IV Всероссийская (национальная) научная конференция с международным участием: «Наука, технологии, общество: Экологический инжиниринг в интересах устойчивого развития территорий» (НТО-IV-2023) при поддержке Краевого фонда науки

10 (2023)

УДК 621-039-542

EDN RAEHXE

Анализ образов будущего с помощью языковых моделей через исследование комментариев в социальных сетях

А.Д. Брагин

Национальный исследовательский Томский политехнический университет, пр. Ленина, 30, Томск, 634050, Россия

E-mail: bragin@tpu.ru

Аннотация. В статье исследуется применение больших языковых моделей (LLMs) для анализа контента в социальной сети Вконтакте, с акцентом на сообщества, посвященные научным и техническим дисциплинам. Исследование разделено на два этапа: сбор данных и их последующий анализ. На первом этапе был проведен сбор постов и комментариев из выбранных сообществ с использованием API Вконтакте. Второй этап включал количественный анализ ключевых слов для выявления главных тем дискуссий и интересов пользователей, а также применение LLMs для более глубокого понимания собранных текстов. Языковые модели помогли выявить общие тренды, тематики и сформировать представление о взглядах пользователей на будущее в контексте научных и технологических инноваций. Результаты исследования демонстрируют потенциал LLMs в качестве мощного инструмента для социальных наук, способного обрабатывать большие объемы данных и выявлять сложные паттерны в общественном сознании и коммуникации.

Ключевые слова: анализ социальных сетей, языковые модели, облако слов, тональность текста.

The future images analysis using language models through the study of comments on social networks

A.D. Bragin

National Research Tomsk Polytechnic University, Lenin Ave., 30, Tomsk, 634050, Russia

E-mail: bragin@tpu.ru

Abstract. This article explores the use of large language models (LLMs) for content analysis on the social network VKontakte, with a focus on communities dedicated to scientific and technical disciplines. The study is divided into two stages: data collection and subsequent analysis. At the first stage, posts and comments were collected from selected communities using the VKontakte API. The second phase involved quantitative keyword analysis to identify the main topics of discussion and user interests, as well as the application of LLMs to gain a deeper understanding of the collected texts. Language models helped identify common trends, topics, and form an understanding of users' views on the future in the context of scientific and technological innovation. The results of the study demonstrate the potential of LLMs as a powerful tool for the social sciences, capable of processing large amounts of data and identifying complex patterns in public consciousness and communication.

Keywords: social network analysis, language models, word cloud, text sentiment.

1. Введение

В последние десятилетия мы стали свидетелями стремительного развития технологий в области искусственного интеллекта, особенно в сегменте языковых моделей (Language Models, LLM). Эти модели, обученные на огромных массивах текстовых данных, демонстрируют поразительную способность к пониманию и генерации естественного языка, что открывает новые горизонты в самых разнообразных областях применения. Одной из таких областей, где LLM находят особенно активное применение, являются социальные сети — динамично развивающиеся платформы, где ежедневно генерируются терабайты текстовой информации.

Социальные сети стали неотъемлемой частью современной жизни, формируя новые формы коммуникации и взаимодействия между людьми. Они предоставляют беспрецедентные возможности для обмена информацией, выражения мнений и распространения идей. Однако, с ростом объемов данных, возникает потребность в их структурировании, анализе и интерпретации. Именно здесь языковые модели вступают в игру, предлагая мощные инструменты для обработки и анализа текстовой информации на масштабе, недоступном для человеческого аналитика.

2. Постановка задачи (Цель исследования)

Целью данной статьи является не только демонстрация эффективности языковых моделей в анализе социальных сетей и постов, связанных с образами будущего, но и обсуждение возможностей и ограничений, связанных с их применением. Языковые модели могут способствовать более глубокому пониманию социальных процессов, происходящих в цифровом пространстве, и какие новые перспективы они открывают для исследователей, маркетологов, политологов и других специалистов, работающих с социальными медиа.

3. Методы и материалы исследования

В обзоре литературы, посвященном применению Больших Языковых Моделей (LLMs) в области программной инженерии (Software Engineering, SE) [1], авторы Xinying Hou, Yanjie Zhao, Yue Liu и их коллеги представляют систематический анализ научных работ, опубликованных с 2017 по 2023 годы. Исследование направлено на понимание

того, как LLMs могут быть использованы для оптимизации процессов и результатов в программировании.

В рамках исследования были сформулированы и исследованы четыре ключевых вопроса. Первый вопрос касается классификации различных LLMs, используемых в задачах SE, их особенностей и применений. Второй вопрос анализирует методы сбора, предварительной обработки данных и применения моделей, подчеркивая важность хорошо структурированных наборов данных для успешного внедрения LLM в SE. Третий вопрос исследует стратегии, используемые для оптимизации и оценки производительности LLM в SE. Наконец, четвертый вопрос рассматривает конкретные задачи SE, где LLMs уже показали свою эффективность, демонстрируя их практический вклад в область.

Исследование подчеркивает, что, несмотря на значительный прогресс, понимание применения, эффектов и возможных ограничений LLM в SE все еще находится на начальном этапе. Авторы обсуждают текущее состояние искусства, выявляют пробелы в существующих исследованиях и указывают на перспективные направления для будущих работ.

Авторы Giselle Gonzalez Garcia и Christian Weilbach в своей работе "If the Sources Could Talk: Evaluating Large Language Models for Research Assistance in History" [2] исследуют, как LLMs могут быть использованы для анализа исторической памяти, представленной в различных типах документов.

Исследование подчеркивает, что LLMs могут быть обогащены векторными вложениями из специализированных академических источников, что делает их ценным инструментом для историков. Это позволяет исследователям использовать LLMs для анализа корпусов данных, состоящих из первичных и вторичных источников, а также их комбинаций. Такой подход может значительно ускорить процесс исследования, предоставляя более глубокое понимание контекста и связей между историческими событиями.

Авторы также сравнивают LLMs с традиционными поисковыми интерфейсами цифровых каталогов, такими как метаданные и полнотекстовый поиск, и обнаруживают, что LLMs предлагают более богатый, разговорный стиль взаимодействия. Это позволяет исследователям задавать вопросы и получать ответы, а также извлекать и

организовывать данные более эффективно. В частности, LLMs демонстрируют способность к семантическому поиску и рассуждению по задачам, специфичным для проблемы, что может быть применено к большим текстовым архивам, даже если они не были частью обучающих данных модели.

Таким образом, исследование подтверждает, что LLMs могут быть адаптированы к конкретным исследовательским проектам и использоваться исследователями для частных запросов, что открывает новые перспективы для гуманитарных наук. Это позволяет предположить, что в будущем LLMs могут стать неотъемлемым инструментом в арсенале гуманитарных исследователей, способствуя более глубокому и масштабному анализу исторических данных.

Настоящее исследование направлено на изучение потенциала применения языковых моделей в анализе социальных сетей. На первом этапе работы был проведен сбор и систематизация данных из различных социальных платформ. Этот этап включал в себя не только извлечение текстовой информации, но и ее предварительную обработку для дальнейшего анализа. На втором этапе исследования применялись различные методы анализа данных, включая современные языковые модели, для выявления тенденций, закономерностей и инсайтов, скрытых в больших объемах текстовой информации, генерируемой пользователями социальных сетей.

Первый этап исследования был направлен на формирование базы данных для анализа, что предполагало сбор сообществ и контента из социальной сети Вконтакте. Социальная сеть Вконтакте была выбрана в силу её популярности и богатства данных, а также доступности инструментов для сбора информации через API.

Основной фокус был сосредоточен на сообществах, посвященных научным и техническим тематикам (рис. 1), так как они представляют интерес для изучения специфических дискурсов и распространения знаний в социальных сетях. Для отбора сообществ использовались ключевые слова и теги, связанные с научными дисциплинами и технологическими отраслями. Критерии отбора включали активность сообщества, количество подписчиков и регулярность публикаций.

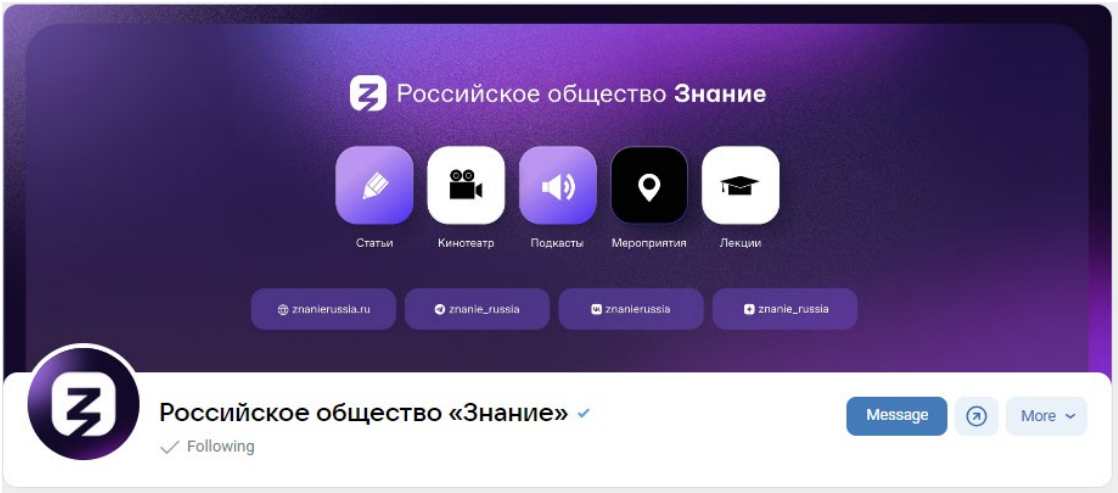

Рисунок 1. Пример сообщества Вконтакте «Российское общество Знание».

Следующий этап заключался в сборе контента в соответствующих сообществах. С помощью официального API Вконтакте были собраны посты и комментарии к ним в автоматизированном режиме. Для каждого сообщества был собран архив постов за определенный период времени, включая тексты, изображения и другие мультимедийные данные, а также метаданные, такие как дата публикации, количество лайков и репостов.

Важной частью процесса сбора данных было соблюдение этических норм и политики конфиденциальности. Сбор данных соответствует условиям использования API Вконтакте и не нарушает прав пользователей на приватность. Личная информация пользователей была исключена из сбора или анонимизирована в соответствии с принципами защиты данных.

Полученные данные были систематизированы и подготовлены к анализу. Это включало очистку данных от спама, неинформативных постов и дубликатов, а также их категоризацию по тематикам и форматам. Подготовленный таким образом датасет стал основой для следующего этапа исследования — анализа с помощью языковых моделей.

Используя методы обработки естественного языка (Natural Language Processing, NLP), был проведен анализ ключевых слов для выявления наиболее часто упоминаемых терминов и концепций в сообществах. Это позволило определить основные темы дискуссий и интересы пользователей. Для этого использовались алгоритмы автоматического извлечения ключевых слов, которые анализируют частоту слов и фраз, их взаимосвязь в тексте, а также контекст их использования.

Дополнительно к анализу ключевых слов, были использованы большие языковые модели для более глубокого понимания контента. Языковые модели, такие как GPT (Generative Pretrained Transformer), были обучены на больших корпусах текста и способны выявлять более сложные паттерны в данных, такие как сентимент, интенции и эмоциональную окраску сообщений. Эти модели использовались для анализа постов с целью выявления общих трендов и тематик, которые могли бы указать на формирование образов будущего у пользователей социальной сети.

Используя LLM, мы смогли не только определить ключевые темы дискуссий, но и выявить скрытые взаимосвязи между тематиками, а также проследить динамику интересов сообщества во времени. Это дало нам возможность понять, как пользователи социальной сети Вконтакте воспринимают и обсуждают научные и технические инновации, и какие представления о будущем они формируют через эти дискуссии.

4. Полученные результаты

Рассмотрим некоторые примеры результатов работы языковой модели. Например, для следующей новости: "На складах Amazon работают роботы-грузчики. Роботы Digit теперь носят коробки с товаром на складах американского маркетплейса. Таких работников в компании уже более 750 тысяч.". А также следующих комментариев:

- "Тормозные пока что роботы."
- "Не дай бог проявить неучтивость к этому работнику… Что будет… Восстание роботов."
- "Давно пора, зачем человек с мозгами должен мучиться на трудных работах. Человек должен работать умом, и для физической работы должен придумывать своим умом роботов."
- "Со временем их запретят, когда уровень безработицы достигнет пика и начнутся бунты со сломанными ножками и головами роботов."
- "Демонстрационный вариант. Носят пустые коробки. Мексиканцы будут проворнее."
- "С такими работниками посылки не дождешься."

Составлен следующий отклик: «Из комментариев видно, что многие комментаторы видят будущее, в котором роботы будут выполнять физическую работу

вместо человека, освобождая его для более творческих и умственных задач. Однако есть и опасения, что введение роботов может привести к увеличению безработицы и социальным проблемам. Некоторые комментаторы также выражают скептицизм по поводу эффективности роботов в настоящее время.».

Также для анализа тональности из комментариев выделялись ключевые слова, результат визуализации ключевых слов в виде облака слов представлен на рисунке 2.

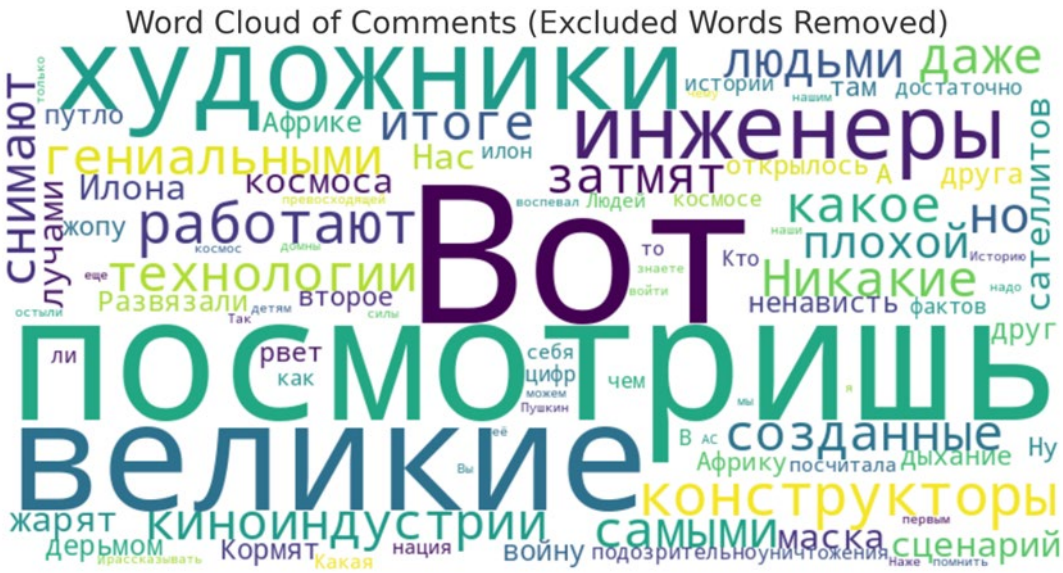

Рисунок 2. Визуализация ключевых слов в комментариях к постам в социальной сети ВКонтакте.

5. Выводы

В ходе исследования было продемонстрировано, как использование больших языковых моделей может ускорить и автоматизировать процесс анализа данных из социальных сетей. Это особенно актуально для задач, где необходимо обрабатывать большие объемы текстовой информации, как, например, извлечение тематически значимых постов на тему "образы будущего" и анализ комментариев под ними для понимания общественного восприятия этой темы.

С помощью продвинутых технологий обработки естественного языка, включая поиск по ключевым словам, классификацию текста, семантический анализ и синтез информации, такие модели способны эффективно фильтровать релевантные данные и предоставлять их в структурированном виде. Это позволяет исследователям и

аналитикам сосредоточиться на интерпретации результатов, а не на рутинной обработке данных.

В будущем можно ожидать дальнейшего усовершенствования этих технологий, что позволит проводить ещё более тонкий анализ, выявляя скрытые закономерности и тренды, предсказывая тенденции и помогая формировать стратегические решения на основе данных из социальных сетей. Это открывает новые горизонты для компаний, исследователей и политических аналитиков в понимании и формировании общественного мнения.

Благодарности

Проект № FSWW-2023-0012 реализован в Национальном исследовательском Томском политехническом университете по итогам отбора научных проектов, проведённых Министерством высшего образования и науки РФ и ЭИСИ.

Список литературы

1. Гонсалес Гарсия Г., Вайльбах К. Если бы источники могли говорить: оценка больших языковых моделей для помощи в исследованиях по истории [Электронный ресурс] // arXiv:2310.10808. – 2023. – Режим доступа: https://arxiv.org/pdf/2310.10808, свободный. – Загл. с экрана.
2. Хоу С., Жао Я., Лю Ю. и др. Систематический обзор литературы по большим языковым моделям для программной инженерии (LLM4SE) [Электронный ресурс] // arXiv:2308.10620. – 2023. – Режим доступа: https://arxiv.org/pdf/2308.10620, свободный. – Загл. с экрана.