МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра дискретной математики и информационных технологий

РЕАЛИЗАЦИЯ МЕТОДОВ АНАЛИЗА ТОНАЛЬНОСТИ НА ОСНОВЕ ТЕЗАУРУСНОГО ПОДХОДА

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Сторожевой Аллы Алексеевны

Научный руководитель
доцент, к.э.н _____ Г. Ю. Чернышова

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2023

ВВЕДЕНИЕ

В настоящее время анализ настроений имеет множество практических применений из-за быстрого роста объема информации в Интернет, многие тексты выражают мнения на обзорных сайтах, форумах, блогах и в социальных сетях. Мнения лежат в основе многих видов человеческой деятельности и стимулируют поведение субъектов. Поиск сайтов общественного мнения в Интернет и контроль за ними, а также фильтрация содержащейся на них информации остаются сложной задачей из-за распространения подобных ресурсов. Каждый сайт, как правило, содержит большой объем текста с мнениями, который нелегко расшифровать в блогах и на форумах. Пользователю трудно идентифицировать релевантные сайты, а также получать и обобщать соответствующие мнения, поэтому возникает необходимость в использовании систем анализа настроений.

На данный момент существуют различные подходы к решению задачи анализа тональности. Однако наиболее изученными являются подход на основе тезаурусов и подход, использующий машинное обучение.

Анализ тональности, основанный на словарном подходе, состоит в анализе оценочной окраски слов и фраз представленных в тексте. Использование машинного обучения для анализа тональности заключается в предварительном обучении модели классификатора и в последующем его использовании.

Целью данной квалификационной работы является анализ эффективности использования тезаурусного подхода для классификации тональности для корпусов текстов.

Для достижения поставленной цели необходимо решить следующие задачи:

— сравнительный анализ подходов к решению задачи определения тональности;
— реализация методов классификации настроений для текстовых данных при помощи словарного подхода и машинного обучения;
— оценка качества классификации, анализ эффективности использования словарного и автоматического подходов.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе рассматривается актуальность задачи анализа тональности, ее особенности, и проводится сравнение методов классификации тональности, реализуемых при помощи различных подходов.

Анализ тональности является дисциплиной, целью которой является извлечение качественных характеристик из текстовых данных пользователя, таких как настроения, мнения, мысли и поведенческие намерения, с использованием методов обработки естественного языка [1].

Для сбора мнений доступно множество методологий, но используются три основные группы подходов: подход с использованием подхода машинного обучения, подход, основанный на лексике и гибридный подход.

Подходы, основанные на лексике, также известные как подходы, основанные на знаниях, предварительно разрабатываются вручную и относятся к анализу семантических и синтаксических шаблонов. В то время как первое относится к созданию словаря путем пометки слов тегами, второе предполагает рассмотрение синтаксических шаблонов.

Подход, основанный на лексике, делится еще на 2 подтипа: тезаурусный подход и подход, основанный на корпусе.

При подходе с использованием машинного обучения первоначальная классификация выполняется путем взятия двух разных сборок документа: обучающих данных и тестовых данных. Далее из текста извлекаются его особенности и характеристики и затем, он подразделяется на контролируемую систему и неконтролируемую.

При контролируемой системе среди различных наборов данных существует размеченный обучающий набор. При этом каждый тип класса имеет свое собственное свойство, преимущество и связанную с ним метку, которая может использоваться в этой системе. Каждое слово классифицируется по метке в зависимости от его типа и связанных с ним характеристиках.

Основными видами классификаторов, используемых при подходе с машинным обучением, являются вероятностные классификаторы, линейные классификаторы, классификатор дерева решений и классификатор на основе правил.

Гибридный подход сочетает в себе неконтролируемый алгоритм машинного обучения с методами обработки естественного языка, а также исполь-

зуется обработки анализа отзывов и пометки по частям речи для получения синтаксической структуры предложения. Полученная синтаксическая структура, наряду с использованием словарей, позволяет определить семантическую направленность рецензии при помощи алгоритма оценки.

В данной работе для анализа эффективности классификации тональности текстов применены 2 подхода: тезаурусный подход и подход с использованием машинного обучения.

Во втором разделе приводятся основные подходы к составлению словарей, рассматриваются методы классификации тональности при помощи тезаурусов и методы их оценивания.

Список слов и фраз оценочной лексики называется лексикой настроений, тональным словарем или словарем оценочной лексики [2]. Существуют три основных способа построения словарей оценочной лексики: ручной, автоматический и гибридный.

При ручном создании словарей для анализа тональности текстов выделяется два этапа. На первом этапе формируется список слов-кандидатов на вхождение в словарь оценочной лексики. На втором этапе происходит разметка отобранных слов-кандидатов.

При автоматическом создании словарей оценочной лексики, как правило, проходят три этапа. На первом этапе строится пространство поиска, в котором будет производиться разметка слов по тональности. На втором этапе выбирается начальное множество оценочных слов с известной тональностью. На третьем этапе происходит разметка слов в пространстве поиска на основе выбранного начального множества слов (бутстрэппинг).

В гибридном способе совместно используются ручные и автоматические методы разметки [3][4].

Существует множество методов классификации текстов по эмоциональной окраске при помощи тезаурусов. Одними из них являются метод подсчета слов, метод подсчета по тексту и метод соотношения количества позитивных и негативных слов.

При методе подсчета слов (метод 1) происходит вычитание из общего числа слов с позитивной оценкой число слов с негативной оценкой.Если полученное в результате подсчетов число является положительным, тогда оценка текста – положительная, если полученное число отрицательное, то оценка

текста – отрицательная. В случае, если полученное число равно нулю, тогда текст является нейтральным.

Метод подсчета по тексту (метод 2) является более оптимальным для большого объема данных. Этот метод включает в себя метод подсчета слов, однако результат делится на общее количество слов в предложении.

В методе соотношения количества позитивных и негативных слов (метод 3) количество положительных слов делится на количество отрицательных слов, к которому также прибавляется единица. Добавление единицы к знаменателю необходимо для устранения появления ошибки деления на ноль. Если результат подсчета возвращает число равное единице, тогда текст является нейтральным. В ином случае, если результат больше единицы, тогда текст является позитивно окрашенным, иначе отрицательно окрашенным.

В настоящее время существует множество различных показателей для оценки классификации, однако наиболее известными являются следующие четыре показателя: точность(accuracy), точность(precision), полнота(recall) и f1-мера.

Простейшей метрикой для определения доли правильно классифицируемых документов является точность (accuracy). Она присваивает всем документам одинаковый вес, что может быть некорректно в случае если распределение документов в обучающей выборке сильно смещено в сторону какого-то одного или нескольких классов.

Точность системы в пределах класса — это доля документов, действительно принадлежащих данному классу, относительно всех документов, которые система отнесла к этому классу.

Полнота системы — это доля определенных классификатором документов, принадлежащих классу, относительно всех документов этого класса в тестовой выборке.

Чем выше точность (precision) и полнота (recall), тем лучше. Однако в реальной жизни максимальная точность и полнота не достижимы одновременно и приходится искать некий баланс. Поэтому, существует метрика, объединяющая в себе информацию о точности и полноте алгоритма. Такой метрикой является f1-мера:

В данной работе были применены словари КартаСловСент и LisniCrowd, составленные при помощи подхода, основанного на разметке слов усилиями

людей (аннотаторов). Для классификации данных при помощи тезаурусного подхода будут использованы методы подсчета слов, метод подсчета по длине текста и метод соотношения количества позитивных и негативных слов. Для оценки классификации используются показатели точности(accuracy), точности(precision), полноты(recall), f1-меры [5].

В третьем разделе приведен пример применения разработанного приложения для анализа тональности, осуществлен вычислительный эксперимент с использованием различных подходов и типов выборок.

Для разработки интерфейса были использованы библиотеки tkinter и openpyxl. tkinter — это одна из основных библиотек языка Python, предоставляющая набор инструментов и виджетов для создания графических пользовательских интерфейсов и работы с ними. openpyxl — библиотека, позволяющая взаимодействовать с электронными таблицами Excel, манипулировать данными и создавать новые файлы Excel программным путем.

Функционал разработанного приложения позволяет: реализовать функции выбора файла из системы, просмотр данных загруженного файла, выбор словаря тональности, выбор метода классификации тональности для выбранного словаря.

Возможность выбора файла из системы была реализована при помощи функции open_file_dialog(), которая на вход получает файл формата .xlsx, а затем вызывает функцию open_window(), которая выводит на экран пользователя содержимое загруженного файла.

После загрузки файла пользователь может просматривать его содержимое в специальной текстовой области, также, для того чтобы иметь возможность просматривать файл полностью, были созданы специальные ползунки.

Для того чтобы вызвать функцию open_file_dialog(), была реализована кнопка с названием “Открыть файл”.

Для выбора используемого метода классификации тональности была создана структура, называемая combobox. При нажатии на данный объект у пользователя появляется выпадающий список, в котором можно выбрать необходимый метод.

Для получения результата классификации была создана кнопка “Применить”, которая вызывает функцию get_result(), вычисляющую результат классификации тональности.

Данная функция передает выбранные пользователем параметры в класс дочерний Res, который использует их для выбора необходимого алгоритма.

Разработанный интерфейс обеспечивает возможность применения данных методов специалист в бизнес-аналитике без дополнительных квалификационных требований.

В данной работе было использовано два датасета, полученных из базы данных сайта kaggle [6]. Первый набор данных (выборка 1) содержит отзывы на мобильные телефоны. Каждому отзыву соответствует оценка степени удовлетворенности покупкой, что упрощает проверку классификации тональности. Количество объектов составляет 458433. Второй набор данных (выборка 2) содержит отзывы на проживание в отеле. Каждому отзыву соответствует оценка степени удовлетворенности проживанием. Количество объектов составляет 49902. Тексты в выборках является неструктурированными, однако для удобства их использования данные в них разделены табуляцией. Все данные в файлах представлены в виде таблицы эксель.

Для выполнения экспериментальной части работы для тезаурусного подхода была разработана следующая схема вычислительного эксперимента:

1. выполнение предобработки текстовых данных:
   а) очистка данных от url- и html-ссылок, лишних символов, приведение данных к нижнему регистру;
   б) обработка данных при помощи методов NLP-анализа: токенизация, стемминг, лемматизация, удаление стоп-слов;
2. выбор словаря тональности: КартаСловСент или LinisCrowd;
3. выбор метода классификации тональности данных;
4. оценка полученных результатов при помощи мер: точность (accuracy), точность (precision), полнота (recall), f1-мера.

Для выполнения экспериментальной части работы для подхода, основанном на машинном обучении, используется схема, похожая на схему для словарного подхода, однако в этом случае пункт выбора словаря заменяется пунктом выбора, обучения и применения модели для классификации тональности данных.

Перед применением классификации тональности при помощи словарного подхода необходимо было выполнить процесс предобработки датасетов.

Для первоначальной обработки текстов были использованы функции

очистки данных от url и html-ссылок, а также от лишних знаков и символов. Данные действия были выполнены при помощи применения регулярных выражений к текстовым данным. Далее весь текст документа был приведен к нижнему регистру.

Следующим шагом необходимо произвести обработку текстов при помощи методов NLP-анализа: токенизации, стемминга, лемматизации, а также удалить встречающиеся в тексте стоп-слова.

Конечным этапом обработки текста является удаление слов, не несущих информационную нагрузку. Они также называются стоп-словами. Выполнение данного действия повышает качество решаемой задачи.

В данной работе для реализации словарного подхода были использованы словари тональностей КартаСловСент и LinisCrowd [7][8].

Для классификации тональности данных тезаурусным подходом в каждом текстовом документе при помощи словарей происходил подсчет количества слов, имеющих позитивную и негативную окраску, соответственно. Затем, согласно полученным данным, происходила оценка тональности документа при помощи трех методов: метода подсчета слов (метод 1), метода подсчета по тексту (метод 2) и метода соотношения количества слов, имеющих противоположную оценку (метод 3).

Для точной оценки классификации было принято решение для столбца оценок первой выборки считать, что, если оценка текстов равна четырем или пяти баллам, то текст будет считаться позитивно окрашенным, иначе негативно окрашенным.

Для второй выборки было принято решение считать положительно окрашенными тексты, имеющие оценку от четырех до пяти и негативно окрашенными от единицы до трех.

Полученные результаты классификации тональности текстовых документов были оценены при помощи показателей точности (accuracy), точности (precision), полноты (recall), f1-меры.

Для словаря КартаСловСент для первой выборки средний результат точности классификации тональности при помощи всех методов составил 0,78, а для словаря LinisCrowd — 0,61. Для второй выборки средний результат точности классификации тональности при помощи всех методов для словаря КартаСловСент составил 0,95, а для словаря LinisCrowd — 0,88.

Сравнивая полученные оценки качества классификации тональности текстовых документов, можно сделать вывод о том, что словарь КартаСловСент показал лучшие результаты на обеих выборках, чем словарь LinisCrowd.

Для анализа эффективности работы тезаурусного подхода для классификации тональности текстов был реализован подход с использованием машинного обучения, включающий в себя метод опорных векторов и метод логистической регрессии.

Перед обучением моделей машинного обучения было необходимо провести такую же предобработку текстовых данных, как и для словарного подхода.

Далее необходимо было применить к данным модель “мешок слов”. Данный метод используется для извлечения объектов из текстов для дальнейшего их использования в методах машинного обучения.

Следующим этапом после применения модели “мешок слов” является разбиение датасета на обучающую и тестовую выборку в соотношении 70 к 30, соответственно. Затем происходит обучение модели опорных векторов логистической регрессии и оценка точности классификации.

Для метода опорных векторов для первой выборки средний результат точности классификации тональности составил 0,64, а для логистической регрессии — 0,60. Для второй выборки средний результат точности классификации тональности при помощи всех методов для метода опорных векторов — 0,59, а для логистической регрессии — 0,593.

Согласно полученным результатам можно сделать вывод о том, что классификация текстовых документов при помощи стандартных алгоритмов машинного обучения работает хуже, в отличие от методов тезаурусного подхода. Это может быть связано с качеством обучающих выборок и недостаточной точностью настройки моделей машинного обучения.

При использовании машинного обучения к анализу тональной информации охватывается больший спектр слов для исследования, в то же время более точную оценку слов содержит в себе тезаурусный подход.

ЗАКЛЮЧЕНИЕ

В настоящее время анализ настроений стал важным инструментом в современном информационном обществе из-за обилия текстовой информации в Интернет. Анализ поведения, эмоций и намерений людей играют существенную роль жизни как различных организаций, так и в жизни других людей. Однако фильтрация такого объема информации без использования специализированных средств является сложной задачей.

Существуют различные подходы к анализу тональности, одними из которых являются тезаурусный подход и подход, основанный на машинном обучении.

В рамках данной выпускной квалификационной работы были выполнены следующие задачи.

Был проведен сравнительный анализ подходов к решению задачи определения тональности.

Реализованы методы классификации настроений для текстовых данных при помощи тезаурусного подхода и машинного обучения. Для подхода, основанного на лексике были использованы словари КартаСловСент и LinisCrowd, созданные при помощи метода краудсорсинга. Для реализации подхода, основанного на машинном обучении были использованы методы опорных векторов и логистической регрессии.

Было разработано приложение для реализации классификации тональности текстовых данных при помощи тезаурусного подхода.

Была произведена оценка качества полученных моделей оценки тональности. В результате сравнения подходов с помощью различных метрик для оценки качества модели было выявлено, что при использовании тезаурусного подхода словарь КартаСловСент получил лучшие результаты, в отличие от словаря LinisCrowd. Это может быть связано с тем, что в словаре КартаСловСент содержится большее количество оцененных слов, данное число в шесть раз превышает количество слов в словаре LinisCrowd. Для повышения качества работы словаря LinisCrowd рекомендуется расширить количество слов в словаре, путем добавления в него синонимов и антонимов, а также последующей их оценке.

В результате анализа результатов классификации тезаурусного подхода и машинного обучения можно сделать вывод о том, что более точным мето-

дом классификации тональности текстов является словарный метод, так как оценка слов, содержащихся в нем является более точной, чем оценки, полученные при помощи методов машинного обучения.

Таким образом, все поставленные задачи были выполнены, цель выпускной квалификационной работы достигнута.

Основные источники информации:

1 Лукашевич, Н. В. Автоматический анализ тональности текстов: проблемы и методы // Интеллектуальные системы. Теория и приложения. 2022. Т. 26, № 1. С. 50-61.

2 How To Determine the Sentiment Score [электронная ссылка] URL: https://www.determ.com/blog/how-to-determine-the-sentiment-score/ (дата обращения 20.04.2023)

3 Choosing Performance Metrics [электронная ссылка]URL: https://towardsdatascience.com/choosing-performance-metrics-61b40819eae1 (дата обращения 20.04.2023)

4 Котельников, Е.В., Разова, Е. Современные словари оценочной лексики для анализа мнений на русском и английском языках // Научно-техническая информация. 2020. №2. С 16-33.

5 Choosing Performance Metrics [электронная ссылка]URL: https://towardsdatascience.com/choosing-performance-metrics-61b40819eae1 (дата обращения 20.04.2023)

6 KAGGLE [Электронный ресурс] URL: https://www.kaggle.com/datasets/theovall/phonereviews (дата обращения: 02.02.2023)

7 Тональный словарь русского языка КартаСловСент [Электронный ресурс] URL: https://github.com/dkulagin/kartaslov/tree/master/dataset/kartaslovsent (дата обращения: 30.03.2023)

8 Linis Crowd [Электронный ресурс] URL: http://www.linis-crowd.org (дата обращения: 30.03.2023)