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 ПРИЛОЖЕНИЯ ДЛЯ АНАЛИЗА ЭКОНОМИЧЕСКИХ НОВОСТЕЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Тарана Евгения Вячеславовича

Научный руководитель
должность, степень, звание _____ Г. Ю. Чернышова

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

Саратов 2025

ВВЕДЕНИЕ

Актуальность темы

В условиях цифровой трансформации общества объемы неструктурированных текстовых данных, генерируемых в социальных сетях, новостных ресурсах и блогах, растут экспоненциально. Ручной анализ таких данных становится невозможным из-за их масштаба и разнообразия, что делает технологии автоматизированного парсинга и анализа тональности критически важными. Парсинг позволяет извлекать структурированную информацию из источников, таких как социальная сеть Вконтакте, а анализ тональности даёт возможность интерпретировать эмоциональную окраску текстов, выявляя общественные настроения, предпочтения пользователей или реакцию на события.

Особую актуальность эта тема приобретает для русскоязычного контента, где наблюдается дефицит качественных обучающих выборок для анализа тональности, особенно в узких областях, таких как экономика или политика. Интеграция парсинга с методами машинного обучения и словарными подходами позволяет не только собирать релевантные данные, но и создавать адаптивные инструменты для их интерпретации, что востребовано в маркетинге, управлении репутацией, социологических исследованиях и прогнозировании трендов.

Цель дипломной работы – разработка приложения, объединяющего парсинг текстовых данных из социальной сети Вконтакте с последующим анализом тональности.

Поставленная цель определила **следующие задачи**:

- анализ и классификация методов парсинга социальных сетей, выбор оптимального подхода для работы с Вконтакте;
- разработка модуля парсинга, формирующего текстовые выборки на основе пользовательских запросов;
- разработка модуля анализа тональности текстов с применением словарных методов;
- апробация приложения с целью анализа сформированных данных о региональных экономических новостях и интерпретация результатов.

Практическая значимость бакалаврской работы заключается в разработке полнофункционального приложения для сбора и анализа тональности новостных сообщений для различных периодов и различных объектов.

Бакалаврская работа состоит из введения, 4 разделов, заключения, списка использованных источников и 3 приложений. Общий объем работы – 80 страниц, из них 43 страниц – основное содержание, включая 13 рисунков и 1 таблицу, список использованных источников информации – 20 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе проведён детальный обзор существующих методов и инструментов парсинга, а также возможности работы с API социальных сетей.

Парсинг веб-сайтов — это процесс извлечения данных из веб-сайтов. Процесс парсинга веб-сайтов включает в себя отправку запросов на получение веб-страницы и извлечение из нее машиночитаемой информации.

В настоящее время существует несколько подходов к осуществлению парсинга. Одной из задач исследовательской работы стал обзор наиболее популярных методов парсинга и выбор наиболее продуктивного способа для использования его при создании приложения, осуществляющего сбор текстовых данных.

Существует большое количество способов парсинга сайтов. Под различные платформы, такие как Python, Java, Ruby, Java, PHP, JavaScript, существует множество средств, обеспечивающих реализацию парсинга. Для рассмотрения нескольких инструментов веб-скрапинга был выбран язык программирования Python, так как он широко распространен и содержит в себе множество библиотек, осуществляющих реализацию дальнейшей обработки данных, в частности Text Mining. Наиболее популярными инструментами для реализации веб-скрапинга на Python являются библиотека BeatifulSoup и фреймворк Scrapy [1].

Парсинг с использованием различных инструментов языков программирования и скрапинг при помощи средств автоматизации браузера могут быть достаточно эффективными методами сбора информации, эти способы предназначены для сбора информации с практически любых источников. Но существует инструмент парсинга данных, который применим лишь к некоторым информационным ресурсам, создатели которых по собственному желанию предоставляют наиболее простой и наиболее быстрый способ получения большого количества информации с интернет-ресурса. Этот инструмент парсинга называется Application Programming Interface(API).

Так как именно парсинг с использованием API является самым быстрым и удобным инструментом для извлечения данных, было принято решение разрабатывать приложение, которое будет реализовывать именно этот метод вебскрапинга. Поскольку целью работы является разработка инструментального средства для формирования текстовой выборки появилась задача поиска такого информационного ресурса, имеющего собственный API. который содержит в себе большое количество различных текстовых данных.

Источником, содержащим в себе большое количество различных текстовых данных, может являться социальная сеть. В силу того, что наиболее популярной русскоязычной социальной сетью является ВКонтакте, многие новостные ресурсы региональных официальных ведомств и СМИ активно ведут новостные сообщества на платформе этого ресурса. Социальная сеть ВКонтакте имеет свой собственный открытый и документируемый API, а также позволяет работать с широким охватом новостных источников, что и стало определяющим фактором в решении обращаться и работать именно с API ВКонтакте.

Во втором разделе рассмотрены классические алгоритмы анализа тональности, такие как деревья решений, метод k-ближайших соседей и случайный лес, а также их сравнение со словарными методами. При выборе методов анализа тональности была рассмотрена дилемма между классическими алгоритмами машинного обучения и словарными подходами.

Деревья решений находят широкое применение в задачах анализа тональности текстов благодаря своей прозрачности и относительной простоте интерпретации. Этот метод классификации особенно полезен при работе с текстовыми данными, где важно понимать, какие именно слова или сочетания слов влияют на итоговую эмоциональную оценку. Основной принцип работы деревьев решений заключается в последовательном разделении текстовых данных на группы по наиболее значимым признакам. В контексте анализа тональности такими признаками обычно выступают отдельные слова или их сочетания. Алгоритм начинает с корневого узла, где выбирается наиболее информативный признак, и затем рекурсивно разделяет данные на подмножества, пока не достигается заданный критерий остановки [2].

Случайный лес (Random Forest) представляет собой ансамблевый алгоритм машинного обучения, который использует комбинацию множества деревьев решений для повышения точности прогнозирования. В основе этого под-

хода лежит техника бутстрэп-агрегирования (bagging), позволяющая снизить эффект переобучения и увеличить устойчивость модели.

Метод k-ближайших соседей (k-Nearest Neighbours, k-NN) представляет собой простой, но действенный алгоритм машинного обучения, применимый для анализа тональности текстов. Основная идея метода заключается в предположении, что схожие по характеристикам объекты обычно принадлежат к одним и тем же классам [3].

Применительно к анализу тональности текстов метод k-ближайших соседей может показывать высокую эффективность в случаях, когда данные обладают сложной структурой и между текстами существуют выраженные тематические связи. Однако данный подход имеет ряд существенных ограничений, среди которых следует отметить высокую вычислительную сложность при работе с большими объемами данных и признаковыми пространствами высокой размерности.

Описанные методы k-ближайших соседей и случайного леса были реализованы для проведения вычислительного эксперимента оценки точности алгоритмов.

Словарные методы анализа тональности основываются на использовании предопределенных словарей, которые содержат слова и фразы с заранее определенной эмоциональной окраской (положительной, отрицательной или нейтральной). Преимущества словарных методов включают:

- простоту реализации и интерпретации результатов;
- отсутствие необходимости в обучающих данных;
- быстроту работы по сравнению с методами машинного обучения
- прозрачность процесса анализа (легко понять, какие слова повлияли на результат);
- возможность адаптации под конкретную предметную область путем дополнения словаря.

Словарные подходы к анализу тональности обладают рядом существенных ограничений при работе с экономическими новостными текстами. Основные проблемы включают: ограниченный охват специализированной экономической лексики, отсутствие учета контекста употребления слов (что критично для многозначных терминов), неспособность корректно обрабатывать идиоматические выражения и устойчивые словосочетания. Дополнительные сложности

создает динамичность новостного языка, требующая регулярного обновления словарных ресурсов. Большинство существующих словарей созданы для общего использования и не учитывают специфику экономической тематики, что существенно снижает качество анализа в данной предметной области [4].

Учитывая указанные особенности, для реализации анализа тональности в рамках бакалаврской работы был выбран словарный подход как методологически наиболее соответствующий поставленным задачам. Данный выбор обусловлен следующими объективными факторами: отсутствием необходимости в размеченных обучающих данных, низкими вычислительными требованиями, высокой степенью интерпретируемости результатов. Хотя словарные методы теоретически уступают в точности современным нейросетевым подходам, их преимущества в условиях ограниченных временных и вычислительных ресурсов, типичных для учебных исследований, делают их оптимальным выбором. Кроме того, модульная структура словарных подходов позволяет в дальнейшем легко комбинировать их с другими методами, создавая гибридные системы анализа тональности.

Таким образом, проведенный анализ стал причиной выбора словарных методов для решения поставленной задачи. Их применение позволяет получить научно значимые результаты при соблюдении требований к методологической строгости и в рамках доступных ресурсов, что особенно важно для учебно-исследовательской работы уровня бакалавриата.

Третий раздел посвящён разработке программного продукта, объединяющего модули парсинга и анализа тональности. Программный продукт разработан на Python с использованием модульной архитектуры, что упрощает его расширение. Графический интерфейс, реализованный на базе CustomTkinter, включает четыре функциональные вкладки: парсинг, управление API-токенами, анализ отдельных выборок и мониторинг динамики тональности.

Так как именно парсинг с использованием API является самым быстрым и удобным инструментом для извлечения данных, было принято решение разрабатывать приложение, которое будет реализовывать именно этот метод веб-скрапинга посредством работы с API социальной сети Вконтакте [5].

У Вконтакте есть собственный API (VKontakte API), созданный для создания Standalone-приложений и приложений внутри самой соцсети. С помощью VK API можно получать доступ к данным пользователя, сообщества или группы,

а также выполнять различные действия, такие как публикация постов, отправка сообщений, получение информации о друзьях и т.д..

Методы работы с VK API по сути реализуют ряд операций с базами данных. Эти методы разделены на секции в зависимости от того, в какой сфере и к чему метод применяются.

Парсер и другие модули приложения были реализованы на языке Python 3.9 с использованием интегрированной среды разработки PyCharm Professional Edition. Для разработки графического интерфейса была выбрана библиотека Tkinter и пакет CustomTkinter. Tkinter представляет собой стандартную библиотеку Python для создания графических интерфейсов, основанную на кроссплатформенной библиотеке Tk, изначально созданной для языка Tcl, но затем адаптированной для Python и других языков программирования. CustomTkinter расширяет функциональность стандартной библиотеки, предлагая дополнительные виджеты и возможности кастомизации интерфейса, включая предустановленные стили оформления.

Основная функция интерфейса вкладки «Парсинг» активируется при нажатии кнопки "Запуск". Она выполняет сбор данных из всех полей ввода, преобразует даты в требуемый формат, блокирует элементы управления на время выполнения операции.

Разработанные элементы графического интерфейса и механизмы его взаимодействия с парсером предоставляют пользователю возможность настраивать параметры для классификации текстовых данных и инициировать процесс сбора информации из социальной сети Вконтакте.

Разработанный функционал для сбора новостных сообщений позволяет задать следующие параметры поиска: ключевые слова, временной период, ключевые слова для поиска сообществ, количество анализируемых сообществ по каждой теме, режим работы, минимальная длина текста.

Таким образом, функционал вкладки «Парсинг» позволяет после отработки одного цикла сбора и отбора данных сформировать срез данных. Срез данных — это подмножество данных, полученное из более крупного массива или набора данных путём выборки определённых элементов по заданным критериям. Критериями для формирования срезов в приложении являются ключевые слова, которые обязаны присутствовать в тексте, минимальное количество слов в тексте и временной промежуток, в который новостное сообщение должно быть

опубликовано.

Для того, чтобы начать пользоваться функциями сбора пользователю необходимо получить user_token(токен пользователя), который можно получить через OAuth-авторизацию. Для получения и ввода токена пользователя в приложении была реализована специальная вкладка.

Таким образом, модуль парсинга позволяет пользователю получить необходимый для работы приложения токен пользователя VK API, выбрать параметры поиска источников текстовых сообщений, установить критерии отбора текстов для финальной выборки и сохранить эту выборку в файл по указанному пути.

Одной из задач исследовательской работы является реализация анализа тональности текстов с применением словарных методов. Для достижения данной задачи был произведен поиск подходящих словарей тональности русских текстов. В качестве универсальных вариантов были выбраны словарь RuSentiLex и словарь KartaslovSent [6].

В ходе создания приложения было принято решение использовать бинарную классификацию текстов на позитивные тексты(1) и объединение негативных сообщений с нейтральными(0). Кроме того, в ходе применения словаря был выработан порог для классификации. Если средняя оценка больше 0.2, текст признается позитивным, иначе он считается нейтральным иди негативным.

Для того чтобы убедиться в точности классификации текстов с использованием рассматриваемых словарей, было принято решение разметить вручную новостные сообщения об экономике, собранные с помощью модуля парсинга. Была построена специальная тестовая выборка, содержащая 422 новостных сообщения, среди которых было признано позитивными 260 и негативными или нейтральными 162.

В целях проверки обоснованности использования словарей тональности в качестве инструмента для анализа тональности новостных сообщений на экономическую тематику, было осуществлено сравнение словарных методов с рассмотренными в работе классическими алгоритмами классификации на специальной тестовой выборке. Сравнение проводилось с помощью стандартных метрик оценки классификации текстов, таких как Accuracy, Precision, Recall и F1-score [7]. На основании значения метрик, полученных в ходе проверки точности словарей и классических алгоритмов обучения на тестовой выборке можно сде-

лать вывод о том, что использование словарей позволяет достичь сравнимую с классическими методами точность классификации. Особенная необходимость использования словарных методов для анализа тональности может проявляться в ситуациях отсутствия размеченных текстовых выборок.

Для интеграции в приложение функции тональной оценки была создана вкладка «Анализ среза», позволяющая указать путь к сгенерированному приложением .xlsx файлу, выбрать один или два словаря тональности с помощью которых будет проведен анализ и сгенерировать файл с размеченными данными и файл с результатами анализа тональности среза.

Файл с размеченными данными представляет собой промежуточный .xlsx файл, который в первом столбце содержит исходные тексты, во втором столбце предобработанные тексты, а в третьем столбце находиться тональная оценка, полученная применением одного из словарей.

Файл с результатами анализа тональности среза представляет собой .xlsx файл, в котором содержится информация о результатах оценки среза данных одним из словарей с указанием количества позитивных и негативных текстов и их процентного соотошения. После формирования файла с результатами анализа тональности генерируется круговая диаграмма, которая выводится в пользовательском интерфейсе и сохраняется в указанную пользователем папку.

Для добавления возможности сбора сводной статистики о результатах работы функции из вкладки «Анализ среза» была создана вкладка «Мониторинг срезов». На данной вкладке пользователю предлагается выбрать каталог, в котором могут содержаться .xlsx файлы или подкаталоги с .xlsx файлами, сгенерированными с помощью приложения. Важно, чтобы пользователь сохранял наименования файлов в том виде, в котором их генерирует приложение.

После выбора такого каталога и нажатия кнопки запуска в выбранную папку будет сохранен .xlsx файл, содержащий сводную информацию обо всех срезах, которые содержаться в иераркической структуре каталогов, что делает удобным дальнейший анализ и визуализацию полученных данных. Кроме того, в выбранной папке сформируется каталог, содержащий графики, характеризующие срезы данных по количеству собранных текстов в каждый временной период и процент текстов оцененных одним или двумя словарями как положительные.

Результатом работы функции вкладки «Мониторинг срезов» является файл формата .xlsx, в котором на каждый подкаталог иерархической файло-

вой структуры приходиться один Excel лист, на котором приведены результаты анализа тональности всех срезов данных из этого подкаталога.

Таким образом, модуль приложения, отвечающий за анализ собранных приложением срезов данных позволяет пользователю осуществить оценку тональности новостных сообщений, выбрать один или два словаря для проведения этой оценки, получить итоговую статистику и ее визуализацию по каждому срезу отдельно или для срезов объединенных в одну структуру каталогов.

В четвёртом разделе проведена апробация приложения на реальных данных. Была поставлена задача сбора, формирование статистики и анализа тональности годовых срезов экономических новостей по регионам Приволжского федерального округа за последние 5 лет(2020 - 2024). Было принято решение в качестве поискового запроса использовать название региона и осуществлять сбор из 5 новостных сообществ этого региона. Сбор данных осуществлялся с применением фильтрации по минимальном количеству слов в тексте (20) и большому перечню ключевых слов, которые должны присутствовать в отбираемом тексте. Ключевые слова были подобраны в соответствии с целями устойчивого развития ООН.

Результаты сбора новостей были собраны в структуру, представляющую собой каталог «ПФО»,в котором содержится 14 подкаталогов, носящие названия 14 регионов ПФО. В каждом подкаталоге содержится пять .xlsx файлов, содержащих новости конкретного региона за конкретный год(срезы).

Для анализа тональности .xlsx файлов собранных в структуру, описанную ранее было принято решение воспользоваться функциональной составляющей вкладки «Анализ среза» следующим образом: проводить анализ с помощью двух словарей, после чего генерируемые файлы разметки, итогов и графики сохранять в ту же папку, в которой храниться анализируемый файл.

В результате анализа графиков, построенных для всех 14 регионов ПФО, можно отметить некоторые общие тенденции по годовым срезам. Практически во всех регионах год от года растет количество новостей на экономическую тематику. Единственным исключением из этой тенденции является Саратовская область, где наблюдается обратная ситуация. Другой общей тенденцией является рост с 2020 по 2024 год доли новостей, оцениваемых как положительные. Исключением в этом тренде являются Кировская область, где рост доли положительных новостей не наблюдается.

ЗАКЛЮЧЕНИЕ

В бакалаврской работе были проанализированы различные подходы к организации парсинга. По результатам анализа способов веб-скрапинга в приложении был реалзиован модуль «Парсинг», осуществляющий сбор текстовых данных через обращения к API популярной социальной сети Вконтакте. Данный модуль позволяет собирать датасеты по различным источникам и тематикам, а также производить отбор текстов по ключевым словам и длине текстовых сообщений.

В бакалаврской работе были рассмотрены различные методы анализа тональности текстов, после чего было решено выбрать словарные методы в качестве реализуемых внутри приложения. В качестве словарей, выбранных для использования в приложении были выбраны русскоязычные словари оценки тональности KartaslovSent и RuSentiLex. Оценка точности этих словарей была проведена на размеченной вручную выборке с применением метрик оценки точности.

В приложении был реализован модуль, представленный вкладками «Анализ среза» и «Мониторинг срезов». Функциональная составляющая вкладки «Анализ среза» позволяет формировать статистические отчеты и графики, характеризующие тональную оценку текстов, в выбранном пользователем. Во вкладке «Мониторинг срезов» пользователю предоставляется возможность получить сводную статистику по сформированному им каталогу с подкаталогами, которые содержат сгенерированные в разрабатываемом приложении .xlsx файлы. Статистика предоставляется в виде одного итогового отчетного .xlsx файла по всем подкаталогам и диаграммам и итоговым .xlsx файлам, которые формируются по данным из каждого подкаталога.

В целях применения и апробации созданного приложения был реализован анализ тональности и формирование статистики о годовых срезах экономических новостей регионов Приволжского федерального округа за период с 2020 по 2024 год. В ходе сбора данных к текстам была применена фильтрация по минимальном количеству слов в тексте и большому перечню ключевых слов, которые должны присутствовать в отбираемом тексте. Собранный датасет и результаты его анализа могут быть использованы в процессе формирования рейтингов региональных субъектов.

Дальнейшим направлением исследования может стать усоворшенствование предобработки собранных данных, процесса подбора и использования ключевых слов для формирования срезов данных. Более точные результаты может дать применение для классификации текстов расширенных словарей и методов глубокого обучения.

Результаты бакалаврской работы были опубликованы в статье Chernyshova, G.; Taran, E.; Firsova, A.; Vavilina, A. Monitoring of Sustainable Development Trends: Text Mining in Regional Media. Sustainability 2025, 17, 3122. https://doi.org/10.3390/su17073122.

Основные источники информации:

1 Ryan, M., Web Scraping with Python: Collecting More Data from the Modern Web – USA: O'Reilly, 2018. – 300 с.

2 Сидоров, А.В., Алгоритмы создания дерева принятия решений [Электронный ресурс] URL: https://s.econf.rae.ru/pdf/2014/03/3245.pdf (дата обращения: 02.04.2024).

3 Метод k-ближайших соседей (K-nearest neighbor) [Электронный ресурс] URL: https://wiki.loginom.ru/articles/k-nearest-neighbor.html (дата обращения: 01.04.2024).

4 Федорова, Е., Рогов, О., Демин, И. Применение словарей тональности для текстового анализа. — Litres, 2022.

5 Документация VK API [Электронный ресурс] URL: https://dev.vk.com/method (дата обращения: 02.03.2025)

6 Лукашевич, Н. В. Создание лексикона оценочных слов русского языка РуСентиЛекс // (OSTIS-2016): материалы VI международной конференции с. 377–382.

7 Метрики в задачах машинного обучения [Электронный ресурс] URL: https://habr.com/ru/companies/ods/articles/328372/ (дата обращения: 06.01.2025)