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 ОЦЕНКИ УСТОЙЧИВОГО РАЗВИТИЯ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Шишкова Никиты Александровича

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

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

Саратов 2026

ВВЕДЕНИЕ

В условиях постоянного роста объемов медиаконтента мониторинг и анализ этого потока становятся практически неосуществимыми. Особую значимость данная задача приобретает применительно к узким тематическим направлениям, таким как проблематика целей устойчивого развития, где требуется не только выделить релевантные публикации из общего массива, но и оценить их количество. Это обусловливает необходимость разработки и внедрения автоматизированных методов извлечения информации из неструктурированных текстовых данных. Возникает объективная потребность в инструментарии, позволяющем агрегировать разрозненные новостные сообщения, идентифицировать скрытые тематические структуры и оценивать характер их освещения.

Целью дипломной работы является реализация усовершенствованной методики оценки социально-экономических показателей устойчивого развития регионов методами Text Mining.

Для достижения поставленной цели были выделены следующие задачи:

— предобработка исходной выборки, включающей новостные тексты;
— построение кластерных моделей для выявления документов в соответствии с заданной тематикой;
— анализ тональности сообщений;
— разработка полнофункционального web-приложения для кластеризации и анализа текстовых документов.

Объектами в дипломной работе являются методы Text Mining для тематической классификации. Предмет представляет из себя программную реализацию методов Text Mining для обработки новостных массивов.

Дипломная работа состоит из 55 страниц, 15 рисунков и 4 таблиц. В первом разделе рассматриваются возможности Text Mining для формирования тематических кластеров. Во втором разделе анализируются методы обработки данных, кластеризации и анализа тональности. В третьем разделе представлены проектирование и разработка приложения, приведен пример кластеризации новостных статей, на основе которой производится выделение тематических групп.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе рассматриваются задачи Text Mining [1], структурные особенности новостных статей и содержательная характеристика целей устойчивого развития.

В отличие от анализа произвольных текстов, классификация новостных сообщений по заранее заданным рубрикам требует учёта многозначности заголовков и возможности пересечения тем. Text Mining позволяет автоматизировать данный процесс путём построения векторных представлений текстов и моделей машинного обучения [2].

В условиях отсутствия априорной разметки, кластеризация позволяет группировать новостные сообщения по смысловой близости. Данный подход применяется для обнаружения событийных кластеров, мониторинга развития информационных поводов во времени и выявления дублирующегося контента из различных источников.

Выделение наиболее значимых терминов, персоналий и организаций позволяет формировать семантические портреты новостных сюжетов. Данный функционал используется для автоматической генерации тегов, аннотирования публикаций и построения графов взаимосвязей между участниками событий [3].

Далее будет применятся комбинированный подход с кластеризацией документов без заданных меток, выделением ключевых слов из полученных кластеров и их дальнейшей классификацией по целям устойчивого развития.

Отслеживание прогресса в достижении ЦУР обосновывается необходимостью представления механизма обратной связи, позволяющего оценивать результативность реализуемых мер и своевременно выявлять отклонения от запланированных траекторий. Наличие количественных индикаторов трансформирует политические обязательства в измеримые параметры, что создаёт основу для объективного анализа достигнутых результатов, а систематический сбор данных на национальном и региональном уровнях обеспечивает информационную базу для корректировки государственных программ и стратегий развития [4]. В рамках дипломной работы предлагается использовать новостной контент в качестве дополнительного способа исследования достижения ЦУР.

Во втором разделе были последовательно рассмотрены ключевые

этапы обработки текстов на естественном языке [5], методы кластеризации, подходы к оценке их качества и анализ тональности. Проведённый обзор методов предобработки показал, что лемматизация с фильтрацией по частям речи и последующей векторизацией TF-IDF [6] формирует признаковое пространство, адекватно отражающее тематическую структуру новостных текстов при сохранении вычислительной эффективности на больших объёмах данных. Сравнительный анализ алгоритмов кластеризации выявил, что плотностные методы, в частности HDBSCAN, обладают преимуществом перед разделительными и вероятностными подходами при работе с текстовыми данными, поскольку способны выделять кластеры произвольной формы, автоматически определять их количество и отсеивать шумовые наблюдения — свойства, критически важные для анализа неразмеченных новостных потоков [7]. Для количественной оценки качества разбиения обосновано применение совокупности внутренних метрик (коэффициент силуэта, индексы Дэвиса-Болдуина и Калински-Харабаша), позволяющих оценить компактность и отделимость кластеров в условиях отсутствия эталонной разметки [8].

Рассмотрение подходов к анализу тональности позволило выделить три ключевых направления: методы на основе правил, методы машинного обучения и гибридные подходы. Установлено, что словарные методы, базирующиеся на лексиконе RuSentilex, обеспечивают прозрачность и интерпретируемость результатов при минимальных требованиях к подготовке данных, что делает их предпочтительным выбором для оценки тонального фона новостных сообщений.

В третьем разделе была осуществлена практическая реализация методики анализа новостного контента на предмет представленности в нём тематики целей устойчивого развития.

Выбор программного инструментария для решения задач обработки естественного языка (Natural Language Processing, NLP) определяется требованиями к аналитическим возможностям, доступным библиотекам и воспроизводимости результатов. В рамках дипломной работы в качестве основной среды разработки используется язык статистического программирования R [9]. Решение основано на соответствии архитектуры и экосистемы языка специфике лингвистических данных и решаемой задаче кластеризации.

На этапе обоснования методов предобработки было установлено, что

для русскоязычных текстов лемматизация с использованием библиотеки udpipe обеспечивает более высокое качество по сравнению со стеммингом, что подтверждено в ходе анализа 93 тысяч новостных статей из 14 регионов Приволжского федерального округа за период 2020–2025 годов. В качестве метода векторизации избран TF-IDF, позволяющий выделить информативные термины при сохранении вычислительной эффективности, а для уменьшения размерности — алгоритм UMAP [10], который сохраняет как локальную, так и глобальную структуру данных при сопоставимых или меньших временных затратах.

Анализ тональности выполнен словарным методом на базе лексикона RuSentilex, что позволило получить количественные оценки эмоциональной окраски новостных сообщений для каждой пары «регион — цель устойчивого развития». Результаты визуализированы в виде тепловой карты, представленной на рисунке 1.

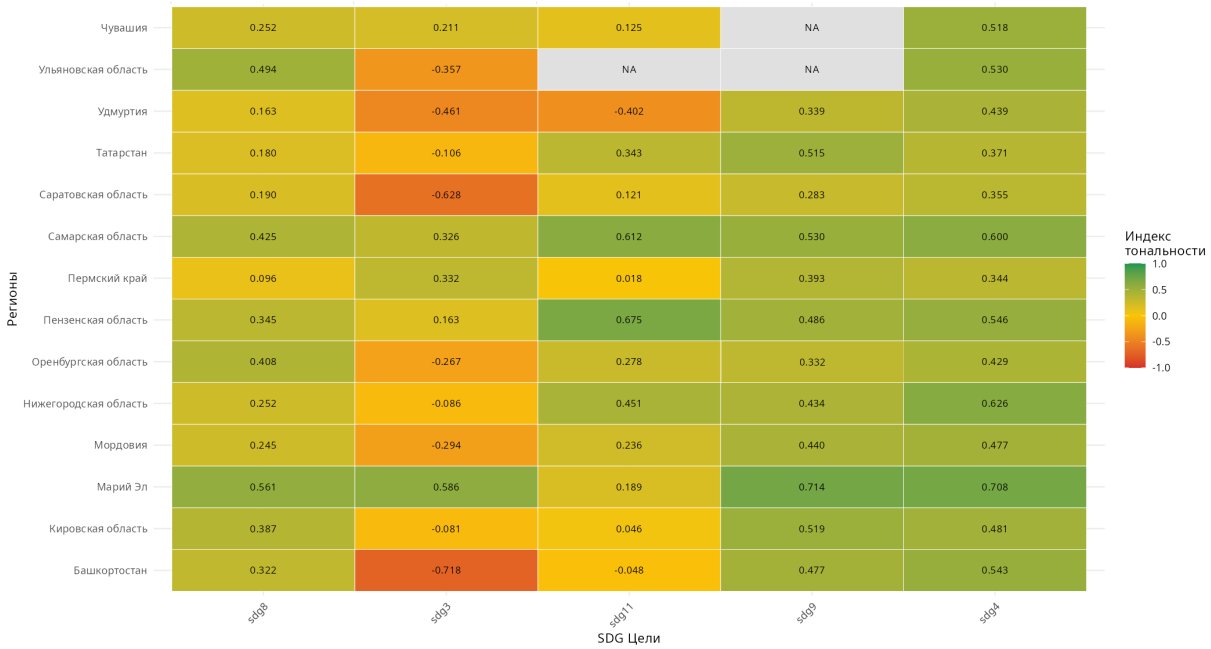

Рисунок 1 – Тепловая карта

Разработанный графический интерфейс на основе пакета Shiny реализует сквозной цикл анализа: от загрузки исходных данных и управления параметрами обработки до интерактивного исследования результатов. Все возможности приложения представленны на рисунке 2 в виде use-case диаграммы.

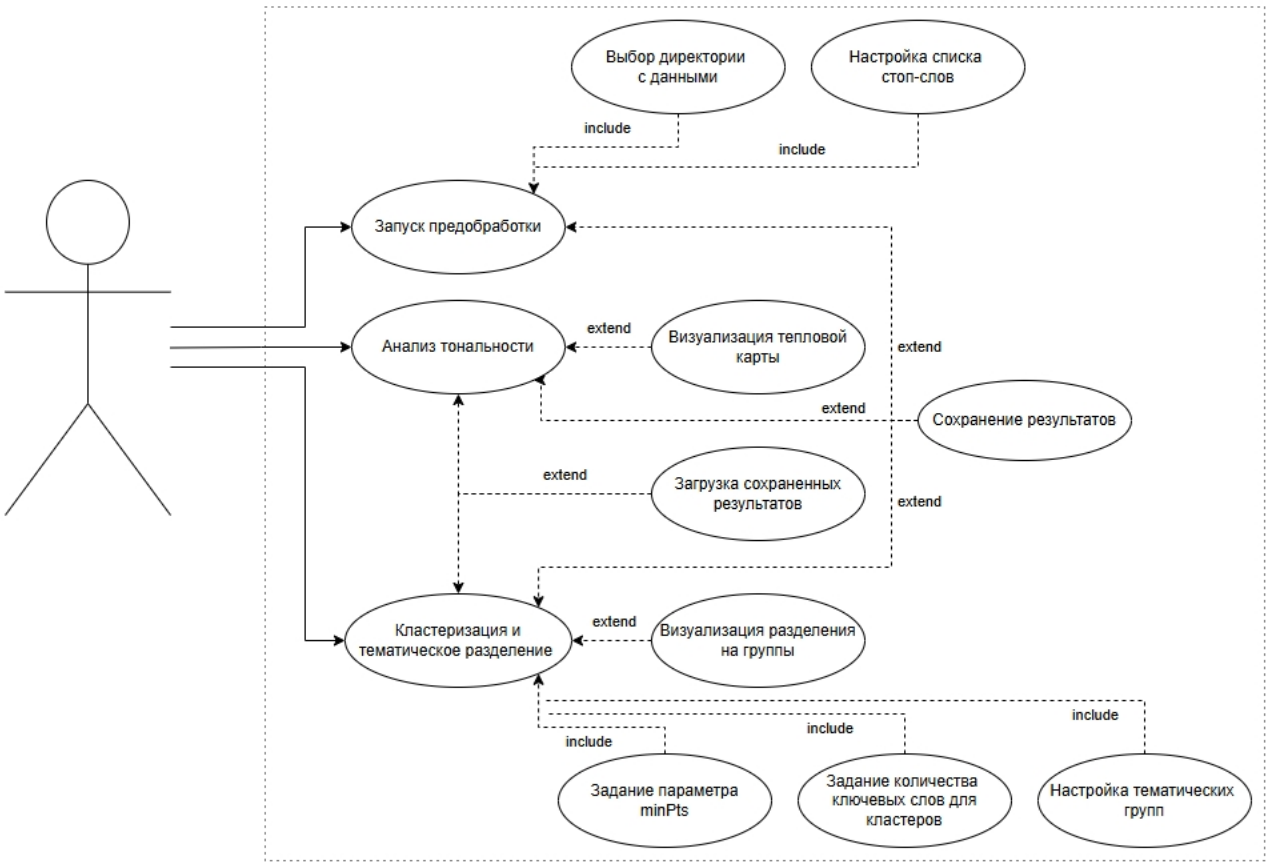

Рисунок 2 – Use-case диаграмма

Приложение включает функции редактирования списков ключевых слов и стоп-слов, загрузки и сохранения промежуточных результатов, фильтрации и сортировки табличных данных, что делает его адаптируемым к изменяющимся аналитическим потребностям.

ЗАКЛЮЧЕНИЕ

В дипломной работе была реализована методика оценки устойчивого развития регионов на основе анализа новостного контента с применением методов Text Mining.

На основе новостных статей, сегментированных по регионам Приволжского федерального округа за период 2020 – 2025 годов, сформирован структурированный набор данных объёмом 93 тысячи документов. Предобработка включала токенизацию, лемматизацию с использованием модели UDPipe для русского языка, фильтрацию по частям речи, удаление стоп-слов и векторизацию методом TF-IDF. Выбор лемматизации перед стеммингом обоснован более высоким качеством выделения канонических форм слов для русскоязычных текстов, что положительно сказалось на точности последующего тематического моделирования и анализа тональности.

Выполнена кластеризация подготовленных текстов, проведено сравнительное исследование различных моделей кластеризации (HDBSCAN, OPTICS и GMM) по внутренним метрикам качества (Silhouette, индекс Дэвиса-Болдуина, индекс Калински-Харабаша). Установлено, что HDBSCAN с параметром minPts = 15 демонстрирует наилучшие результаты: коэффициент силуэта составил 0,804, индекс Дэвиса-Болдуина — 0,393, индекс Калински-Харабаша — 22556,29. Модель GMM показала существенно более низкое качество (Silhouette не превысил 0,304), что объясняется несоответствием предположения об эллиптической форме кластеров реальной геометрии текстовых данных. OPTICS продемонстрировал наихудшие результаты (Silhouette от -0,032 до 0,012) вследствие чувствительности к выбору параметров eps и minPts.

Затем, полученные с помощью наиболее точной модели кластеры соотнесены с различными целями устойчивого развития. Для этого применена мультиязычная модель Sentence-BERT, позволившая сопоставить каждый тематический кластер с соответствующей целью устойчивого развития на основе косинусной близости семантических векторов.

Был проведён анализ тональности новостных сообщений. С использованием словарного подхода на базе RuSentilex вычислены средние оценки тональности для каждой пары «регион — цель устойчивого развития». Результаты визуализированы в виде интерактивной тепловой карты, что позволяет оценить тональность освещения отдельных целей устойчивого развития

в разрезе регионов и выявить как позитивно, так и негативно окрашенные тематические направления.

Разработано web-приложение, позволяющее агрегировать разрозненные новостные сообщения, идентифицировать скрытые тематические структуры и количественно оценивать характер их освещения в региональном разрезе с интерактивным представлением результатов. Реализована возможность редактирования списков ключевых слов и стоп-слов, загрузки и сохранения промежуточных результатов, что расширяет функциональность инструмента и делает его адаптируемым к изменяющимся аналитическим потребностям.

В качестве дальнейшего направления исследования предполагается использование больших языковых моделей для решения задачи тематической классификации новостного контента.

Основные источники информации:

1 Yoon, J., Han, S., Lee, Y., Hwang, H. Text Mining Analysis of ESG Management Reports in South Korea: Comparison With Sustainable Development Goals. Sage Open. 2023, 13, 4, 21582440231202896.

2 Foroudi, P., Marvi, R., Cuomo, M.T., D'Amato, A. Sustainable Development Goals in a regional context: conceptualising, measuring and managing residents' perceptions // Regional Studies. 2024, p. 1–16.

3 Spinder, S.; Frasincar, F.; Matsiiako, V.; Boekestijn, D.; Brandt, T. A text mining approach to identifying sustainability in the private sector. Computers in Industry 2023, 149, 103932.

4 Pravitasari, A.E., Rustiadi, E., Mulya, S.P., Fuadina L.N. Developing Regional Sustainability Index as a New Approach for Evaluating Sustainability Performance in Indonesia. Environ. Ecol. Res. 2018, 6(3), 157–168.

5 Обработка естественного языка в действии. / Хобсон Лейн, Ханнес Хапке, Коул Ховард — СПб.: Питер, 2020. — 576с.

6 Liang-Ching, Ch. An extended TF-IDF method for improving keyword extraction in traditional corpus-based research: An example of a climate change corpus. Data Knowl. Eng. 2024, 153, 102322.

7 Махрусе, Н. Современные тенденции методов интеллектуального анализа данных: метод кластеризации // Московский экономический журнал. 2019. №6. С. 69-77.

8 Внешние метрики качества: [Электронный ресурс] URL:

https : *//deepmachinelearning.ru/docs/Machine−learning/Clustering−evaluation/External − clustering − evaluation − measures* (дата обращения: 24.04.2026) -Загл. с экрана - Яз. рус.

9 Rdocumentation: [Электронный ресурс] URL: *https* : *//www.rdocumentation.org* (дата обращения: 25.04.2026) -Загл. с экрана - Яз. англ.

10 A Survey: Potential Dimensionality Reduction Methods For Data Reduction / F. H. Mean, Y. I. Chang, Institute of Statistical Science Academia Sinica, Taipei, Taiwan 11529