МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ К ЗАДАЧАМ АНАЛИЗА ТЕКСТОВ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Нарватова Вадима Валерьевича

Научный руководитель
доцент, к. ф.-м. н. _____ С. П. Сидоров

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2024

ВВЕДЕНИЕ

Актуальность темы заключается в том, что в современном мире генерируется огромное количество информации, которое распространяется каждую секунду и только растет с каждым годом. Одним из наиболее распространенных источников информации являются текстовые данные. Однако, объем текстовой информации столь велик, что ее анализ вручную становится невозможным заданием. Для сокращения расходов на ее обработку возникает потребность в автоматизации. Это создает необходимость в разработке средств, которые позволят анализировать и обрабатывать большие объемы текстовых данных.

Применение методов машинного обучения к анализу текстовых данных имеет огромный потенциал в различных областях, таких как обработка естественного языка, информационный поиск, анализ социальных медиа, медицинская диагностика, финансовая аналитика и многое другое. С развитием технологий и доступностью больших объемов данных возможности применения методов машинного обучения для анализа текстовых данных становятся все более широкими и перспективными. Такие методы позволяют автоматизировать процессы анализа текста, извлечения информации, классификации текстов и принятия решений на основе текстовых данных, что делает данную тему крайне актуальной для исследований и практического применения.

В работе исследуются два подхода к задаче определения тональности текста: подход, основанный на методах машинного обучения и подход, основанный на использовании словарей тональной лексики. Описан и реализован метод для автоматического извлечения из текста слов, несущих эмоциональную оценку. Описан и реализован метод последующей классификации текстов на основе полученного словаря тональностей.

Целью бакалаврской работы является рассмотрение классов задач и методов анализа текста, проведение анализа тональности текста для конкретной выборки данных и визуализация полученных результатов, исследование, модификация, программная реализация и последующее сравнение качества методов автоматического построения словаря тональностей. Последующее сравнение алгоритмов классификации текстов на основе полученного

словаря и на основе методов машинного обучения.

Объект исследования - средства data-mining для анализа текстовой информации, словарь тональной лексики.

Предмет исследования - отзывы пользователей Интернет-ресурса imhonet.ru по трем предметным областям: книги, фильмы, камеры.

Данные были представлены на Российском семинаре по оценке методов информационного поиска.

Для достижения поставленных целей в работе необходимо решить следующие **задачи** :

- обзор теоретических аспектов анализа текстовой информации и классификации текстов;
- исследование инструментов анализа текстовой информации;
- определение основных моделей, используемых в text mining;
- проведение анализа тональности текста с помощью языка программирования Python и наглядная демонстрация полученных результатов;
- исследование и разработка метода извлечения оценочных слов для заданной предметной области;
- разработка методов классификации текстов на основе построенного словаря;
- провести эксперименты на реальных данных.

Практическая значимость проводимого исследования, состоит в автоматическом извлечение мнений из текстов, то есть определении, содержит ли данный текст субъективную составляющую, а также классификация текстов на основе тональности на два (позитив и негатив) или более классов. За счет скорости поиска, структурирования информации, эффективности использования для конечного потребителя, применение методов машинного обучения упрощает обработку большого объема информации. Под тональностью здесь понимается эмоциональная оценка, выраженная автором относительно некоторого объекта.

Структура и содержание бакалаврской работы. Работа состоит из введения, трех разделов, заключения, списка использованных источников, содержащего 20 наименований, и трех приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы работы, формулируется цель работы и решаемые задачи, отмечается практическая значимость полученных результатов.

В **первом** разделе рассматриваются классы задач анализа текстов и методы анализа текстов. Производится обзор основных подходов к решению данных задач. В главе рассмотрен подход к определению качества работы программ-классификаторов.

Основные задачи анализа текстов:

— Извлечение ключевых слов и фраз: определение наиболее важных слов и фраз, которые наиболее точно описывают содержание текста, и которые могут быть использованы для дальнейшей обработки.

— Классификация текстов: разделение текстов на категории в соответствии с их содержанием или темой. Это может быть полезно для автоматической обработки больших объемов информации.

— Извлечение информации: определение в тексте конкретных фактов, событий, дат, имен, мест и другой информации, которая может быть использована для дальнейшей обработки или анализа.

— Анализ тональности: определение отношения автора к определенной теме или событию, используя анализ лексических, семантических и грамматических признаков.

— Анализ синтаксиса: определение структуры предложений в тексте и их взаимосвязей, что может помочь в понимании содержания и смысла текста.

— Анализ семантики: определение значения слов и фраз в контексте текста, что может помочь в понимании содержания и смысла текста.

— Извлечение именованных сущностей: определение имен, мест, организаций и других именованных сущностей в тексте, что может быть использовано для дополнительной обработки и анализа.

— Визуализация: наглядное представление результатов анализа, используя графики, диаграммы и другие инструменты.

— Интерпретация: проведение анализа результатов.

Также рассмотрена работа классификатора и оценка качества его работы.

Обзор существующих методов классификации текстов включает в себя: наивный метод Байеса, деревья, алгоритм Random Forest, метод опорных векторов (SVM), Модель мешок слов. Проведен сравнительный анализ двух подходов к задаче классификации: подхода, основанного на методах машинного обучения и подхода, основанного на использовании словарей тональности.

Проведен сравнительный анализ двух подходов к задаче классификации: подхода, основанного на методах машинного обучения и подхода, основанного на использовании словарей тональности.

Во **втором** разделе рассмотрены методы используемые в работе. В том числе метод странности слов. Для вычисления признака «Странность» необходимо два корпуса текстов. Возможны два варианта:

— Первый текст нейтральный, второй - тональный (положительный или отрицательный). Таким образом слово несущее тональность (положительную или отрицательную) было бы «странно» встретить в нейтральном корпусе текстов, например, в новостях или обзорных статьях – эмоционально окрашенная лексика встречается в таких текстах крайне редко.

— Два корпуса текстов с противоположными тональностями. Один положительный, другой - отрицательный. В данном случае ситуация аналогична. «Странно» встретить ярко окрашенную положительную лексику в негативных комментариях о фильме или недавно купленной бытовой технике.

Результатом работы алгоритма является словарь, в котором каждому слову поставлены в соответствие следующие величины: частота данного слова в первом корпусе текстов, частота слова во втором корпусе текстов, странность слова, нормализованная (по максимальной величине странности среди всех слов) странность слова, тональность.

В результате работы метода получены словари странности слов для нескольких предметных областей. Построенный таким образом словарь тональности может использоваться для классификации на любое число классов.

Также рассмотрен метод классификации текстов на основе странности

слов и классификация методами машинного обучения с использованием словаря тональности.

А также описывается методология. Рассмотрим подробнее:

1. Тестовые данные, используемые в работе, были представлены на Российском семинаре по Оценке Методов Информационного Поиска (РОМИП). Данные размечены экспертно на три класса – нейтральный, позитивный и негативный;
2. Обработка данных. Каждый текст был представлен в виде набора входящих в него слов. После чего была произведена нормализация текста. Слова текста приведены к нормальной форме. Для этого использовалась библиотека PyMorphy2 на языке программирования Python. Выполнено удаление стоп-слов. Слова, имеющие длину менее 4 символов были удалены, так как заведомо не являются оценочными. Данная операция позволила существенно сократить время работы алгоритма построения словаря тональности.
3. Построение словаря тональности. Каждому слову поставлена в соответствие его частота в тексте. После чего эта величина нормируется.
4. Дополнительная обработка словаря тональности. Данный этап включает в себя два шага: 1. Экспериментальное определение порогового значения в словаре тональности, позволяющее существенно уменьшить размер словаря в основном за счет удаление слов, не несущих эмоциональную оценку (шумов). 2. Дополнительная экспертная разметка слов, входящих словарь. Данная операция позволяет свести зашумленность словаря к абсолютному минимуму.
5. Сравнение словарей тональности, которое позволяет определить плотность оценочных слов в словаре и оценить качество полученных словарей.
6. Классификация на основе агрегированного сентимент-значения слов, входящих в документ.
7. Альтернативный вариант – классификация методами машинного обучения с использованием словаря тональностей. Слова, входящие в словарь тональности выступают в качестве атрибутов классификации. Для применения такого метода необходимо составить матрицу термдокумент, в

данной работе значения атрибутов являются бинарными.

8. Классификация одним из методов машинного обучения, с использованием матрицы.
9. Эксперименты на текстовых данных РОМИП. Тестирование методов для различных наборов входных данных: словарей тональности и документов, принадлежащих разным предметным областям.
10. Сравнение методов на основе полученных результатов. Анализ полученных результатов и формулировка выводов.

В **третьем** разделе описываются результаты тестирования реализованных алгоритмов.

Произведены тесты на реальных данных, в качестве примера использовался набор данных обзоров женской одежды для электронной коммерции от Kaggle (Women's E-Commerce Clothing Reviews).

Используем составную оценку для определения метки настроений. Если совокупный балл больше или равен 0,05, классифицируем его как "Положительный". Если итоговый балл меньше или равен -0,05, классифицируем его как "Отрицательный". В противном случае классифицируем его как "Нейтральный".

После успешного завершения анализа тональности текста на основе имеющихся у данных перейдем их визуальному представлению.

Создадим круговую диаграмму для визуализации распределения меток настроений в наборе данных, что показано на рисунке 1:

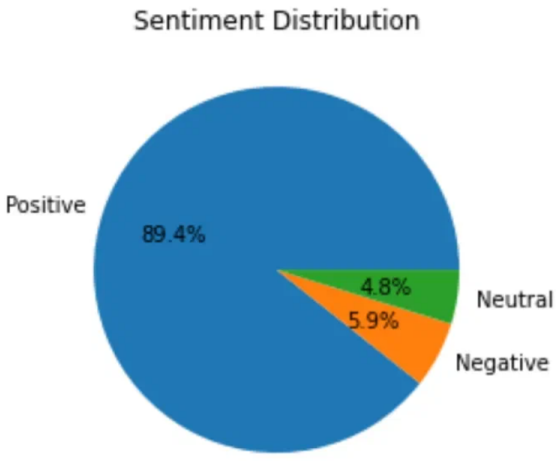

Рисунок 1 – Диаграмма результатов настроений

Выясним, как распределяются настроения для каждого рейтинга, которые наглядно представлены на рисунке 2.

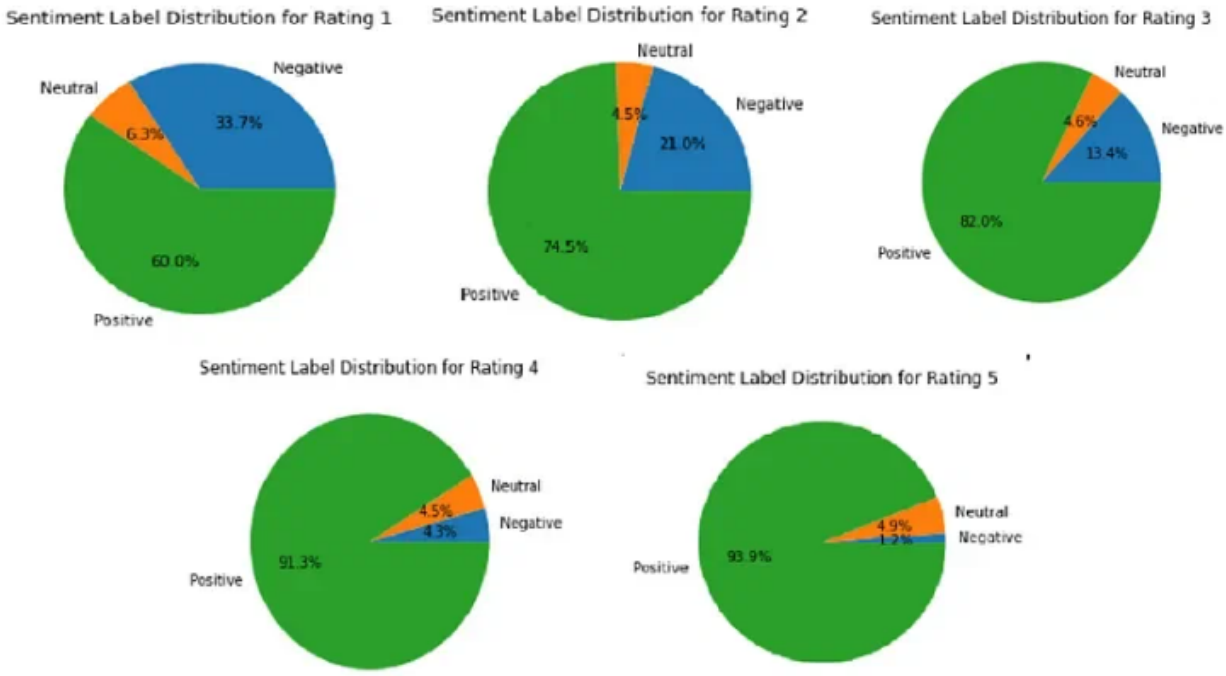

Рисунок 2 – Диаграмма распределения настроения для каждого рейтинга

Также можно сопоставить метку настроений с рейтингом каждого отзыва. Это позволяет увидеть, существует ли какая-либо корреляция между настроениями и рейтинговыми баллами. В результате получим гистограмму, как на рисунке 3:

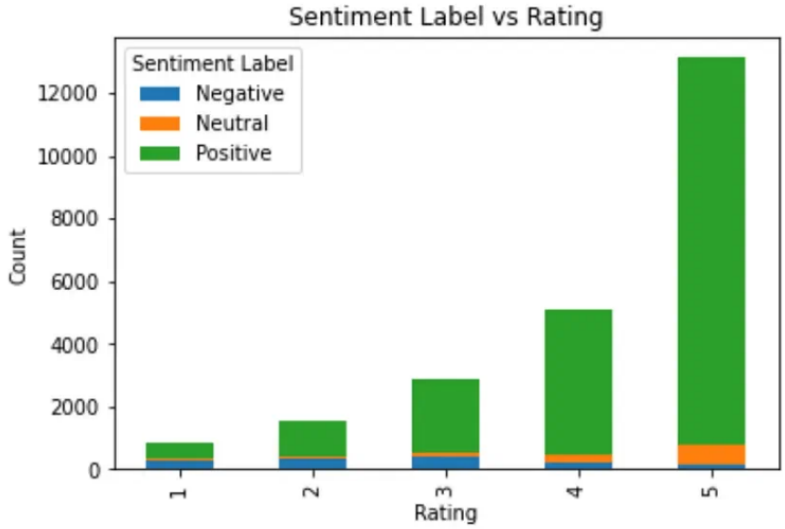

Рисунок 3 – Гистограмма результатов

Таким образом, анализ тональности текста может дать необходимую информацию об отзывах клиентов. Используя такие инструменты, как VADER lexicon и библиотеки Python, такие как pandas и matplotlib, могут выполнять анализ настроений текстовых данных и визуализировать распределение настроений.

Далее в работе представлены эксперименты, направленные на то, чтобы сравнить качество построения словарей тональности для различных предметных областей и различных наборов входных данных.

Тональность текста напрямую зависит от предметной области. В частности, при использовании списка оценочных слов (словаря тональностей) эмоциональная оценка одного и того же слова может меняться в разных предметных областях.

Тексты, содержащие оценочную лексику в каждой из областей, разделены на три части – нейтральные, негативные и позитивные. В дальнейшем в каждом из этих разделов выделены два класса – обучающие и тестирующие тексты. Их доли составляют 90% и 10% соответственно.

Словари строились для трех предметных областей: обзоры пользователей по купленным фотокамерам, отзывы о фильмах, отзывы о книгах. Также был построен словарь тональности по предметной области, состоящей из объединения отзывов о фильмах, камерах и книгах.

Эксперименты показали, что полученные словари чувствительны к шумам, что, безусловно является их существенным недостатком. Построенные списки слов требуют ручной постобработки, что достаточно трудозатратно в силу большого размера полученных словарей (5000-6000 слов).

В главе проведен сравнительный анализ алгоритмов, описанных во второй главе. Методы применились на основе построенного и оптимизированного словаря тональности, а также на размеченном вручную словаре.

Также произведено сравнение методов машинного обучения с методом W на основе результатов тестов.

В результате было обнаружено, что методы машинного обучения в данной работе значительно превзошли метод классификации, основанный на использовании исключительно словарей тональности.

В данной работе производен сравнительный анализ алгоритмов машин-

ного обучения, принимающих в качестве параметров классификации слова из словаря тональностей, и методов, использующих исключительно словарь тональностей.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ:

— Определены основные классы задач. Рассмотрена теоретическая сторона анализа текста;

— Проведена оценка настроения пользователей по отзывам на женскую одежду;

— Построены словари тональности на различных наборах входных данных, произведена оценка качества полученных словарей;

— Произведено сравнение методов машинного обучения при работе со словарем тональностей как с набором атрибутов, сделаны выводы по качеству работы классификатора;

— Произведено сравнение методов машинного обучения с методом классификации с использованием словаря тональностей на основе результатов тестов.