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 РЕАЛИЗАЦИИ МЕТОДОВ АНАЛИЗА ТОНАЛЬНОСТИ НОВОСТНЫХ ЛЕНТ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Абакумовой Натальи Владимировны

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

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

Саратов 2023

ВВЕДЕНИЕ

Text Mining (интеллектуальный анализ текста) представляет собой процесс преобразования неструктурированного текста в структурированный формат для выявления значимых закономерностей и использования методов машинного обучения для дальнейшего применения в прикладных задачах. Различные методы machine learning, такие как наивный байесовский метод, метод опорных векторов и другие алгоритмы глубокого обучения, позволят исследовать и обнаруживать скрытые взаимосвязи в неструктурированных данных.

Интеллектуальный анализ текста занимается решением такой задачи, как анализ тональности текста (Sentiment Analysis). Анализ тональности позволяет выяснить, является ли эмоциональный тон сообщения положительным, отрицательным или нейтральным. Существует несколько подходов к решению данной задачи, в частности, создание системы на основе правил, которая определяет, классифицирует и оценивает конкретные ключевые слова на основе заранее заданных лексиконов. Однако словарь этих лексиконов необходимо постоянно пополнять, и может возникнуть неточность в обработке предложений, имеющих разные значения для разных культур. Другой подход заключается в использовании методов машинного обучения и алгоритмов классификации. В ряде случаев это обеспечивает большую точность модели, если обучить ее на большом количестве примеров. Существует третий гибридный подход, который использует функции двух предыдущих для оптимизации скорости получения и точности прогнозов, но требует больших технических ресурсов.

Анализ тональности можно применить для новостных сообщений, чтобы повысить объективность оценки эмоциональной окраски текстов из информационных источников. Однако можно столкнуться со следующей проблемой: новостные сообщения чаще всего подаются в нейтральной тональности, что затрудняет их классификацию. Актуальность данной работы связана с предлагаемым решением данной проблемы путем предобработки обучающего массива данных и подбора алгоритмов машинного обучения.

Целью бакалаврской работы является создание web-приложения, позволяющего оценить тональность сообщений новостных лент.

Для выполнения данной цели были поставлены следующие задачи:

— анализ подходов к Sentiment Analysis для оценки тональности сообще-

ний;

- сравнительный анализ существующих обучающих выборок для Sentiment Analysis;
- формирование обучающей выборки с учетом особенности новостных сообщений;
- разработка web-приложения для оценки тональности новостных лент;
- применение web-приложения на примере текстовых сообщений на новостную тематику.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Объектом бакалаврской работы является исследование методов машинного обучения для задачи Sentiment Analysis. Предметом бакалаврской работы является оценка тональности новостных сообщений. Бакалаврская работа состоит из трех разделов.

В первой главе описываются подходы к Sentiment Analysis и алгоритмы машинного обучения, которые применяются к решению данной задачи. Извлечение полезной информации из текста с помощью различных типов статистических алгоритмов называется интеллектуальным анализом текста, текстовой аналитикой, машинным обучением на основе текста. Текстовая аналитика становится все более популярной в последние годы из-за повсеместного распространения текстовых данных в Интернете, социальных сетях, электронной почте, цифровых библиотеках и чат-сайтах [1]. Примерами задач интеллектуального анализа текста являются: классификация текстов по настроению или тематикам, обнаружение намерений, извлечение информации.

Интеллектуальный анализ текста — это область, которая объединяет множество технологий, включающих широкий спектр возможностей. В практических приложениях обычно необходимо объединить несколько связанных технологий для выполнения прикладной задачи, а выполнение технологии интеллектуального анализа данных обычно скрыто за прикладной системой. Таким образом, интеллектуальный анализ текста — это не единая технологическая система, а обычно комплексное применение нескольких технологий. Существуют базовые и более сложные технологии, также они называются методами анализа текста, каждый из которых используется для разных целей [2]. Рассмотрим некоторые из них:

Классификация текста — это специфическое применение технологии классификации шаблонов. Задача метода состоит в том, чтобы разделить текст на предопределенные типы текста. Он считается одним из самых полезных методов обработки естественного языка, потому что он универсален и может организовывать, структурировать и классифицировать практически любую форму текста для предоставления значимых данных и решения проблем. Обработка естественного языка (NLP) — это метод машинного обучения, который позволяет компьютерам разбивать и понимать текст так же, как это сделал бы человек.

Одной из распространенных задач классификации является анализ тональности — это процесс анализа цифрового текста для определения того, является ли эмоциональный тон сообщения положительным, отрицательным или нейтральным.

Частотный анализ — это метод анализа текста, который измеряет наиболее часто встречающиеся слова или понятия в данном тексте с использованием числовой статистики TF-IDF (термин частотно-обратная частота документа) [3].

Чтобы автоматически анализировать текст с помощью машинного обучения, необходимо упорядочить и подготовить данные. Автоматический анализ текста использует ряд методов обработки естественного языка, подобных приведенным ниже.

Токенизация — это процесс разбиения строки символов на семантически значимые части (токены), которые можно анализировать (например, слова), при отбрасывании бессмысленных фрагментов (например, пробелов) [4].

Основополагание и лемматизация относятся к процессу удаления всех аффиксов (суффиксов, префиксов и т. д.), прикрепленных к слову, чтобы сохранить его лексическую основу, также известную как корень или основа, или его словарная форма, или лемма. Основное различие между этими двумя процессами заключается в том, что основополагание обычно основано на правилах, обрезающих начало и окончание слов, что приводит к неверному результату, тогда как лемматизация использует словари и гораздо более сложный морфологический анализ.

Чтобы обеспечить более точный автоматический анализ текста, нужно удалить слова, которые несут очень мало семантической информации или вообще не несут смысла. Эти слова также известны как стоп-слова [5].

Текст является одним из наиболее распространенных типов неструктурированных данных. Из-за беспорядочной природы текста анализ, понимание, организация и сортировка текстовых данных сложны и требуют много времени. В таких случаях применяется классификация текстов с помощью машинного обучения.

Автоматическая классификация текста применяет машинное обучение, обработку естественного языка и другие методы под управлением искусственного интеллекта для автоматической классификации текста более быстрым,

экономичным и точным способом. Данный подход является одним из подходов к решению задачи анализа тональности текстов. Анализ тональности на основе машинного обучения полезен тем, что он точно обрабатывает широкий спектр текстовой информации. Пока программное обеспечение проходит обучение с достаточным количеством примеров, анализ тональности машинного обучения может точно предсказать эмоциональный тон сообщений. Однако обученная модель может делать предсказания только для текстов из той же сферы, на которых она была обучена [6].

Для анализа тональности новостных текстовых сообщений предпочтительнее будет применить систему на основе методов машинного обучения при наличии обучающего набора данных и обширного списка алгоритмов классификации, которых можно применить к обработанному тексту.

Во второй главе рассматриваются обучающие выборки, описывается работа с моделями классификации тональности новостных сообщений.

Набор данных с российскими новостями [7] имеет 4 поля: нумерация строк, text, в котором хранятся тексты новостей, id, sentiment – столбец, в котором хранятся метки классов, к которому принадлежат новости. Общее количество 8256 записей.

У вышеописанного набора данных есть недостатки. Во-первых, небольшое количество записей, можно сразу сказать, что обучение не будет качественным, так как для каждого из трех классов в данных наборах небольшое количество образцов. Вторая проблема появляется из первой, ибо в каждом из классов несбалансированное количество записей, в датасете с новостями очень мало образцов с негативной окраской.

Обучающую выборку, состоящую из текстовых записей, нельзя сразу отправить на обработку алгоритмом классификации. Перед обучением датасет должен пройти предобработку, состоящую из нескольких этапов. Сначала нужно привести весь текст в один регистр, убрать все знаки препинания и специальные символы, также необходимо вернуть каждому слову его начальную форму. После этих действий надо применить метод TF-IDF, который переведет все слова в числовые значения. После получения этих числовых значений можно строить модель и обучать ее на обработанной выборке. В конце нужно оценить качество модели на основе изначальных меток и меток, предсказанных обученной моделью [8].

Модели обученные на данной выборке показали низкую точность в определении негативных и позитивных записей, поэтому было принято решение дополнить выборку. Чтобы дополнить готовую выборку был примене парсинг информационного источника Лента.ру. После получения массива данных он был размечен по тональности на три класса: позитивная новость, помечалась как positive, нейтральная новость, помечалась как neutral, негативная новость, помечалась как negative. Разметка проводилась вручную. Изначальный датасет получилось дополнить 1067 записями.

В следующей таблице показано отличие оценки precision у старой модели, обученной на готовой выборке с сайта Kaggle.com, и у модели, которая обучена на выборке, полученной в результате объединения готового набора данных и созданного собственноручно. Обе модели были обучены алгоритмом случайного леса. В таблице 1 представлена разница оценок precision у вышеуказанных выборок для разных классов.

Таблица 1 – Оценка precision для моделей, обученных на рассматриваемых выборках

	Готовая выборка	Объединенная выборка
negative	0.64	0.73
neutral	0.67	0.66
positive	0.66	0.70

На модернезированной выборке точность повышается. Полученная выборка по совокупности оценок точности явно предпочтительнее, чем остальные варианты обучающих выборок. Вероятно, повышение точности связано с увеличением обучающих примеров в классах, обозначенных как positive и negative.

В третьей главе разрабатывается интерфейс приложения для анализа тональности новостных лент и производится его апробация.

Разработанное приложение для анализа тональности новостной ленты имеет следующий функционал:

— загрузка массива данных в формате .csv с двумя столбцами: Text, в котором хранится текстовое сообщение, ID, в котором хранится дата создания сообщения в формате число:месяц:год;

— классификация текстовых сообщений по настроению на три класса: ней-

тральное, позитивное, негативное. Есть возможность выбора одной из четырех моделей, обученных следующими методами машинного обучения: случайный лес, логистическая регрессия, наивный байесовский классификатор и KNN;

— вывод диаграммы, показывающей количество сообщений принадлежащих разным классам в разные даты;
— вывод круговой диаграммы, показывающей процентное соотношение количества записей в разных классах от общего числа.

На рисунке 1 представлен интерфейс приложения во время работы. Графики появляются в отдельных окнах после нажатия соответсвующих кнопок.

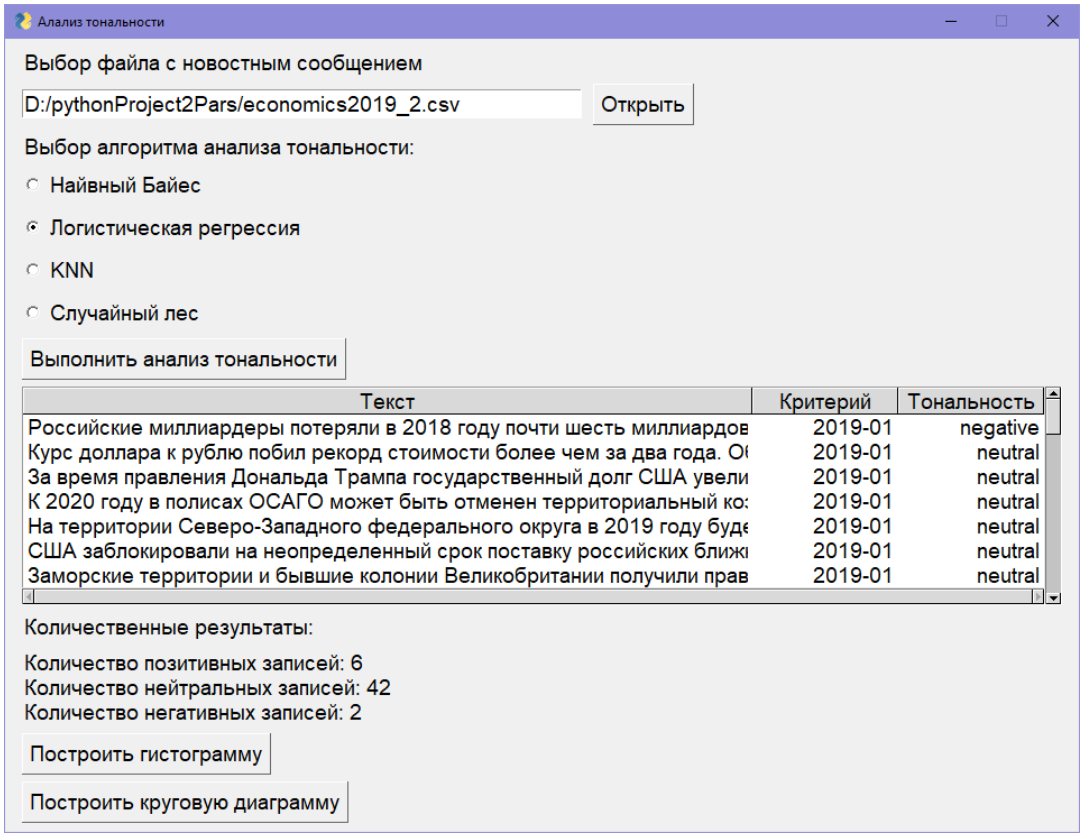

Рисунок 1 – Скриншот приложения во время выполнения анализа тональности данных

В качестве апробации был снова применен парсинг новостей информационного источника Лента.ру. Были собраны новости в период с 2019 по 2021 год из рубрики «Экономика». В таблице 2 собраны данные по этим трем годам, количественные значения для каждого из класса тональности, а также процентное соотношение классов к общему числу данных за год.

На основе данных из таблицы можно сделать вывод, что при общем

Таблица 2 – Результаты анализа тональности новостной ленты за период с 2019 по 2021 год

Тональность	Позитивная		Нейтральная		Негативная	
Год	Кол-во	%	Кол-во	%	Кол-во	%
2019	874	14.4	4979	82.3	196	3.2
2020	888	11.9	6357	85.2	213	2.9
2021	1057	12.8	7068	85.5	138	1.7

увеличении количества новостей, в процентном соотношении количество нейтральных 82-85%, позитивных стало меньше в 2020, но через год результат вырос на 1% и в общем итоге 12-14%, негативных новостей с каждым годом все меньше, примерно 2 – 3 %.

ЗАКЛЮЧЕНИЕ

В бакалаврской работе были выявлены особенности работы методов и алгоритмов Sentiment Analysis для анализа тональности новостных сообщений. Для подобных корпусов текстов характерно представление без ярко выраженной эмоциональной окраски, что затрудняет применение стандартных методик оценки тональности. По результатам вычислительного эксперимента большинство сообщений в выборке определяются как нейтральные.

В работе была применена практика перевода иностранных обучающих выборок на русский язык. Из-за особенностей подачи материала в других культурах, модели, обученные на этих выборках хорошо давали оценку тональности только для новостей из своей страны, на российской новостной ленте был показан недостаточно точный результат. По этой причине был применен другой подход к обучению на российских выборках. Для обновления словаря терминов и расширения примеров в классах с негативными и позитивными записями был использован парсинг информационного источника Лента.ру. В результате был получен массив новостей, которые в дальнейшем были размечены и добавлены в модифицированную обучающую выборку. Модели, обученные на обновленной выборке, показали большую точность.

В процессе работы было обучено 4 модели на основе разных алгоритмов классификации. Разные алгоритмы были выбраны для того, чтобы исследовать зависимость качества модели от предобработки данных и применения на них алгоритмов классификации. В итоге у полученных моделей точность незначительно отличается.

Было разработано приложения на языке Python 3.8 с использованием полученных моделей анализа тональности. Функциональные возможности приложения обеспечивают загружать корпус новостных сообщений, выбирать модель классификации, визуализировать результаты оценки тональности. В качестве апробации были использованы новостные данные из рубрики Экономика в период с 2019 по 2021 год из новостного источника Лента.ру. Модели показали стабильное снижение негативных новостей на фоне общего увеличения числа новостей за год.

Основные источники информации:

1. Machine Learning for Text/ Aggarwal C. — Luxembourg: Springer, 2018. — 516 c.
2. Clinical Text Mining/ Dalianis H. — Luxembourg: Springer, 2018. — 192 c.
3. Text Data Mining/ Chengqing Z., Rui X., Jiajun Z. — Beijing: Tsinghua University Press, 2021. — 363 c.
4. Data Science. Наука о данных с нуля/ Грас Дж. — СПб.: БХВ-Петербург, 2020. — 416 с.
5. Introduction to Information Retrieval/ Manning C., Raghavan P., Schutze H. — United Kingdom: Cambridge University Press, 2008. — 478 c.
6. Reis, J., Olmo, P., Benevenuto, F., Kwak, H., Prates, R., An, J. Breaking the news: first impressions matter on online news / Reis J., Olmo P., Benevenuto F., Kwak H., Prates R., An J. // Журн. ICWSM. 2015. № 15.
7. Sentiment Analysis in Russian [Электронный ресурс] URL: https://www.kaggle.com/competitions/sentiment-analysis-in-russian/data (дата обращения:05.01.2023) — Яз.англ.
8. Анализ данных и процессов/ Барсегян А. А., Куприянов М. С., Холод И. И., Тесс М. Д., Елизаров С. И. . — СПб.: БХВ-Петербург, 2009. — 512 с.: ил. + CD-ROM — (Учебная литература для вузов)