

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра теории функций и стохастического анализа

**ПРИМЕНЕНИЕ ИНСТРУМЕНТОВ ОБРАБОТКИ
ЕСТЕСТВЕННОГО ЯЗЫКА ДЛЯ МОНИТОРИНГА И
АНАЛИЗА УПОМИНАНИЙ КРУПНЕЙШИХ КОМПАНИЙ
САРАТОВСКОЙ ОБЛАСТИ В TELEGRAM-КАНАЛАХ**

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Ерохиной Ксении Михайловны

Научный руководитель
доцент, к. ф.-м. н.

Д. В. Мельничук

Заведующий кафедрой
д. ф.-м. н., доцент

С. П. Сидоров

Саратов 2025

ВВЕДЕНИЕ

Актуальность темы исследования. Современные цифровые технологии кардинально изменили способы распространения информации, сделав социальные сети и мессенджеры ключевыми платформами для общественного обсуждения. Telegram, как одна из самых популярных коммуникационных платформ в России, активно используется для обсуждения новостей, в том числе деятельности региональных компаний. Для бизнеса, особенно для крупнейших компаний Саратовской области, мониторинг упоминаний в Telegram-каналах становится важным инструментом для анализа общественной реакции, выявления потенциальных кризисных ситуаций и оценки эффективности коммуникационных стратегий.

Актуальность работы обусловлена необходимостью автоматизированного мониторинга и анализа текстовых данных для выявления оценки тональности упоминаний в информационном поле. Применение инструментов обработки естественного языка (NLP) позволяет значительно ускорить и упростить эти процессы.

Цель работы: разработка системы мониторинга и анализа упоминаний крупнейших компаний Саратовской области в Telegram-каналах на основе инструментов обработки естественного языка.

Задачи исследования:

1. Изучить теоретические основы обработки естественного языка (NLP).
2. Провести анализ предметной области, включая медиаактивность компаний Саратовской области в Telegram.
3. Разработать систему мониторинга и анализа на основе Telegram-бота.
4. Провести тестирование и валидацию системы.

Практическая значимость работы заключается в разработке программного продукта для автоматизированного мониторинга и анализа упоминаний компаний, что позволяет автоматизировать процесс мониторинга медиапространства, сократить временные затраты на ручной анализ данных и повысить эффективность управления репутацией компаний.

Основное содержание работы

Бакалаврская работа состоит из: введения, двух теоретических и двух практических разделов, заключения, списка использованных источников и приложений.

Введение содержит основные положения: актуальность темы, цель, задачи, практическую значимость исследования, а также структуру работы.

Первый раздел «Теоретические основы обработки естественного языка» описывает теоретические основы обработки естественного языка.

Обработка естественного языка: основные понятия и методы

Обработка естественного языка (Natural Language Processing, NLP) — это междисциплинарная область, объединяющая лингвистику, компьютерные науки и искусственный интеллект. Ее основная цель заключается в разработке алгоритмов и моделей, способных автоматически анализировать, понимать и генерировать человеческую речь.

Ключевые задачи:

NLP решает широкий спектр задач, включая:

- Токенизацию (разбиение текста на слова/предложения),
- Нормализацию (стемминг, лемматизация, удаление стоп-слов),
- Частеречную разметку (POS-tagging),
- Распознавание сущностей (NER),
- Синтаксический и семантический анализ,
- Анализ тональности,
- Машинный перевод и генерацию текста.

Основные подходы:

1. Статистические методы (N-граммы, TF-IDF, HMM) – ранние способы обработки текста на основе частотных закономерностей.

2. Машинное обучение (SVM, логистическая регрессия, LSA, LDA) – классификация, кластеризация и тематическое моделирование.

3. Глубокое обучение – современные нейросетевые архитектуры:

- Word Embeddings (Word2Vec, GloVe) для векторного представления слов,

- RNN, LSTM – для работы с последовательностями,
- Трансформеры (BERT, GPT) – модели на основе механизма внимания.

4. Гибридные методы – сочетание rule-based систем с ML/нейросетями для повышения точности.

Современные инструменты и библиотеки для текстового анализа

Современные инструменты NLP можно разделить на несколько категорий:

1. Универсальные библиотеки

NLTK предлагает широкий набор инструментов для обучения и исследований, spaCy обеспечивает высокую производительность в промышленных задачах (токенизация, NER, синтаксический анализ), а Gensim специализируется на тематическом моделировании и векторных представлениях.

2. Библиотеки для глубокого обучения

Hugging Face Transformers предоставляет доступ к предобученным моделям (BERT, GPT), а TensorFlow Text и PyTorch-NLP интегрируют NLP-методы в соответствующие фреймворки.

3. Специализированные решения

Stanza поддерживает морфологический анализ для 70+ языков, Flair работает с контекстными эмбедингами, а AllenNLP ориентирован на исследовательские задачи.

Ключевые тенденции в развитии инструментов NLP:

- Упрощение использования сложных моделей через API (Hugging Face, LangChain)
- Развитие мультязычных моделей (mBERT, XLM-R)
- Интеграция с облачными сервисами (Google NLP, AWS Comprehend)

Критерии выбора инструментов

При подборе инструментов учитывают:

- Функциональность (поддержка нужных задач — от токенизации до анализа тональности)
- Производительность (скорость обработки больших данных)

- Поддержку языка (особенно для русского)
- Технические аспекты (интеграция в инфраструктуру, масштабируемость)
- Эксплуатационные факторы (документация, сообщество, лицензия)

Второй раздел «Анализ предметной области» содержит исследование медиаактивности крупнейших компаний Саратовской области в Telegram и анализ особенностей платформы как источника данных.

Крупнейшие компании Саратовской области и их медиаактивность

Саратовская область – важный промышленный регион России с развитыми нефтегазовой, энергетической и строительными отраслями. В исследовании анализируется медиаактивность двух компаний региона в Telegram.

1. Т Плюс Саратов, обеспечивающая тепло- и электроснабжение региона, ведет собственный Telegram-канал. Дополнительно информация о компании появляется в местных СМИ и популярных пабликах: «Новости Саратова и Области», «Саратов 24», «ЧП Саратов» и «Саратов | Без цензуры».

2. Газпром трансгаз Саратов, занимающийся транспортировкой газа, также имеет официальный Telegram-канал. Новости компании дублируются в «Новостях Саратова и Области», «ЧП Саратов» и канале Романа Бусаргина.

Особенности Telegram как платформы для мониторинга

Telegram, сочетающий функции мессенджера и медиаплатформы, представляет собой уникальный инструмент для мониторинга публикаций. Его главное преимущество - открытая архитектура: большинство новостных каналов доступны без авторизации, а развитое API позволяет автоматизировать сбор данных. Информация в Telegram распространяется с исключительной скоростью благодаря системе мгновенных уведомлений, что делает платформу идеальной для оперативного отслеживания упоминаний компаний.

Особенность Telegram - поддержка мультимедийного контента: помимо текста, анализировать приходится изображения, видео и документы, что требует комплексных NLP-решений. Политика анонимности, с одной стороны, обеспечивает свободу высказываний, с другой - ставит вопрос достовер-

ности информации. При этом встроенные возможности автоматизации через ботов позволяют внедрять инструменты sentiment-анализа и тематического моделирования непосредственно в платформу.

Таким образом, Telegram предлагает мощный, но сложный инструментарий для мониторинга, требующий адаптированных NLP-методов для работы с разнородным контентом и обеспечения точности анализа.

Потребности заинтересованных сторон в анализе упоминаний

Мониторинг публикаций в Telegram представляет ценность для различных групп, каждая из которых использует его для решения специфических задач. Крупные компании региона, включая промышленные предприятия Саратовской области, активно применяют анализ упоминаний для управления репутацией, выявления негативных отзывов и оценки эффективности маркетинговых кампаний. Для этих целей особенно востребованы инструменты sentiment-анализа, позволяющие оперативно реагировать на изменения общественного мнения.

Государственные структуры используют мониторинг Telegram для отслеживания социально-экономических тенденций и раннего выявления потенциальных кризисных ситуаций. В их случае акцент делается на тематическом моделировании, помогающем выявлять ключевые проблемы, волнующие население. Инвесторы же рассматривают анализ упоминаний как важный элемент due diligence, применяя методы извлечения именованных сущностей для выявления скрытых рисков и прогнозирования рыночных трендов.

Особое значение Telegram имеет для СМИ, которые используют платформу как источник оперативной информации. Журналисты заинтересованы в инструментах верификации контента и сравнительного анализа, позволяющих отделять достоверные сведения от дезинформации. Разнообразие этих задач подчеркивает необходимость развития специализированных NLP-решений, способных учитывать специфику потребностей каждой группы стейкхолдеров при работе с контентом Telegram.

Третий раздел «Разработка системы мониторинга и анализа» описывает архитектуру Telegram-бота, выбор инструментов и реализацию функционала для парсинга каналов, классификации тональности сооб-

щений и визуализации результатов. Система разработана на языке Python с использованием специализированных библиотек для обработки естественного языка, включая `razdel` для токенизации русского текста и `sklearn.feature_extraction.text.TfidfVectorizer` для векторизации текста методом TF-IDF. Основная логика анализа текста реализована в кастомном модуле `nlp_tools`, который включает функции `analyze_text` для обработки контента и `create_summary_pdf` для генерации отчетов. Для визуализации результатов применяются библиотеки `wordcloud` для создания облаков слов и `matplotlib` для построения графиков. Генерация PDF-отчетов осуществляется с помощью `fpdf2`, а взаимодействие с пользователями организовано через Telegram-бота на базе фреймворка `aiogram`. Дополнительно система использует стандартные библиотеки Python: `json` для работы с данными в формате JSON, `re` для обработки текста регулярными выражениями, `datetime` для работы с временными метками, `collections.Counter` для частотного анализа и `logging` для ведения логов работы системы. Этот комплексный набор инструментов обеспечивает полный цикл обработки текстовых данных — от сбора и первичного анализа до визуализации результатов и формирования итоговых отчетов.

Четвёртый раздел «Тестирование и валидация системы» содержит практическое применение инструментов обработки естественного языка (NLP) для мониторинга и анализа упоминаний крупнейших компаний Саратовской области.

Проведённый анализ реальных сообщений из Telegram-каналов, посвящённых деятельности компаний «Т Плюс Саратов» и ООО «Газпром трансгаз Саратов», подтвердил работоспособность и точность разработанной системы мониторинга. В рамках тестирования решались следующие ключевые задачи: автоматическое выявление упоминаний компаний, определение тональности текста (позитивная, негативная, нейтральная), выделение ключевых слов и визуализация данных в виде облака слов и графика TF-IDF.

В результате тестирования система показала высокую эффективность в мониторинге Telegram-каналов: успешно идентифицирует компании, анализирует тональность сообщений (например, чётко определяет негатив в новостях об экологических инцидентах) и выделяет ключевые термины.

Автоматически генерируемые отчёты и визуализации позволяют оперативно оценивать информационный фон вокруг предприятий.

Основные результаты

1. Изучены теоретические основы обработки естественного языка, включая ключевые задачи, методы и современные инструменты NLP. Проведён анализ статистических подходов, методов машинного обучения и нейросетевых архитектур, применяемых для обработки текстовых данных. Рассмотрены критерии выбора библиотек и инструментов для реализации системы мониторинга.

2. Проведён анализ предметной области, в рамках которого определены крупнейшие компании Саратовской области и их медиаактивность в Telegram. Исследованы особенности платформы Telegram как источника данных, включая её открытость, скорость распространения информации и поддержку автоматизации. Проанализированы потребности заинтересованных сторон в мониторинге упоминаний.

3. Разработана система мониторинга и анализа упоминаний компаний в Telegram-каналах на основе инструментов NLP. Описана архитектура Telegram-бота, включая модули для сбора данных, предварительной обработки текста, анализа тональности и визуализации результатов. Реализован функционал для автоматической классификации сообщений и формирования отчётов.

4. Проведены тестирование и валидация системы на реальных данных из Telegram-каналов. Продемонстрированы примеры анализа упоминаний компаний, оценена точность работы алгоритмов и выявлены направления для дальнейшего улучшения системы. Результаты тестирования подтвердили эффективность предложенного решения.

5. Практическая значимость работы заключается в создании инструмента, который позволяет автоматизировать процесс мониторинга медиапространства, сократить временные затраты на ручной анализ и повысить качество управления репутацией компаний. Система может быть доработана за счёт интеграции дополнительных NLP-моделей для более детального анализа контекста упоминаний.