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

РАЗРАБОТКА WEB-ПРИЛОЖЕНИЯ ДЛЯ КЛАССИФИКАЦИИ И АНАЛИЗА ТОНАЛЬНОСТИ СООБЩЕНИЙ НОВОСТНОЙ ЛЕНТЫ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 271 группы
направления 09.04.01 — Информатика и вычислительная техника
факультета КНиИТ
Суховой Надежды Валентиновны

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

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

Саратов 2023

ВВЕДЕНИЕ

Анализ информационных потоков помогает быстро вычленить из огромных потоков данных эмоциональную оценку, которая называется тональностью. Необходимость в данной технологии в настоящее время только растёт с ростом доверия к информации из сети в виде отзывов, комментариев и т. п.

Вместе с тем наблюдается недостаток исследований, связывающих анализ тональности новостных сообщений и социально-экономическую оценку регионального развития. Применение математических моделей и методов может быть полезным для получения внешнего представления о положении дел по различным территориям для быстрого реагирования на изменения ситуации там. Технологии текстового анализа позволят выявить регионы, оцениваемые в положительном контексте, оценить и внедрить используемые в них практики.

Объектом магистерской работы является исследование методов машинного обучения для задачи анализа тональности. Предметом магистерской работы является оценка тональности новостной ленты.

Целью магистерской работы является разработка приложения для анализа тональности.

В магистерской работе поставлены следующие задачи:

— обзор существующих методов анализа тональности;
— сбор и обработка данных новостной ленты в качестве апробации приложения;
— разработка модулей для реализации анализа тональности;
— разработка модуля скрапинга.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Магистерская работа состоит из четырёх разделов. В первом разделе описываются подходы и алгоритмы анализа тональности текстовых документов. Во втором разделе рассматриваются обучающие выборки и описывается работа с моделями классификации тональности новостных сообщений. В третьем разделе описывается разработка интерфейса приложения для анализа тональности новостных лент. В четвёртом разделе описывается апробация приложения и рассматриваются выводы, сделанные с учётом результатов работы приложения. В работе имеется 13 рисунков, 4 таблицы и 34 использованных источника.

В первом разделе магистерской работы был описан обзор методов анализа тональности текстовых документов.

Анализ тональности — это тип обработки текста для отслеживания в нем тональности, т. е. эмоционально окрашенной лексики. Этот анализ включает в себя классификацию и сбор мнений касательно какой-либо тематики. Для автоматизации данного анализа в частности используется машинное обучение.

Оценки тональности могут разделяться на несколько вариантов. Например:

— позитивная;
— негативная;
— нейтральная.

Каждая из оценок отображает процент эмоциональной окраски в тексте, где нейтральная оценка отвечает за отсутствие в тексте эмоциональной окраски. При этом в некоторых случаях текст может классифицироваться по большему числу оценок, либо только по двум: позитивной и негативной, чего недостаточно для полноценного анализа текста.

В основном перед анализом тональности ставятся следующие задачи:

— анализ тональности текста отзывов на какой-либо объект или процесс;
— определение эмоциональной окраски комментариев в новостных лентах.

Все методы анализа тональности можно разделить на три условные категории:

— методы, основанные на наборе правил;
— методы, основанные на машинном обучении;

— гибридные методы.

Первая категория — это метод, который основывается на наборе правил (rule-based). Обычно в данном методе используется некий словарь эмоциональных слов, по совокупности которых делается вывод об эмоциональной окраске текста. Такой метод очень эффективен в анализе тестов с общей тематикой, но считается трудоёмким для создания, особенно когда не находится подходящего словаря, и приходится создавать его вручную.Одной из главных особенностей методов анализа, основанных на машинном обучении, является автоматическое извлечение признаков присутствия той или иной эмоциональной окраски из текста. Третья категория — гибридные подходы, которые сочетают в себе особенности двух ранее описанных методов.

В настоящее время существует много сервисов и приложений, которые занимаются анализом тональности. Наиболее популярными сервисами являются:

— Awario;
— Brandwatch;
— Talkwalker;
— Lexalytics;
— Hootsuite Insights.

Из рассмотренного спектра уже существующих решений предлагается использовать методы машинного обучения в целях разработки специализированного приложения для облегчения обработки новостных массивов. Данное приложение предоставит пользователю результаты анализа новостных текстов для их дальнейшего использования при принятии решений о социально-экономическом состоянии регионов по их новостной ленте.

Во втором разделе описан процесс реализации обучения моделей для анализа тональности. Для выполнения поставленных задач и изучения имеющихся технологий, был выбран язык Python. У Python можно выделить следующие преимущества:

— Python лёгок в изучении;
— открытый код, позволяет реализовать как backend так и интерфейс программы.

В данной работе используется метод обучения моделей на тренировочных данных с последующим использованием обученных моделей для прогно-

зирования тональности текста. Были применены следующие методы:

— логистическая регрессия;

— стохастический градиентный спуск;

— дерево принятия решений;

— метод случайного леса.

Для использования и обучения данных моделей использовалась универсальная библиотека `scikit-learn`, которая помогла в создании модулей приложения для анализа. Каждый модуль реализует анализ с помощью одной из вышеуказанных моделей, обучая модель и анализируя с её помощью текст, указанный в качестве входных данных модуля. Выбор данных методов осуществлялся исходя из особенностей выборки, имеющей достаточно ограниченный объем.

Для данных методов анализа был проведен эксперимент с подбором наилучших параметров для модели. Эксперимент предполагал провести с помощью цикла подбор параметров для различных методов таким образом, чтобы достигнуть наилучшей возможной оценки точности работы методов. С этими параметрами модель проходит обучение на данных, полученных из системы организации конкурсов по исследованию данных[1], в которых находилось около 8000 новостных текстов с результатом тональности.

Для анализа новостных лент был разработан модуль скрапинга для сбора новостных сообщений. Скрапинг — средство автоматизированного извлечения данных с веб-страницы. Благодаря удобству работы с тегами, использование библиотеки BeautifulSoup будет целесообразно в целях удобства написания скрапинга.

Для корректной работы моделей необходим набор данных, состоящий из текстов новостных лент, которые модели будут анализировать. Этот набор данных формируется путём парсинга новостной ленты информационного агентства «РИА Новости», которое было выбрано по ряду причин. Полное наименование — Федеральное государственное унитарное предприятие «Российское агентство международной информации РИА Новости». Данная медиагруппа является крупнейшим государственным информационным агентством, а ria.ru является одним из крупнейших новостных ресурсов Европы и ведущим новостным сайтом страны. Сайт подходит для выполнения по-

[1] https://www.kaggle.com/competitions/sentiment-analysis-in-russian/data

ставленных задач, поскольку новости разделяются по регионам, что упрощает процесс парсинга. Также сайт предоставляет возможность просматривать опубликованные за конкретный день новости.

Информационное агентство предоставляет пользователям возможность искать новости по дате. Данный архив новостей находится в свободном доступе и представляет из себя страницу с перечнем ссылок на опубликованные новости. Таким образом, модуль для веб-скрапинга должен позволять совершать следующие действия:

— получать на вход временной промежуток;
— извлекать новостные тексты за данный промежуток.

В третьем разделе описывается разработка интерфейса приложения для анализа тональности новостных лент. Функционал приложения можно представить в следующем списке:

— получение от пользователя данных для формирования выборки;
— формирование выборки новостных текстов через парсинг новостного источника;
— анализ сохраненных новостных текстов с помощью выбранных пользователем методов;
— сохранение результата работы приложения по желанию пользователя.

Преимуществом разработанного приложения является решение конкретной задачи обработки новостных сообщений, включая парсинг новостной ленты и применение методов машинной обработки данных. Код интерфейса был создан с помощью библиотеки `streamlit`. Данная библиотека создана с помощью языка Python, её особенность в том, что она значительно упрощает работу с интерфейсной частью приложения. Эта библиотека помогает развернуть приложение, используя написанный код, отображая изменения в интерфейсе в реальном времени. Помимо библиотеки `streamlit` так же используются библиотеки `pandas`, `transliterate`, а также модули `Counter` и `Path`.

`Counter` — класс, предназначенный для быстрого подсчета повторяющихся элементов в последовательности, который формирует переменную со словарем, сохраняет повторяющиеся элементы как ключи словаря и количество повторяющихся элементов как значения. Модуль `Path` помогает в работе с путями файлов. В данном случае этот модуль поможет в процессе сохранения CSV-файла с результатом работы приложения.

В результате работы пользователь видит таблицу, содержащую в себе новостные тексты и их тональность; гистограмму, отображающую количественное соотношение новостей разных тональностей; список соотношения новостей, отображенного в процентах и точность работы метода, сокращенную для удобства отображения до трёх знаков после запятой.

Архитектурный шаблон — это архитектурная конструкция, которая ориентирована на решение проблемы, поставленной в процессе разработки. Учитывая концепцию приложения и техническое задание, можно заключить, что к данной работе больше всего подходит, так называемый, шаблон посредника. В этом случае приложение играет роль посредника, получающего запрос от клиента и обращающегося за выполнением этого запроса к подходящей службе из реестра. В данном случае пользователь использует приложение как посредника, который производит парсинг и анализ данных, взятых со стороннего ресурса.

В четвёртом разделе описывается апробация приложения. Вычислительный эксперимент заключался в сравнении нескольких регионов, основываясь на их новостной ленте за 2020, 2021 и 2022 годы.

Были выбраны следующие регионы:

— Саратов;
— Самара;
— Казань;
— Волгоград.

Количество процентов позитивно окрашенных новостных текстов по регионам и годам отображено в таблице 1.

Количество процентов негативно окрашенных новостей по регионам и годам отображено в таблице 2.

На основе результатов анализа можно сказать, что 2021 год выдался контрпродуктивным для всех выбранных регионов, однако ближе к 2022 ситуация стабилизировалась, об этом свидетельствуют таблицы 1 и 2. Регионом с наилучшими показателями считается Казань, так как на момент 2022 года в этом регионе опубликовалось больше новостей с позитивным окрасом и меньше с негативным, чем в остальных регионах.

Таблица 1 – Оценка позитивных новостей по регионам и временному периоду

Период / Регион	2020 г.	2021 г.	2022 г.
Саратов	18.7%	9%	8.9%
Самара	18.8%	11.6%	19.9%
Казань	16.6%	14.1%	29.6%
Волгоград	16%	9.4%	16%

Таблица 2 – Оценка негативных новостей по регионам и временному периоду

Период / Регион	2020 г.	2021 г.	2022 г.
Саратов	27.4%	38.6%	40.1%
Самара	18.3%	38.7%	30.4%
Казань	13.7%	23.7%	15.7%
Волгоград	29.1%	32.5%	24.9%

ЗАКЛЮЧЕНИЕ

В ходе магистерской работы был осуществлён анализ современного состояния исследований тональности текстов, был проведён обзор существующих технологий для анализа тональности, а также инструментов анализа тональности для аналитиков маркетинговых компаний.

В процессе обзора и анализа существующих технологий были сделаны выводы, что несмотря на существование инструментов для анализа тональности новостных лент, их функционал не позволяет разделять новостные посты по регионам, делая при этом выводы о ситуации в различных регионах.

Исследование проводилось с использованием языка программирования Python, с помощью которого проведён анализ тональности новостей регионов и отображение результатов в виде статистических данных.

В магистерской работе были разработаны модули для реализации четырех возможных методов анализа тональности. Каждый модуль является опциональным при выборе пользователем методов анализа.

Разработан модуль скрапинга для приложения. С помощью разделения работы цикла модуля на потоки, время работы приложения уменьшилось, что в общем позитивно сказалось на работе приложения.

Был разработан интерфейс для приложения для анализа тональности. С помощью этого интерфейса пользователь формирует выборку новостей и анализирует её с помощью выбранных методов.

Проведена апробация приложения с последующим сравнением состояния регионов в течение трёх лет. Для наглядности были предоставлены гистограммы и таблицы с указанными процентными соотношениями новостей. На основе оценок тональности были сделаны выводы о социально-экономическом состоянии регионов за три года. Полученные результаты позволяют утверждать, что из выбранных регионов (Саратов, Самара, Казань и Волгоград) наиболее успешным за период от 2020 до 2022 года считается Казань.

Данное исследование было представлено в рамках тринадцатой научно-практической конференции «Presenting Academic Achievements to the World» (Саратов, «Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского», Факультет иностранных языков и лингводидактики, 11.04.2022).

Апробация данного исследования была проведена в рамках студенче-

ской научной конференции факультета КНиИТ ««Компьютерные науки и информационные технологии» (Саратов, «Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского», Факультет компьютерных наук и информационных технологий, 04.05.2023).

Основные источники информации:

2 Сметанин, С. И. The applications of sentiment analysis for russian language texts: Current challenge and future perspectives / С. И. Сметанин // IEEE Access. - 2020. - Vol. 8. - Pp. 110693–110704

3 Павлов, Ю. Н. Сравнение методов оценки тональности текста / Ю. Н. Павлов // Молодой учёный. - 2016. - Vol. 12. - P. 59.

6 The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives [Электронный ресурс]. - URL: https://ieeexplore.ieee.org/document/9117010 (Дата обращения 13.12.2021). Загл. с экр. Яз. англ.

17 Samuels, A. News Sentiment Analysis / A. Samuels, J. Mcgonical. — California: University of Southern California, Caltech, 2020.

22 Getting Started [Электронный ресурс]. - URL: https://scikit-learn.org/stable/getting_started.html (Дата обращения 05.10.2022). Загл. с экр. Яз. англ.

23 Hackeling, G. Stochastic Gradient Descent in Theory and Practice / G. Hackeling. - Великобритания: Packt Publishing, 2014.

25 Спиринцев, В. В. Анализ современных архитектурных шаблонов, используемых при проектировании приложений в среде ios / В. В. Спиринцев, Н. А. Шитик // Вестник Херсонского национального технического университета. - 2016. - Vol. 3. - P. 58.

29 Конова, П. С. Библиотека streamlit как инструмент обработки и визуализации больших данных / П. С. Конова // Столыпинский вестник. - 2022. - Vol. 3. - Pp. 665–667.

30 Streamlit library [Электронный ресурс]. - URL: https://docs.streamlit.io/library/api-reference/text/st.markdown (Дата обращения 21.04.2023). Загл. с экр. Яз. англ.