ИНФОРМАТИКА

COMPUTER SCIENCES

Вестник Сыктывкарского университета.
Серия 1: Математика. Механика. Информатика. 2025.
Выпуск 2 (55)
Bulletin of Syktyvkar University.
Series 1: Mathematics. Mechanics. Informatics. 2025; 2 (55)

Научная статья

УДК 517.9, 539.3
https://doi.org/10.34130/1992-2752_2025_2_8

АНАЛИЗ КОНТЕНТА СОЦИАЛЬНЫХ СЕТЕЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Кирилл Павлович Колпаков,
Владимир Александрович Устюгов,
Иван Иванович Лавреш
Сыктывкарский государственный
университет им. Питирима Сорокина, ustyugov@syktsu.ru

Аннотация. Роль социальных сетей в современном обществе трудно переоценить. Они стали основным инструментом коммуникации, позволяющим мгновенно обмениваться информацией независимо от геолокации пользователей. Более того, социальные сети играют значительную роль в формировании общественного мнения, мобилизации населения для участия в социальных и политических акциях, а также в развитии бизнеса через маркетинг и прямую связь с потребителями. Тем не менее вместе с возможностями социальные сети несут и определенные вызовы. Вопросы, связанные с интернет-зависимостью, нарушением приватности, распространением фальсифицированной информации и цифровым насилием, становятся все более острыми. Эти проблемы выделяют критическую необходимость в разработке и внедрении эффективных механизмов контроля за контентом, что требует

активного взаимодействия между разработчиками, исследователями и пользователями для создания безопасного цифрового пространства.

Основная цель исследования — разработка и реализация системы, которая может применяться для мониторинга аккаунтов в социальных сетях работников государственных учреждений, образовательных организаций и коммерческих компаний, и способной автоматически определять негативные и положительные посты, а также выявлять потенциально опасный контент.

Ключевые слова: анализ контента, социальные сети, обработка естественного языка, NLP, машинное обучение, LiveJournal, ВКонтакте, тональность текста, информационная безопасность, деструктивное поведение, трансформеры, ruT5

Для цитирования: Колпаков К. П., Устюгов В. А., Лавреш И. И. Анализ контента социальных сетей с использованием методов обработки естественного языка // *Вестник Сыктывкарского университета. Сер. 1: Математика. Механика. Информатика.* 2025. Вып. 2 (55). С. 8–19. https://doi.org/10.34130/1992-2752_2025_2_8

Article

SOCIAL MEDIA CONTENT ANALYSIS USING NATURAL LANGUAGE PROCESSING TECHNIQUES

Kirill P. Kolpakiov, Vladimir A. Ustyugov, Ivan I. Lavresh
Pitirim Sorokin Syktyvkar State University, ustyugov@syktsu.ru

Abstract. The role of social media in modern society cannot be overestimated. They have become the main communication tool that allows you to instantly exchange information regardless of the geolocation of users. Moreover, social networks play a significant role in shaping public opinion, mobilizing the population to participate in social and political actions, as well as in business development through marketing and direct communication with consumers. Nevertheless, along with the opportunities, social networks also bring certain challenges. Issues related to Internet addiction, privacy violations, the spread of falsified information and digital violence are becoming increasingly acute. These issues highlight the critical need to develop and implement effective content control mechanisms, which requires active collaboration among developers, researchers, and users to create a secure digital space.

The main purpose of the research is to develop and implement a system that can be used to monitor social media accounts of employees of government agencies, educational organizations, and commercial companies, and is able to automatically identify negative and positive posts, as well as identify potentially dangerous content.

Keywords: content analysis, social networks, natural language processing, NLP, machine learning, LiveJournal, VKontakte, text tonality, information security, destructive behavior, transformers, ruT5

For citation: Kolpakov K. P., Ustyugov V. A., Lavresh I. I. Social media content analysis using natural language processing techniques. *Vestnik Syktyvkarskogo universiteta. Seriya 1: Matematika. Mekhanika. Informatika* [Bulletin of Syktyvkar University, Series 1: Mathematics. Mechanics. Informatics], 2025, no 2 (55), pp. 8–19. (In Russ.) https://doi.org/10.34130/1992-2752_2025_2_8

1. Введение в обработку естественного языка

Среди многочисленных опций и методов, предоставляемых современными технологиями, для решения задачи анализа публикаций в социальных сетях было выбрано конкретное направление в машинном обучении, а именно Natural Language Processing (NLP) — обработка естественного языка [1]. Блок-схема, детализирующая процессы NLP, приведена на рис. 1. Выбор технологии обусловлен спецификой задачи исследования, связанной с анализом текстовых данных и выявлением скрытых закономерностей в языковых структурах.

Рис. 1. Блок-схема, детализирующая процессы, связанные с обработкой естественного языка (Natural Language Processing, NLP)

Обозначим этапы, необходимые для анализа текстов согласно технологии NLP [2; 3] (схема взаимосвязи приведена на рис. 2):

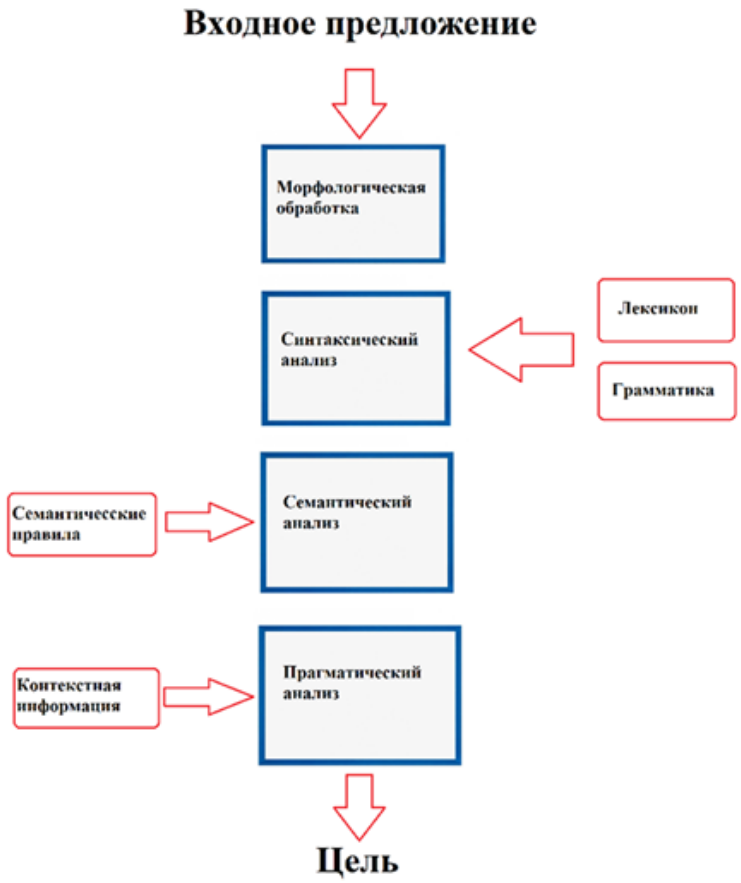

Рис. 2. Необходимые этапы анализа естественного языка

1) Лексический анализ представляет собой начальный этап обработки текста, на котором происходит разбиение текста на токены. *Токенизация* — это процесс, в ходе которого непрерывный текст преобразуется в дискретные единицы, что упрощает дальнейшую обработку. *Стемминг* и *лемматизация*, важные компоненты лексического анализа, вносят вклад в нормализацию текста путем приведения слов к их корневой форме, например, слова «бежит», «бежал», «бегу» будут преобразованы к «бег». Дополнительно существует технология удаления шумовых слов, она представляет собой алгоритм, исключающий из текста слова, не несущие значительной смысловой нагрузки.

2) Синтаксический анализ, или *парсинг*, это процесс анализа предложений с целью определения их грамматической структуры и взаимо-

связей между словами. Результатом является построение дерева зависимостей, которое показывает, как слова сочетаются для формирования смысловых единиц.

3) *Семантический анализ* направлен на понимание значения слов и их взаимодействия в предложениях. Модели векторного представления слов, такие как Word2Vec или GloVe, представляют слова в виде элементов в многомерном пространстве, где семантически похожие слова располагаются ближе друг к другу. Это позволяет количественно оценивать семантическую близость слов, анализировать связи и отношения между ними.

4) *Прагматический анализ* расширяет семантический анализ, то есть выходит за рамки лексического, синтаксического и семантического анализа, учитывая контекст и цель использования языка. Этот аспект NLP исследует, как контекст влияет на интерпретацию слов, что особенно важно для понимания иронии, сарказма, двусмысленности и других нюансов общения.

5) *Методы машинного обучения* — дерево принятия решений, логистическая регрессия, кластеризация, методы опорных векторов, применяемые в области обработки естественного языка (NLP), обеспечивают фундамент для разработки систем, способных анализировать, понимать человеческий язык, а также генерировать тексты в его естественной форме.

2. Методы и материалы

Перечислим библиотеки языка Python, выбранные в качестве основополагающих в реализации нашей системы.

Библиотека **Natasha** — это набор специализированных инструментов для обработки текстов на русском языке, она включает модули для разбиения текста на слова и предложения, анализа структуры предложений, определения грамматических характеристик слов, выделения значимой информации на основе заданных правил, а также визуализации результатов анализа.

Библиотека **Dostoevsky** — инструмент для анализа тональности текста на русском языке с использованием предобученных моделей FastText, классифицирующий тексты по эмоциональной окраске (позитивные, негативные, нейтральные и др.). Схема анализа тональности текста с использованием данной библиотеки приведена на рис. 3. Библиотека предоставляет простой API для интеграции с Python-проектами, что позволяет легко внедрять функционал анализа тональности в различные приложения.

Рис. 3. Схема работы библиотеки Dostoevsky

Библиотека **Transformers** — библиотека для работы с современными моделями обработки естественного языка, такими как BERT, GPT, T5 и другими, она выполняет следующие функции:

1. Загрузка предобученной модели — библиотека позволяет использовать уже обученные языковые модели (в нашем случае ai-forever/ruT5-base) без необходимости их самостоятельного обучения.

2. Токенизация текста — библиотека преобразует текст в числовой формат, понятный нейросети, обеспечивая совместимость с моделью.

3. Классификация текста — AutoModelForSequenceClassification применяет предобученную модель для анализа входного текста и возвращает результаты в виде логитов (сырых оценок вероятностей).

Библиотека **Pymystem3** — библиотека для лемматизации текстов на русском языке. Основные возможности: преобразование слов в их начальную форму (лемматизация), поддержка русского языка.

Библиотека **Datasets** — библиотека для загрузки и обработки датасетов, часто применяемых в NLP задачах. Главные возможности и преимущества данной библиотеки заключаются в предоставлении легкого доступа к популярным NLP датасетам и простоте интерфейса для загрузки, обработки и разделения данных на тренировочные и тестовые выборки. Также в процессе разработки системы были использованы библиотеки telebot, requests, xmltodict, torch. Эти библиотеки обеспечили необходимую функциональность для обработки текстов, создания и управления интерфейсом системы в виде Telegram-бота, а также для выполнения машинного обучения.

Для анализа текстов на политическую тематику и другие темы использовалась модель, ранее обученная на открытых данных с платформы Pikabu – ruT5-base [4].

Для целевого обучения модели использовался датасет с сайта HuggingFace, который содержит посты из платформы Pikabu [5]. Этот датасет размечен различными тегами. Для решения задачи преобразуем теги так, чтобы остались только два: «политика» и «иное». Это необходимо для упрощения анализа и повышения точности.

Обучение модели — это очень затратный с точки зрения вычислений процесс, поэтому использовалось подмножество данных, которое способно обработать наше оборудование, а именно содержание 20000 постов с сайта Pikabu.

Обучение модели включало 20 циклов, каждый цикл занимал около двух часов. После каждого цикла обучения необходимо проверять результаты и, при необходимости, корректировать параметры обучения, чтобы улучшить качество модели. Наконец, необходимо прописать функцию инференса (применение обученной модели).

Теперь мы имеем обученную модель, готовую к использованию. На рис. 4 представлены результаты обучения модели — на левом графике отображены потери на тренировочных и тестовых данных по мере обучения модели, а на правом — точность модели на тестовых данных. Эти графики помогают понять, как изменялись показатели модели во время обучения и насколько хорошо модель обучилась распознавать нужные категории.

С полным кодом работы можно ознакомиться по ссылке https://clck.ru/3Mke6R.

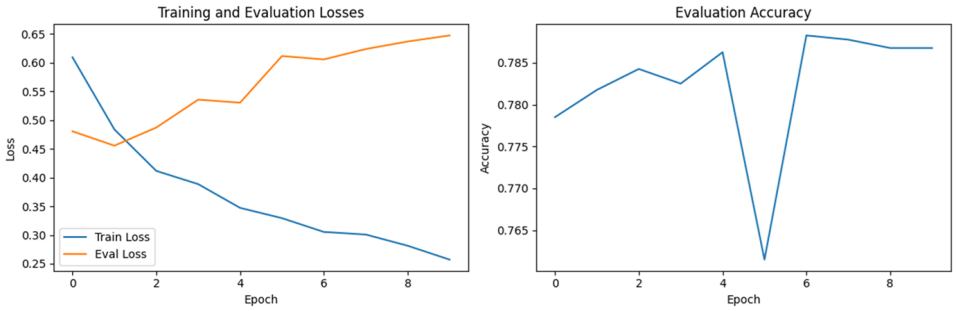

Рис. 4. Результаты обучения модели

3. Результаты и обсуждение

После того как все компоненты системы разработаны, важно убедиться, что они работают правильно. Тестирование включает проверку парсинга данных, анализа тональности, инференса модели и взаимодействия с Telegram-ботом

Работа конечного пользователя с разработанным программным обеспечением выглядит следующим образом. Бот запрашивает уникальный ID или никнейм исследуемого аккаунта, и в реальном времени пользователю предоставляется плашка с процентом готовности анализа (процесс обработки происходит на сервере). Иллюстрация работы системы приведена на рис. 5.

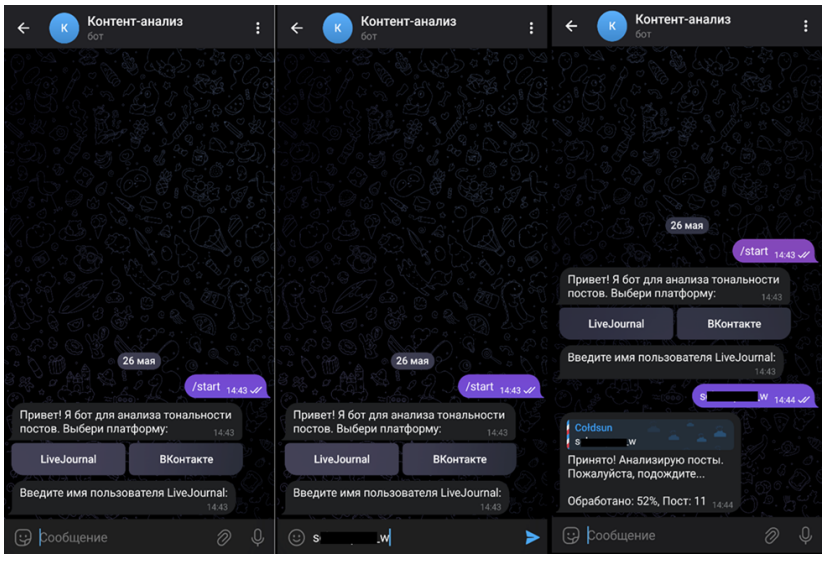

Рис. 5. Процедура поиска, сбора и анализа постов выбранного пользователя сайта LiveJournal

Результат предоставляется в виде сообщений, каждое из которых включает ссылку на пост, категорию, теги, ключевые слова и значение настроения. Пример отображения результатов приведен на рис. 6.

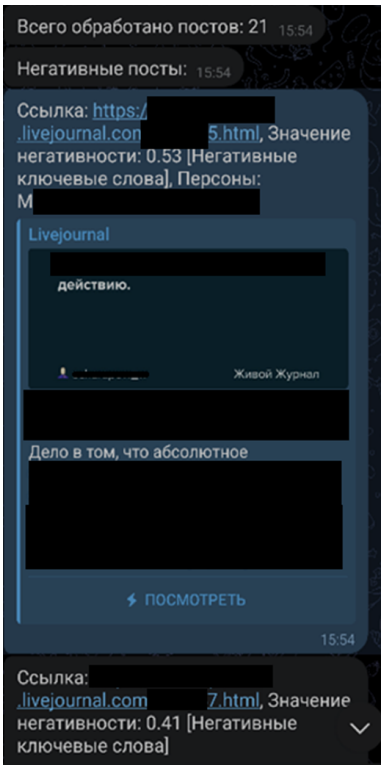

Рис. 6. Пример результата анализа публикаций пользователя сайта LiveJournal

В случае если пользователь недоволен результатом, он может продолжить сбор постов выбранного пользователя с последующим их анализом, нажав кнопку «Продолжить поиск».

Для дополнительной проверки точности системы была проведена выборка из 100 постов одного аккаунта «ВКонтакте», из которых 50 % постов были политического содержания, а остальные 50 % — нейтрального. Посты были вручную классифицированы для оценки их тематической принадлежности. Полученные результаты сравнивались с автоматической классификацией, проведенной разработанной системой.

Процедура ручной проверки точности работы системы состояла из следующих этапов:

1. Были выбраны 100 последних постов одного аккаунта, опубликованных на момент исследования.

2. Каждый пост анализировался вручную и классифицировался на две категории: политический пост содержит выраженные мнения или обсуждения политических событий, действий властей и т.п.; либо нейтральный пост не содержит выраженной политической окраски и посвящен общим темам (быт, хобби, развлечения).

Результаты ручной классификации сравнивались с выводами системы. Было выявлено, что система продемонстрировала высокую точность, классифицируя 96 % политических и 98 % нейтральных постов правильно. Различие в 2–4 % объясняется сложными случаями, где текст содержит двусмысленные формулировки, затрудняющие как ручную, так и автоматическую классификацию.

4. Заключение

По итогам проведенного исследования была разработана и протестирована система анализа контента социальных сетей на основе современных методов обработки естественного языка и машинного обучения. Интеграция библиотек Natasha, Dostoevsky, Transformers, Pymystem3 позволила обеспечить высокую точность распознавания и интерпретации текстовых данных.

Реализация клиентской части системы в виде Telegram-бота продемонстрировала ее практическую пригодность и оперативность в вопросе взаимодействия пользователя с системой. Результаты тестирования показали высокое качество поиска, сбора, классификации и анализа, тем самым подтвердив эффективность предложенного подхода.

Значимость данной работы заключается в демонстрации возможности масштабируемого и точного анализа неструктурированных данных социальных сетей с помощью современных NLP-технологий. Полученные выводы и созданный прототип расширяют инструментарий исследований в области автоматизированной обработки текстов, что особенно актуально для задач мониторинга социальных сетей и анализа мнений пользователей.

В перспективе дальнейшие исследования включают оптимизацию алгоритмов для повышения производительности, а также углубление анализа сложных языковых конструкций, таких как сарказм и контекстно-зависимые выражения, с целью повышения точности интерпретации собранного материала.

Список источников

1. **Chakrawarti R. K.** Natural Language Processing for Software Engineering. Wiley. 2025. 544 p.

2. **Lee R. S. T.** Natural Language Processing. A Textbook with Python Implementation. Springer. 2024. 469 p.

3. **Лейн Х., Хапке Х., Ховард К.** Обработка естественного языка в действии. СПб.: *Питер*, 2020. 576 p.

4. ruT5-base [Электронный ресурс]. URL: https://huggingface.co/ai-forever/ruT5-base (дата обращения: 21.08.2025).

5. Pikabu dataset [Электронный ресурс]. URL: https://huggingface.co/datasets/IlyaGusev/pikabu (дата обращения: 21.08.2025).

References

1. **Chakrawarti R. K.** *Natural Language Processing for Software Engineering*. Wiley. 2025. 544 p.

2. **Lee R. S. T.** *Natural Language Processing. A Textbook with Python Implementation*. Springer, 2024. 469 p.

3. **Lein H., Hapke H., Howard K.** *Natural language processing in action*. St. Petersburg: Piter, 2020. 576 p. (In Russ.)

4. *ruT5-base* [Electronic resource]. Available at: https://huggingface.co/ai-forever/ruT5-base (accessed: 21.08.2025).

5. *Pikabu dataset* [Electronic resource]. Available at: https://huggingface.co/datasets/IlyaGusev/pikabu (accessed: 21.08.2025).

Сведения об авторах / Information about authors
Колпаков Кирилл Павлович / Kirill P. Kolpakov
магистрант / magister degree student

Сыктывкарский государственный университет имени Питирима Сорокина / Pitirim Sorokin Syktyvkar State University
167001, Россия, г. Сыктывкар, Октябрьский пр., 55 / 55, Oktyabrsky Ave., Syktyvkar, 167001, Russia

Устюгов Владимир Александрович / Vladimir A. Ustyugov
к.ф.-м.н., доцент, заведующий кафедрой информационной безопасности / Candidate of Science in Physics and Mathematics, Associate Professor, Head of the Information Security Department
Сыктывкарский государственный университет имени Питирима Сорокина / Pitirim Sorokin Syktyvkar State University
167001, Россия, г. Сыктывкар, Октябрьский пр., 55 / 55, Oktyabrsky Ave., Syktyvkar, 167001, Russia

Лавреш Иван Иванович / Ivan I. Lavresh
к.т.н., доцент кафедры информационной безопасности / Candidate of Sciences in Technology, Associate Professor of the Information Security Department
Сыктывкарский государственный университет имени Питирима Сорокина / Pitirim Sorokin Syktyvkar State University
167001, Россия, г. Сыктывкар, Октябрьский пр., 55 / 55, Oktyabrsky Ave., Syktyvkar, 167001, Russia

Статья поступила в редакцию / The article was submitted 19.05.2025
Одобрена после рецензирования / Approved after reviewing 29.05.2025
Принята к публикации / Accepted for publication 26.06.2025