлечение признаков из текста. В простых подходах для представления текста в векторном пространстве обычно используется модель «мешок слов» (bag of words). В более сложных системах для генерирования эмбеддингов слов применяются модели дистрибутивной семантики, например, Word2Vec,

GloVe или FastText. Одним из их главных недостатков с точки зрения генерирования эмбеддингов является потребность в больших массивах текстов для обучения. Однако, это справедливо для всех методов машинного обучения, потому что всем алгоритмам обучения с учителем нужны для обучения размеченные наборы данных.

Третья группа — гибридные подходы. Они объединяют в себе подходы двух предыдущих видов. С одной стороны, комбинация методов на основе правил и машинного обучения обычно позволяет добиться более точных результатов. А с другой — гибридные подходы наследуют трудности и ограничения составляющих их алгоритмов.

В процессе исследование будут исследованы подходы к анализу текста второй группы. Для сравнения подходов, принадлежавших ко второй группе, будут использоваться следующие метрики:

1. Истинно положительные (true positives, TP) – число комментариев, которые модель правильно предсказала как положительные.
2. Ложноположительные (false positives, FP) – число комментариев, которые модель неверно предсказала как положительные, хотя на самом деле они были негативными.
3. Истинно отрицательные (true negatives, TN) – число комментариев, которые модель правильно предсказала как негативные.
4. Ложноотрицательные (false negatives, FN) – число комментариев, которые модель неверно предсказала как негативные, хотя на самом деле они были положительными.

На основе четырех описанных статистических данных вычисляются две метрики: точность и полноту. Эти метрики являются показателями эффективности модели классификации:

Точность (precision) – отношение истинно положительных результатов ко всем элементам, отмеченным моделью как положительные (истинные и

ложные срабатывания). Точность 1.0 означает, что каждый отзыв, отмеченный моделью как положительный, действительно относится к положительному классу: $precision = \frac{TP}{TP + FP}$

Полнота (recall) – это отношение истинно положительных отзывов ко всем фактическим положительным отзывам, то есть количество истинно положительных отзывов, деленных на суммарное количество истинно положительных и ложноотрицательных отзывов: $recall = \frac{TP}{TP + FN}$

F1-мера – среднее гармоническое точности и полноты. Максимизация F1-меры приводит к одновременной максимизации этих двух критериев: $F_1 = \frac{2 * precision * recall}{precision + recall}$

Список используемой литературы.

1. SENTIMENT ANALYSIS ON TWITTER POSTS // https://www.researchgate.net/ URL: https://www.researchgate.net/publication/362491603_SENTIMENT_ANALYSIS_ON_TWITTER_POSTS
2. Sentiment Analysis of Twitter Data // https://www.researchgate.net/ URL: https://www.researchgate.net/publication/365618365_Sentiment_Analysis_of_Twitter_Data
3. Performance Analysis of Different Neural Networks for Sentiment Analysis on IMDb Movie Reviews // https://www.researchgate.net/ URL: https://www.researchgate.net/publication/343046458_Performance_Analysis_of_Different_Neural_Networks_for_Sentiment_Analysis_on_IMDb_Movie_Reviews