CC BY 4.0

УДК 004.05
DOI: http://dx.doi.org/10.21686/1818-4243-2025-2-55-70

А.А. Погуда, Тапе Хабиб Жан Макс

Национальный исследовательский Томский государственный университет,
Томск, Россия

Разработка интеллектуальной системы для обработки слабоструктурированных данных: отраслевая структуризация и расширенный анализ информации, извлеченной из комментариев к видеороликам в социальных сетях

Научная актуальность исследования. *В эпоху стремительного увеличения объемов данных, генерируемых пользователями социальных сетей, анализ текстовых данных, таких как комментарии, становится одной из ключевых задач современной науки. Комментарии представляют собой ценный источник информации, позволяя выявлять общественные настроения, анализировать мнения пользователей и отслеживать социальные тренды. Однако из-за слабоструктурированного или полностью неструктурированного характера этих данных их обработка требует применения инновационных подходов.*

Целью данного исследования *является разработка интеллектуальной системы для обработки слабоструктурированных данных, получаемых из комментариев на видео в социальных сетях, с использованием алгоритмов структуризации, ориентированных на различные отрасли. Исследование направлено на создание эффективного метода анализа тональности, кластеризации и извлечения ключевых тем из комментариев с целью оценки воздействия видео-контента на аудиторию. В результате исследования будет предложен подход к автоматическому выделению и структурированию данных по отраслям, что позволит более точно и глубоко анализировать восприятие контента и его влияние на различные социальные и профессиональные сферы.*

Методы: *Разработка интеллектуальной системы для анализа слабоструктурированных данных требует применения инновационных методов и подходов, сочетающих в себе обработку естественного языка (NLP), алгоритмы машинного обучения и методы анализа больших данных. Эти методы включают: автоматическое извлечение данных через API, предварительную обработку, адаптированную для трех языков (французского, английского и русского), глубокий анализ настроений с помощью продукта Bert и вероятностного алгоритма для статистических расчетов, а также кластеризацию с помощью алгоритмов K-Means, DBSCAN и Agglomerative.*

Материалы *основываются на комментариях из социальных сетей (TikTok, Instagram, Twitter, Facebook, YouTube, Reddit, ВКонтакте) на русском, английском и французском языках. Для предобработки применялись библиотеки SpaCy и NLTK, а модель Hugging Face Transformers работала с предобученными моделями для анализа настроений. Использованы методы машинного обучения, включая кластеризацию и обработку естественного языка. Данные структурированы с помощью тематического моделирования и языковых моделей, реализованных с помощью Python-библиотек.*

Результаты исследования. *Разработка интеллектуальной системы для обработки слабо структурированных данных позволила улучшить анализ комментариев к видеороликам в социальных сетях благодаря комбинации различных моделей машинного обучения и алгоритмов. Результаты исследования позволили нам разработать прототип инструмента для анализа комментариев, который эффективно собирает и структурирует данные из различных социальных сетей. Эта структуризация данных привела к лучшей организации и повышенной доступности информации, что облегчило их использование. Используя методы обработки естественного языка (NLP), мы выявили ключевые темы и эмоции комментариев, проводя анализ настроений, который освещает основные эмоциональные тренды. Методы кластеризации, такие как K-средние, сгруппировали комментарии по схожим темам. Кроме того, мы создали визуализации, показывающие распределение настроений, что позволяет пользователям быстро интерпретировать данные. Интеграция методов визуализации преобразует сложные аналитические результаты в интуитивно понятные графики, что облегчает понимание взаимодействия пользователей с контентом. Таким образом, наша система оказывается эффективной для предоставления ценных инсайтов и оптимизации стратегий взаимодействия с аудиторией.*

Заключение. *Результаты исследования показали, что предложенный подход значительно улучшает точность классификации и структурирования слабо структурированных данных, особенно когда речь идет о комментариях, извлеченных из видеороликов в социальных сетях. Разработанная система использует алгоритмы обработки естественного языка для анализа данных с учетом их отраслевой принадлежности, что позволяет автоматически структурировать комментарии в зависимости от их содержания и проводить подробный анализ тональности. Эффективность данного подхода была подтверждена на примере анализа комментариев с различных социальных платформ, что продемонстрировало его способность извлекать и структурировать релевантную информацию, а также оценивать влияние видеороликов через реакции пользователей.*

Ключевые слова: *обработка слабоструктурированных данных, api, комментарии к видеороликам, социальные сети, структурирование данных, влияние видео.*

Alexey A. Poguda, Tape Habib Jean Max

National Research Tomsk State University, Tomsk, Russia

Development of an Intelligent System for Processing Semistructured Data: Industry Structuring and Advanced Analysis of Information Extracted from Comments to Video Clips in Social Networks

Scientific relevance of the study. *In the era of rapidly increasing volumes of data generated by social media users, analyzing textual data such as comments is becoming one of the key challenges of modern science. Comments are a valuable source of information, allowing us to identify public sentiment, analyze users' opinions, and track social trends. However, due to the semistructured or completely unstructured nature of these data, their processing requires innovative approaches.*

Purpose of research. *The aim of this research is to develop an intelligent system for processing semistructured data from comments on social media videos using structuring algorithms targeting different industries. The research aims to create an efficient method to analyze tone, clustering and extract key themes from comments in order to evaluate the impact of video content on the audience. The research will propose an approach to automatically extract and structure data by industry, which will allow for a more accurate and in-depth analysis of content perception and its impact on different social and professional domains.*

Methods. *Developing an intelligent system for analyzing semistructured data requires innovative methods and approaches that combine natural language processing (NLP), machine learning algorithms and big data analytics techniques. These methods include: automatic data extraction via API, preprocessing adapted for three languages (French, English and Russian), deep sentiment analysis using the Bert product and a probabilistic algorithm for statistical calculations, and clustering using K-Means, DBSCAN and Agglomerative algorithms.*

The materials *are based on comments from social networks (TikTok, Instagram, Twitter, Facebook, YouTube, Reddit, VKontakte) in Russian, English and French. SpaCy and NLTK libraries were used for preprocessing, and the Hugging Face Transformers model worked with pre-trained models for sentiment analysis. Machine learning techniques including clustering and natural language processing were used. Data was structured using topic modeling and language models implemented using Python libraries.*

The results of the study. *The development of an intelligent system for processing semistructured data has improved the analysis of comments on videos in social networks through a combination of various machine learning models and algorithms. The results of the study allowed us to develop a prototype of a comment analysis tool that effectively collects and structures data from various social networks. This data structuring led to better organization and increased accessibility of information, facilitating its utilization. By using natural language processing (NLP) methods, we identified key themes and emotions in the comments while conducting sentiment analysis that highlights major emotional trends. Clustering methods, such as K-means, grouped the comments by similar themes. Additionally, we created visualizations that show sentiment distribution, allowing users to quickly interpret the data. The integration of visualization techniques transforms complex analytical results into intuitive graphs, making it easier to understand user interactions with the content. Thus, our system proves effective in providing valuable insights and optimizing audience interaction strategies.*

Conclusion. *The results of the study showed that the proposed approach significantly improves the accuracy of classification and structuring of semistructured data, especially when it comes to comments extracted from social media videos. The developed system uses natural language processing algorithms to analyze the data with respect to its industry, which allows for automatic structuring of comments depending on their content and detailed tone analysis. The effectiveness of this approach was validated by analyzing comments from various social platforms, which demonstrated its ability to extract and structure relevant information, as well as assess the impact of videos through user reactions.*

Keywords: *semistructured data processing, api, video comments, social networks, data structuring, impact of the video.*

Введение

В настоящее время цифровые данные производятся в огромных количествах. Подавляющее большинство данных являются неструктурированными или слабоструктурированными. Учитывая, что лишь небольшая часть данных может считаться структурированной, необходимость структурирования неструктурированных данных имеет огромное значение. В зависимости от источника данных, операции или намерений можно работать с самыми разными данными. Традиционные методы структурирования данных имеют огромные ограничения, связанные с их жесткостью. В отличие от них, мы проводим исследования по разработке алгоритма структурирования данных, способного работать со слабоструктурированными данными. Постоянная цель структурирования данных – сделать данные, коллекции данных и, наконец, базы данных более доступными, удобными для использования и поиска. В данной работе проблема структурирования данных рассматривается с нескольких точек зрения. Нам хорошо известны методы обработки очень больших коллекций данных. Большие коллекции данных содержат множество файлов. Целевыми данными этого исследования действительно являются файлы. Слабоструктурированные данные представляют собой информацию, лишённую ясной организации или формата, в составе которой могут находиться разнообразные типы данных, такие как текстовые фрагменты, изображения, аудиофайлы, видеофайлы, таблицы и графики. Основная сложность анализа подобных данных заключается в их гетерогенности и отсутствии чётких

паттернов, что значительно усложняет процессы обработки, интерпретации и последующего использования. Именно поэтому создание математического алгоритма для работы с данными подобного характера приобретает значительную важность для современной науки и технологий, так как позволяет систематизировать, анализировать и извлекать полезную информацию из этих данных. Результаты таких решений могут применяться в различных сферах, включая медицину, финансы, маркетинг, науку, искусственный интеллект и городское планирование [1–3]. Это открывает новые горизонты для исследований и разработок, способствуя ускорению развития и технологического прогресса всего общества.

Анализ данных из комментариев в социальных сетях представляет собой сложную задачу, поскольку такие данные зачастую являются неструктурированными, содержат сленг, сокращения, эмодзи и ошибки, что требует применения передовых методов обработки естественного языка [4–5]. Дополнительные трудности возникают из-за многоязычности комментариев, что требует создания мультиязычных моделей, способных учитывать контекст и тональность текста [6]. Одной из ключевых задач является определение эмоционального окраса и контекста, где сложные выражения, ирония и сарказм могут стать источником ошибок в анализе [7]. Также важно выделение ключевых тем и трендов с помощью тематического моделирования, такого как LDA или более современных нейросетевых подходов [8]. В условиях огромного объема данных и их непрерывного поступления необходимо использовать масштабируемые алгоритмы, обеспечивающие обработку в режиме реального времени [9]. Решение этих задач позволяет разработать интеллектуальные системы, которые помогут извлекать ценную информацию для бизнеса, науки и управления, способствуя принятию более точных и оперативных решений.

Обоснование необходимости разработки интеллектуальной системы

Современный мир генерирует беспрецедентные объемы данных, большая часть которых относится к слабоструктурированным или неструктурированным форматам, таким как текстовые комментарии, изображения, аудиофайлы и видеоконтент. Например, данные из социальных сетей, отзывы пользователей о продуктах и онлайн-обсуждения предоставляют ценные сведения для анализа общественного мнения, изучения предпочтений и прогнозирования рыночных трендов. Однако обработка таких данных сопряжена с рядом сложностей, включая разнородность, отсутствие структуры, а также высокую скорость и объем их поступления [10]. Текущие технологии, такие как машинное обучение и методы обработки естественного языка (NLP), уже доказали свою эффективность при работе с текстами. [6, 12]. Тем не менее их возможности часто ограничены отсутствием учета отраслевой специфики. Например, для анализа данных из медицинской или финансовой сфер требуется адаптация алгоритмов под профессиональные термины и контексты. Помимо этого, глобальная природа социальных сетей создает вызовы для анализа многоязычных текстов, что требует создания мульти язычных моделей обработки.

Обзор литературы

В этом разделе мы обобщим ряд литературы, посвященной алгоритмам структурирования данных. Сюда входят исследования по следующим темам: структурирование данных, многоструктурные данные, адаптивное моделирование структуры данных и моделирование структуры данных на основе репрезентативности. Наконец, мы приводим характеристики слабоструктурированных данных.

За последние два десятилетия было разработано множество методов структурирования данных. В то время как ранний структурированный отчет о техниках и методах сосредоточен на традиционных методах концептуального моделирования, в новом отчете представлены несколько методов, которые находят все большее внимание как характеризующиеся адаптивностью к изменяющимся требованиям. Поиск новых методов был вызван, главным образом, необходимостью анализа больших, неполных, неоднородных и очень изменчивых текстовых и веб-данных. Методы можно отнести к одной из следующих групп: структурно анализирующие различные результаты поиска по заданным ключам и понятиям; структурно извлекающие общие структурные блоки и представителей из большого количества документов с похожими названиями, но несколько отличающихся по содержанию; или поддерживающие пользователя, структурно анализирующего неструктурированные или многоструктурные данные.

Начиная с работы Рамоса Гаргантильи и других [13], авторы подчеркивают ограничения существующих лингвистических структур при извлечении структурированной информации из текстов на естественном языке. Они вводят понятие «лингвистической схемы», которая повышает выразительность и снижает неявные ограничения, что способствует лучшему пониманию текста и поиску информации. Эта фундаментальная работа закладывает основу для дальнейших исследований сложностей структурирования данных.

Галкин и другие [15] сосредоточились на автоматическом извлечении веб-таблиц, выступая

за использование методов машинного обучения для анализа структуры таблиц. Их работа подчеркивает потенциал извлечения знаний из неструктурированных источников данных, предполагая, что интеграция машинного обучения со структурированием данных может значительно улучшить процесс извлечения.

Джунчилья и другие [16] решают проблему семантической неоднородности в управлении данными, предлагая стратифицированный подход к представлению данных. Этот подход позволяет легче интегрировать различные типы данных, тем самым упрощая сложности, связанные со структурированием данных

Танг и другие [17] рассматривают слабоконтролируемое обучение, подчеркивая необходимость новых методов моделирования в условиях отсутствия сильного контроля. Они выявляют ограничения традиционных подходов к моделированию данных и выступают за интеграцию различных культур моделирования для решения проблем, связанных с большими и сложными наборами данных.

Ку и Ким [18] представляют обширный обзор генеративных диффузионных моделей для структурированных данных, обсуждая применение этих моделей в различных задачах. Они выявляют проблемы, с которыми сталкиваются традиционные методы машинного обучения при работе со структурированными данными, и подчеркивают потенциал глубоких генеративных моделей для преодоления этих препятствий.

Лю и другие [19] исследуют распознавание структурных функций академических документов, подчеркивая важность контекстной информации для улучшения идентификации структурных функций. Их работа подчеркивает важность понимания структур, лежащих в основе академических документов, что может помочь в разработке алгоритмов для структурирования данных.

Статья «Базы данных экспериментов: создание новой платформы для исследований мета-обучения» авторов [21] представляет собой важный вклад в область машинного обучения, особенно в контексте анализа алгоритмов и их производительности на различных наборах данных. Основная идея работы заключается в предложении подхода, основанного на стандартизированном представлении и хранении метаданных из предыдущих экспериментов, что может послужить основой для формирования сообщества, занимающегося анализом алгоритмов обучения. Авторы отмечают, что несмотря на наличие множества эмпирических исследований, до сих пор остаются нерешёнными вопросы, касающиеся причин успеха или неудачи алгоритмов на определённых наборах данных. Это объясняется тем, что метаданные, генерируемые в ходе исследований в области машинного обучения, часто собираются и хранятся в разных форматах, что затрудняет их совместное использование и повторное применение. В этой связи предложенный подход, направленный на унификацию описания экспериментов и их хранение в едином репозитории, может значительно упростить доступ к данным. Критический анализ материала позволяет отметить, что предложенная система репозитория не только облегчает процесс поиска и сравнения алгоритмов, но и открывает новые возможности для теоретической интерпретации полученных результатов. Это может привести к новым инсайтам и улучшению существующих алгоритмов. Однако стоит отметить возможные ограничения, связанные с качеством и полнотой собранных данных. Если метаданные окажутся недостаточно информативными или неполными, это может повлиять на точность выводов, сделанных на их основе.

Статья «Понимание проблем качества данных в динамичных организационных средах – обзор литературы» авторов [22] представляет собой важный вклад в исследование качества данных, особенно в контексте быстро меняющихся организационных условий. Современные технологии, ускоряя рост объема, разнообразия и скорости данных, создают новые трудности для организаций, вынужденных оценивать качество данных и учитывать последствия их низкого качества. Основная идея работы заключается в том, что широкое разнообразие данных, в том числе ранее недоступных и сложных для анализа неструктурированных данных, требует пересмотра подходов к обеспечению их качества. Авторы указывают, что около 85% данных в организациях представляют собой неструктурированные данные, и в будущем они станут доминировать, что поставит перед организациями новые задачи по их эффективному управлению. Это ставит вопросы о том, как справиться с качеством таких данных и какие возможные риски могут возникнуть, если их качество будет неудовлетворительным. Оценив статью, можно выделить, что она подчеркивает важность дальнейших исследований в области качества данных, особенно с учетом неструктурированных данных. Однако, несмотря на подробный анализ существующей литературы, авторы не предлагают конкретных математических моделей или алгоритмов для структурирования и анализа таких данных. Это создает пробел в исследовании, так как практических инструментов для решения проблем качества данных в отношении неструктурированных данных все еще недостаточно.

Статья «A survey on pre-processing techniques: Relevant issues in the context of environmental data mining» авторов [25] представляет собой значимый вклад в исследование предварительной

обработки данных, особенно в рамках анализа экологической информации. В ней авторы акцентируют внимание на том, что качество исходных данных играет ключевую роль в аналитических процессах. Это связано с тем, что данные, полученные из реального мира, часто характеризуются наличием шума, ошибок, неопределенностей, а также избыточной или нерелевантной информации. Такое состояние данных делает процесс их качественной предварительной обработки не просто желательным, а крайне необходимым для создания достоверных моделей и принятия эффективных решений. Одной из основных идей статьи является то, что отсутствие четко структурированного подхода к предварительной обработке может привести к созданию малоэффективных моделей, особенно если они опираются на неполные или ошибочные данные. Эта проблема становится особенно критичной при работе с экологическими системами, которые представляют собой сложные и динамичные структуры с множеством взаимосвязанных элементов. Авторы подчеркивают, что для более глубокого понимания таких систем, а также для их эффективного управления, необходимо совершенствование подходов к моделированию и созданию систем поддержки принятия решений.

Вместе с тем, стоит отметить, что статья уделяет недостаточно внимания конкретным методам предварительной обработки данных, что может усложнить их применение в практических задачах. Также подчеркивается необходимость разработки более современных и мощных алгоритмов, способных эффективно работать с неструктурированными данными. Это направление является крайне актуальным и требует дальнейшего развития в области анализа данных.

Статья «Алгоритмы и подходы для обработки больших данных « авторов [23] представляет собой полезный анализ различных методов, применяемых для обработки и анализа больших объемов данных, что является крайне актуальным в условиях современного общества, где данные играют ключевую роль. Основной акцент работы сделан на проблемах обработки слабоструктурированных данных, которые часто встречаются в таких областях, как геномика, метеорология и экология. Авторы отмечают, что традиционные базы данных не способны эффективно справляться с обработкой сложных и объемных наборов данных, которые имеют разнообразные форматы и структуры. В связи с этим возникает потребность в разработке новых математических алгоритмов, способных извлекать полезные знания из данных. В статье рассматриваются различные методы, такие как генетические алгоритмы, машины опорных векторов, деревья решений и кластерный анализ, которые могут помочь выявить скрытые закономерности в больших данных. При оценке работы стоит отметить, что хотя обзор алгоритмов предоставляет полезную информацию о текущих тенденциях в области обработки больших данных, он не предоставляет глубокого анализа каждого из методов. Это может затруднить выбор наиболее подходящего алгоритма для решения конкретных задач. Кроме того, отсутствие практических примеров реального применения алгоритмов ограничивает возможность их внедрения в практическую деятельность.

Миттал и другие [20] предлагают новую структуру, которая объединяет семантические запросы с SQL, облегчая анализ неструктурированных данных в реляционных базах данных. Их подход представляет собой значительное достижение в интеграции анализа неструктурированных данных в существующие системы управления данными.

Статья «Cost-effective data structural preparation» авторства [33] рассматривает важные аспекты подготовки данных для эффективного анализа, особенно в контексте слабоструктурированных данных. Основная идея статьи заключается в том, что выбор структуры представления данных напрямую влияет на эффективность алгоритмов, которые будут применяться к этим данным. Это подчеркивает необходимость предварительной подготовки данных, чтобы они соответствовали требованиям конкретных аналитических процедур. Автор вводит концепцию независимости проектирования, что позволяет создавать алгоритмы, которые остаются эффективными независимо от выбранных представлений данных. Это особенно важно в условиях, когда ручная подготовка данных может быть затратной и время затратной. Статья предлагает алгоритм для задачи поиска сходства, который удовлетворяет свойству независимости проектирования. Это дает возможность расширить применение алгоритма на другие структурные варианты, что является значительным вкладом в область анализа данных. Одним из ключевых предложений является использование существующей онтологии для добавления структурной информации к неструктурированным наборам данных. Это позволяет улучшить эффективность алгоритмов, работающих с данными, путем аннотирования. Таким образом, статья не только подчеркивает важность структурирования данных, но и предлагает практические решения для достижения этой цели. Критически оценивая материал, можно отметить, что работа предоставляет полезные идеи и алгоритмы, которые могут значительно упростить процесс подготовки данных. Однако, хотя авторы и утверждают, что их алгоритмы могут быть применены к различным структурным вариантам, необходимо больше эмпирических

данных для подтверждения универсальности предложенных решений. Также стоит обратить внимание на потенциальные ограничения, связанные с использованием онтологий, так как их создание и поддержка могут потребовать значительных ресурсов.

Статья «Unsupervised Data Extraction from Computer-generated Documents with Single Line Formatting» авторов Бернштейна и Афанасенкова (2020) представляет собой значимый вклад в развитие технологий обработки слабоструктурированных данных, особенно в аспекте автоматизации извлечения информации из компьютерных документов. Основной целью исследования является разработка методологии, позволяющей осуществлять неуправляемое и полностью автоматическое извлечение данных независимо от формата документа. Эта проблема становится особенно актуальной в эпоху больших данных. Авторы выделяют три основные подхода, используемые в настоящее время для извлечения данных: ручной ввод, использование скриптов и специализированных инструментов. Они отмечают, что ручной ввод, хотя и применяется на практике, является трудозатратным, дорогостоящим и подверженным ошибкам, что делает его неэффективным для работы с большими объемами информации. Скрипты, с другой стороны, обеспечивают надежность и эффективность, но их создание и поддержка требуют значительных ресурсов. Эти ограничения подтверждают необходимость разработки более автоматизированных решений. Кроме того, в статье рассматриваются недостатки существующих технологий, включая методы, основанные на машинном обучении. Авторы подчеркивают, что современный прогресс в области искусственного интеллекта в значительной степени сосредоточен на контролируемом обучении, что требует существенного человеческого вмешательства, особенно на этапе настройки моделей [26]. Это ограничивает возможности автоматизации процессов, особенно при работе с документами, которые имеют сложное или произвольное форматирование. Предложенная методология направлена на преодоление этих барьеров. Она базируется на предположении, что форматирование документа отражает его исходную структуру данных. Такой подход открывает новые возможности для автоматизированного обмена данными, снижая необходимость в ручной работе и повышая общую эффективность обработки информации. Эта методология представляет собой перспективное решение для задач анализа слабоструктурированных данных и их интеграции в различные области применения.

В работе Грира [14], посвященной исследованию концептуальных деревьев, автор говорит о необходимости создания динамических структур, которые адаптируются к изменяющейся природе данных. Включая элементы времени и случайности, этот подход позволяет более тонко понять полу-структурированные данные, тем самым внося свой вклад в более широкий разговор об эффективных алгоритмах структурирования данных.

Статья Оксаны Комарницкой «Методы автоматизированного семантического анализа природноязычной информации» Она представляет собой важный вклад в изучение обработки слабоструктурированных данных, особенно в контексте анализа текстов, полученных из комментариев к видеороликам на социальных платформах. В центре исследования находится оценка методов латентно-семантического анализа (ЛСА) и их способности выявлять скрытые смысловые связи между словами и фрагментами текста. Автор отмечает, что ЛСА является полезным инструментом для обработки текстовых данных, но имеет и свои ограничения, такие как отсутствие когнитивных особенностей, присущих человеческому восприятию, а также игнорирование синтаксиса и морфологии. Этот момент особенно важен, учитывая, что в комментариях часто используются разнообразные и неформальные языковые структуры, что может значительно повлиять на точность анализа. Для решения этой проблемы Комарницкая предлагает интеграцию лингвистических методов с подходами из области статистики, что может привести к улучшению обработки естественного языка. Такой подход открывает новые возможности для создания интеллектуальных систем, которые будут учитывать как смысловые, так и синтаксические особенности текстов. Внедрение технологий, таких как явный семантический анализ, ЛСА, теории нечёткой логики и искусственного интеллекта, имеет огромный потенциал для решения задач автоматизированного анализа смысла. Однако при критической оценке статьи стоит отметить, что, хотя она дает полезные рекомендации для будущих исследований в области обработки слабоструктурированных данных, в ней не приводится конкретных примеров успешного применения предложенных методов в реальных ситуациях. Это оставляет некоторую лакуну в исследовании, так как практическое применение теоретических подходов могло бы значительно повысить актуальность и ценность работы [24].

Анализ существующих методов обработки слабоструктурированных данных

Сегодня существует несколько подходов к обработке слабоструктурированных данных, включая семантический анализ, классификацию, кластеризацию и методы обработки естественного языка (NLP). Эти методы позволяют эффективно структурировать и анализировать

данные, изначально не имеющие четкой структуры, такие как текстовые комментарии в социальных сетях, аудио или видеоматериалы. Например, семантический анализ используется для извлечения смысловой информации из текстов, выявления скрытых значений и взаимосвязей, что особенно актуально для работы с неформальными языковыми конструкциями, как жаргон или эмодзи. Тем не менее, данный метод имеет свои ограничения, связанные с трудностью обработки многозначных фраз и сарказма. В свою очередь, кластеризация и классификация помогают группировать данные на основе их сходства, что делает их полезными для выявления паттернов в больших объемах информации. Однако выбор метода классификации или кластеризации зависит от конкретной задачи и характеристик данных, требующих точной настройки алгоритмов[27–28].

Современные методы обработки естественного языка, такие как NLP, включая анализ настроений и распознавание именованных сущностей, также играют важную роль в извлечении ключевых аспектов текста и оценки эмоциональной окраски. Эти методы позволяют анализировать не только содержание, но и контекст данных, что критически важно для понимания реальных намерений пользователей. Вместе с тем, применение машинного обучения и глубокого обучения, например, с использованием моделей вроде BERT, значительно повышает точность обработки сложных языковых конструкций, включая сарказм и многозначность, что расширяет возможности для более глубокого анализа текста. Тем не менее, такие подходы требуют значительных вычислительных ресурсов и данных для обучения моделей. Однако, эти методы могут давать менее точные результаты по сравнению с более специализированными техниками, такими как анализ настроений или тематическое моделирование. Применение этих методов в различных областях, таких как медицина, финансы, маркетинг и другие, открыло новые возможности для улучшения понимания и интерпретации данных. Например, в медицине это может привести к более точному анализу отзывов пациентов, а в маркетинге – к лучшему пониманию предпочтений потребителей. Однако каждый метод имеет свои преимущества и ограничения, и их успешное применение зависит от правильного выбора подхода в зависимости от задачи [30–31]. Таким образом, подходы к обработке слабоструктурированных данных продолжают развиваться, и необходимость постоянного усовершенствования этих методов является важнейшей задачей для решения возникающих проблем в анализе данных. Это подчеркивает важность разработки новых и более эффективных методов, способных справиться с различными вызовами, которые эти данные представляют. Такой процесс прогресса в области анализа слабоструктурированных данных является ключевым фактором для дальнейших достижений в их использовании и интерпретации [32]

Методология

Методология данного исследования заключается в разработке интеллектуальной системы для обработки и анализа слабоструктурированных данных, полученных из комментариев пользователей в социальных сетях. В основе подхода лежат современные методы обработки естественного языка, машинного обучения и специфические подходы для работы с такими данными. Описание основных этапов разработки системы и использованных инструментов приведено ниже

Архитектура интеллектуальной системы

В этой секции рассматривается архитектура интеллектуальной системы, разработанной для обработки и анализа слабо структурированных данных, полученных из комментариев к видеороликам в социальных сетях. Система состоит из нескольких ключевых компонентов, каждый из которых выполняет свою роль в процессе извлечения, обработки и анализа информации.

Компоненты системы

Сбор данных:

- Система использует API различных социальных сетей, таких как TikTok, Instagram, Twitter, Facebook, YouTube, Reddit и ВКонтакте, для извлечения комментариев к видеороликам.
- Для получения данных в реальном времени применяются библиотеки, такие как praw для Reddit, tweepy для Twitter и instaloader для Instagram.

Предобработка данных:

- Комментарии очищаются от лишних символов, URL, хештегов и упоминаний с помощью регулярных выражений.
- Для улучшения качества текстовых данных используются инструменты обработки естественного языка (NLP), такие как spaCy, которые выполняют лемматизацию и удаление стопслов.

Структурирование комментариев:

- На этом этапе комментарии организуются в структурированный формат, что позволяет легче анализировать их содержание.
- Используются методы, такие как создание таблиц или баз данных, где каждый комментарий связывается с метаданными, такими как дата публикации, автор и контекст видео. Это помогает в дальнейшей аналитике и визуализации.

Таблица сравнительного анализа различных методов структуризации слабоструктурированных данных
Table of comparative analysis of different methods of structuring semistructured data

Метод	Описание	Тип данных	Алгоритмы и инструменты	Преимущества	Недостатки	Области применения
Кластеризация [34, 18]	Группировка данных на основе сходства без предварительной метки	Текст, изображения, звук	K-means, DBSCAN, иерархическая кластеризация	Простота, выявление скрытых паттернов, автоматизация обработки	Требует предварительного выбора параметров (например, числа кластеров)	Анализ текста, обработка изображений, обработка аудиофайлов
Обработка на основе графов [25]	Использование графов для представления взаимосвязей между объектами	Текст, изображения, сети	Графовые нейронные сети (GNN), алгоритм PageRank	Моделирование сложных связей, высокая интерпретируемость	Требует значительных вычислительных ресурсов, сложность построения графа	Социальные сети, анализ данных, биоинформатика
Машинное обучение [37, 39]	Применение алгоритмов машинного обучения для выявления паттернов	Текст, изображения, звук	Рандомные леса, SVM, нейронные сети, ансамбли	Хорошая точность, возможность работы с большими данными	Требует большого объема данных для обучения, сложность настройки моделей	Прогнозирование, классификация, анализ текста, обработка изображений
Обработка естественного языка (NLP) [37]	Использование методов для работы с текстами и извлечения информации	Текст	TF-IDF, Word2Vec, BERT, LSTM	Умение работать с неструктурированным текстом, высокая точность	Сложность в обучении и настройке моделей, ресурсоемкость	Анализ тональности, извлечение сущностей, автоматический перевод
Сетевые методы [38]	Использование сетевых технологий для моделирования и анализа данных	Текст, изображения, сети	NMF (неотрицательное матричное разложение), PCA	Моделирование взаимосвязей, выявление скрытых закономерностей	Трудность интерпретации моделей, чувствительность к параметрам	Рекомендательные системы, анализ больших данных, медицина
Правила ассоциации [34]	Выявление взаимосвязей между различными элементами данных	Текст, покупательские данные	Алгоритм Apriori, FP-growth	Легкость в интерпретации, находит полезные связи в данных	Могут быть неинформативными или трудными для интерпретации в больших наборах данных	Рекомендательные системы, маркетинг, анализ данных
Инструменты глубокого обучения (DL) [38]	Использование нейронных сетей для глубокого анализа данных	Текст, изображения, видео	CNN, RNN, GAN, автоэнкодеры	Высокая точность, способность извлекать признаки из больших данных	Огромные требования к вычислительным ресурсам, сложность в обучении	Обработка изображений, видеоанализ, генерация текста, распознавание речи

Анализ настроений:

• Чтобы определить эмоциональную окраску комментариев, система применяет модель глубокого обучения на основе трансформеров, например, nlptown/bert-base-multilingual-uncased-sentiment.

• Модель классифицирует комментарии по шкале от 1 (очень негативный) до 5 (очень позитивный), что позволяет оценить общее восприятие видео.

Кластеризация:

• Комментарии группируются с помощью алгоритмов машинного обучения, таких как KMeans, DBSCAN и агломеративная кластеризация. Это помогает выявить скрытые паттерны и повторяющиеся темы в комментариях.

Визуализация данных:

• Результаты анализа и кластеризации визуализируются с помощью библиотек matplotlib и plotly.

• Создаются графики, иллюстрирующие распределение настроений, и круговые диаграммы, что помогает пользователям лучше понять эмоциональную реакцию на видео.

Отчет об воздействии видео:

• На основе собранных данных и проведенного анализа формируется отчет, который подчеркивает влияние конкретного видеоконтента.

• Отчет включает в себя ключевые метрики, такие как общее количество комментариев, средний уровень настроений, а также выявленные темы и паттерны.

• Также анализируются корреляции между характеристиками видео (такими как длительность, жанр и время публикации) и реакцией аудитории, что позволяет оценить, как различные факторы влияют на восприятие контента.

Предлагаемый алгоритм: Математическая основа

В основе предложенного алгоритма лежат методы математической обработки данных, машинного обучения и алгоритмы кластеризации. Ниже приведено описание математических аспектов основных этапов алгоритма.

Шаг 1: Извлечение комментариев

$C_i = \{c_{i1}, c_{i2}, ..., c_{im}\}$ через API. (1)

Объединить комментарии: $C = \bigcup_{i=1}^{k} C_i$

API социальных сетей для извлечения комментариев.

Результат: множество комментариев

$$C = \{c_1, c_2, ..., c_n\} \quad (2)$$

Шаг 2: Предобработка комментариев

$$\boldsymbol{P(C) = \{\text{Preprocess}(c_i) \mid c_i \in C\}} \quad (3)$$

Объяснение и демонстрация

Общая формула алгоритма предварительной обработки комментариев, как описано, представляет собой последовательность трансформаций f_k, применённыхк набору необработанных комментариев $C = \{c_1, c_2, ..., c_n\}$. Формула может быть записана следующим образом:

$$C' = f_K\left(f_{K-1}\left(\ldots f_2\left(f_1(C)\right)\ldots\right)\right) \quad (4)$$

где: $C = \{c_1, c_2, ..., c_n\}$ – исходный набор необработанных комментариев,

f_K – это K-я трансформация, применённая на шаге обработки,

C' – это набор комментариев после применения всех последовательных трансформаций.

Индивидуальная трансформация: Для каждого комментария $c_i \in C$ применяется последовательность функций $f_1, f_2, ..., f_k$ к исходному тексту t_i. Результат после k-го шага выражается как:

$$T_i^{(k)} = f_k\left(T_i^{(k-1)}\right), \text{ где } T_i^{(0)} = t_i \quad (5)$$

Структурированный комментарий: Финальный структурированный комментарий $c_i^{\text{processed}}$ формируется путем объединения предварительно обработанного текста и метаданных:

$$c_i^{\text{processed}} = \{\text{'Original Comment':}t_i, \text{'Processed Comment':}T_i^{(K)}, \text{'Metadata':}c_i[\text{"Metadata'}]\}$$

Глобальная трансформация: Для всех комментариев C результирующий список структурированных комментариев P задается как:

$$P = \left\{f_6\left(t_i, c_i[\text{'Metadata'}]\right) \mid c_i \in C\right\} \quad [11](6)$$

Детализированные шаги преобразований f_k

1. Очистка (f_1): Удаление ненужных элементов, таких как URL, хештеги и упоминания.

$$f_1(t) = \text{remive}_{\text{pattems}}(t)$$

2. Преобразование в строчные буквы (f_2): Преобразование текста в строчные буквы.

$$f_2(t) = \text{to lowercase}(t)$$

3. Токенизация (f_3): Разбиение текста на токены.

$$f_3(t) = \text{tokenuze}(t)$$

4. Лемматизация и фильтрация (f_4): Преобразование токенов вих леммы, удаление стоп-слов ипунктуации.

$$f_4(T) = \left\{\text{lemme}(t) \mid t \in T, t \notin \text{stopwords}, t \notin \text{punctuation}\right\} \quad (7)$$

5. Реконструкция (f_5): Сборка токенов обратно в строку.

$$f_5(T) = \text{join}(T, \text{sep} = \text{""}) \quad (8)$$

6. Финальная структура (f_6): Объединение предварительно обработанного текста с метаданными. Финальный результат

Для комментария c_i, предварительно обработанный текст будет:

$$T_i^{(K)} = f_5\left(f_4\left(f_3\left(f_2\left(f_1(t_i)\right)\right)\right)\right) \quad (9)$$

А сам структурированный комментарий:

$$c_i^{\text{processed}} = f_6\left(t_i, c_i\left['\text{Metadata'}\right]\right) \quad [6]\ (10)$$

Для всех комментариев результатом будет:

$$\boldsymbol{P} = \left\{\boldsymbol{c}_i^{\text{processed}} \boldsymbol{c}_i \in \boldsymbol{C}\right\} \quad \textbf{(2)}$$

Шаг 3: Анализ тональности с использованием BERT

Использовать модель пIptown/bert-base-multilingual-uncased-sentiment f_{BERT} которая возвращает распределение вероятностей s_i по 5 классам тональности:

$$s_i = f_{\text{BERT}}\left(p_i\right), s_i = \left[s_{i1}, s_{i2}, \ldots, s_{i5}\right] \quad [36](11)$$

Предсказать тональность как индекс класса с максимальной вероятностью:

$$\hat{s}_i = \arg\max_{j \in \{1,2,3,4,5\}} s_{ij} \quad (12)$$

где $\hat{s}_i$ – предсказанная тональность для комментария p_i и s_{ij} – вероятность принадлежности комментария p_i.

Классы тональности: j = 1: очень негативный; j = 2: негативный; j = 3: нейтральный; j = 4: позитивный; j = 5: очень позитивный.

Шаг 4: Векторизация комментариев

Преобразовать предварительно обработанных комментариев p_i в векторы x_i с использованием метода TF-IDF или аналогичного:

$$x_i = \text{Vectorize}(p_i) \quad (13)$$

Создать матрицу признаков X:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad (14)$$

Шаг 5: Обучение модели ML

Обучить модель машинного обучения f_{ML} для предсказания тональности, используя X и предсказания BERT $\hat{S} = \left\{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_n\right\}$:

$$f_{\text{ML}} = \text{TrainModel}\left(X, \hat{S}\right) \quad (15)$$

Шаг 6. Предсказание для новых комментариев

Для нового комментария c_{new}

$p_{\text{new}} = \text{Preprocess}(c_{\text{new}})$ Провести предобработку

$x_{\text{new}} = \text{Vectorize}(p_{\text{new}})$ Преобразовать в вектор

$\hat{s}_{\text{new}} = f_{\text{ML}}\left(x_{\text{new}}\right)$ Предсказать с помощью модели ML

Шаг 7. Расчет воздействия видео на основе тональности

Воздействие видео I_{video} можно вычислить как сумму вкладов всех обработанных комментариев в зависимости от их тональности:

$$I_{\text{video}} = \sum_{i=1}^{n}\left(w_i * \hat{s}_i\right) \quad (16)$$

где: w_i – вес каждого комментария $c_i^{\text{processed}}$, который может зависеть от параметров, таких как длина комментария или вовлеченность.

Итоговый расчет воздействия Воздействие видео можно также представить как взвешенное суммирование тональностей всех комментариев:

$$I_{\text{video}} = \sum_{j=1}^{5}\left(p_j * s_j\right)$$

I_{video} =Argmax(P1 + P2, P3, P4 + P5)

p_j пропорция комментариев, относящихся к классу j и s_j предсказанная тональность комментария

$p_{\text{positive}} = \frac{\sum_{i=1}^{n}\left(s_{i4} + s_{i5}\right)}{n} * 100\%$ p_{positive} процент позитивных комментариев,

$p_{\text{negative}} = \frac{\sum_{i=1}^{n}\left(s_{i1} + s_{i2}\right)}{n} * 100\%$; процент негативных комментариев,

$p_{\text{neutral}} = \frac{\sum_{i=1}^{n}\left(s_{i1} + s_{i2}\right)}{n} * 100\%$; процент нейтральных комментариев.

Условия для определения влияния

$p_{\text{positive}} > p_{\text{negative}} + p_{\text{neutral}}$, то влияние положительное.

$p_{\text{negative}} > p_{\text{positive}} + p_{\text{neutral}}$, то влияние отрицательное.

$p_{\text{neutral}} \geq 50\%$, то влияние нейтральное.

Практический этап и результаты работы

Описание программы

Программа представляет собой веб-приложение, созданное с использованием фреймворка Dash, который упрощает разработку интерактивных веб-приложений для анализа данных. Приложение предназначено для извлечения комментариев из различных социальных плат-

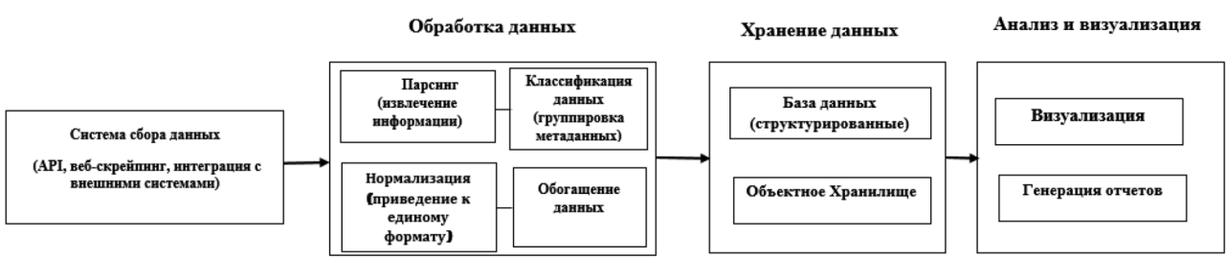

Рис. 1. Архитектура системы

Fig. 1. System architecture

форм, анализа их эмоционального окраса, группировки в кластеры и оценки влияния видео на аудиторию. Основная цель – предоставить пользователю удобный инструмент для работы с большими объемами данных из социальных медиа.

Основные характеристики

Приложение отличается интуитивно понятным интерфейсом, созданным с использованием Dash и Bootstrap. Оно включает интерактивные элементы, такие как выпадающие списки, поля ввода и кнопки. Программа поддерживает интеграцию с несколькими платформами, включая TikTok, Instagram, Twitter, Facebook, YouTube, Reddit и ВКонтакте, извлекая комментарии через API каждой из них. Анализ чувств осуществляется с помощью алгоритмов, которые оценивают эмоциональную окраску комментариев и визуализируют результаты в виде графиков. Пользователи могут настраивать параметры кластеризации, выбирая количество кластеров и тип алгоритма (например, KMeans или DBSCAN).

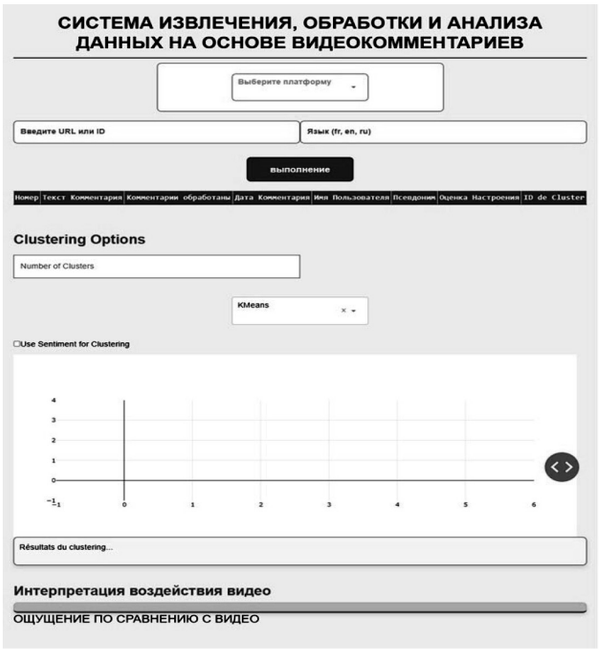

Рис. 2. Разработка графического интерфейса системы

Fig. 2. Development of the graphical interface of the system

Преимущества дизайна

Программа построена модульно, что обеспечивает удобство ее сопровождения и расширения. Каждый функциональный блок (извлечение данных, анализ чувств, кластеризация) структурирован отдельно. Благодаря интерактивности пользователи могут в реальном времени менять платформы, вводить URL и настраивать параметры анализа. Визуализация данных позволяет легко интерпретировать результаты анализа, делая приложение доступным даже для непрофессионалов. Масштабируемость архитектуры позволяет добавлять поддержку новых платформ и функций без значительных изменений.

Функциональность

Приложение предоставляет пользователю комплексный набор инструментов: извлечение комментариев с заданного URL, анализ их эмоционального содержания, групповую кластеризацию и оценку влияния видео на аудиторию. Результаты анализа представлены в таблице и графиках, что облегчает их восприятие. Интерфейс поддерживает управление параметрами

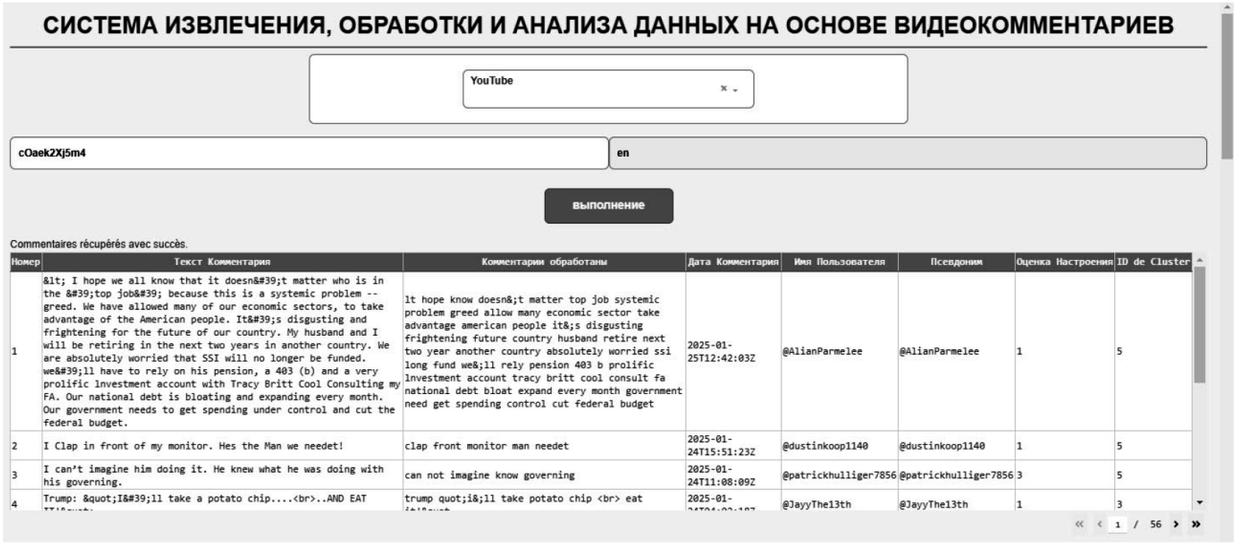

Рис. 3. Результат структурирования комментариев

Fig. 3. Result of comment structuring

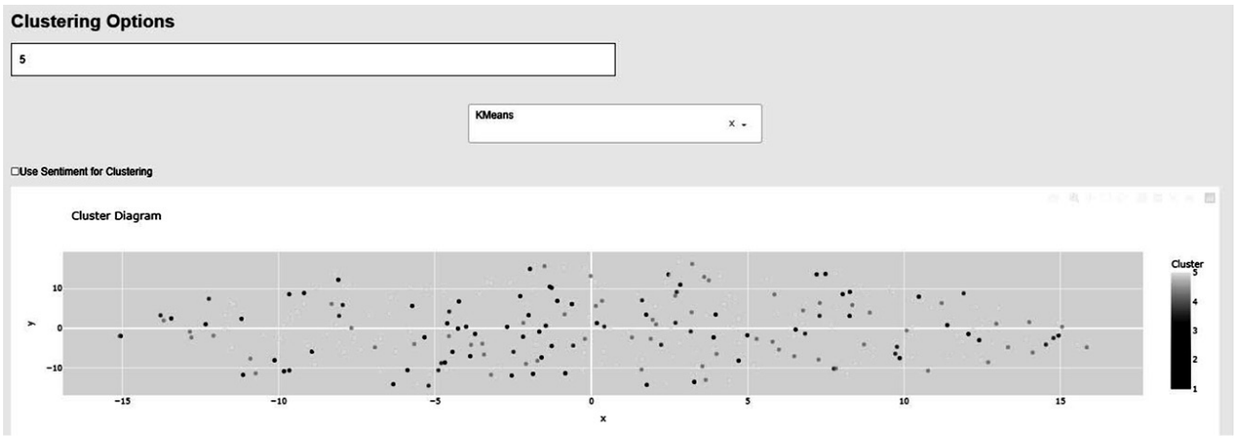

Рисунок 4. Результаты кластеризации обработанных комментариев в соответствии с оценками
Figure 4. Results of clustering of the processed comments according to scores

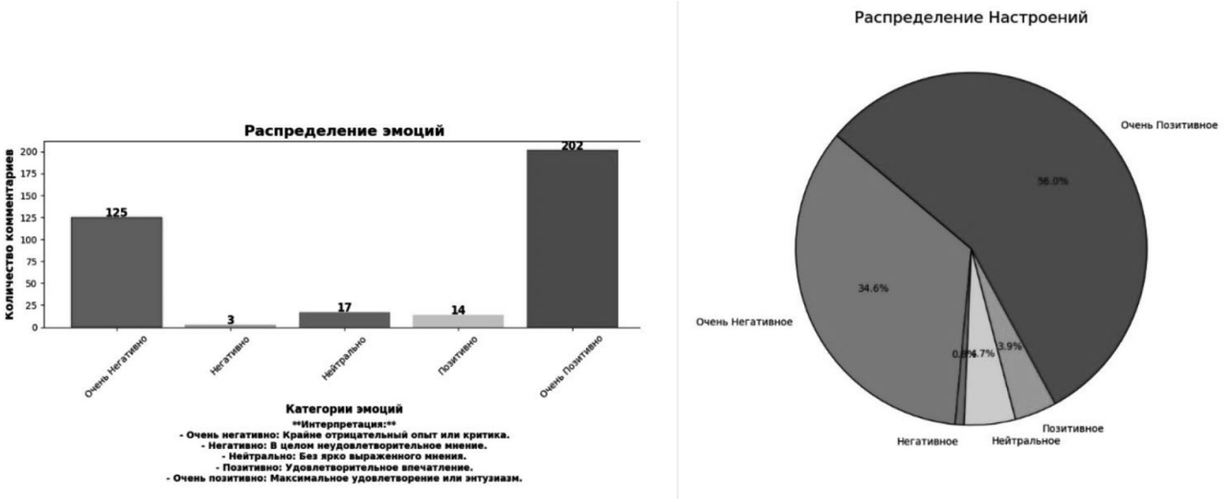

Рисунок 5. Распределение результатов анализа настроения обработанных комментариев
Figure 5. Distribution of the results of sentiment analysis of the processed comments

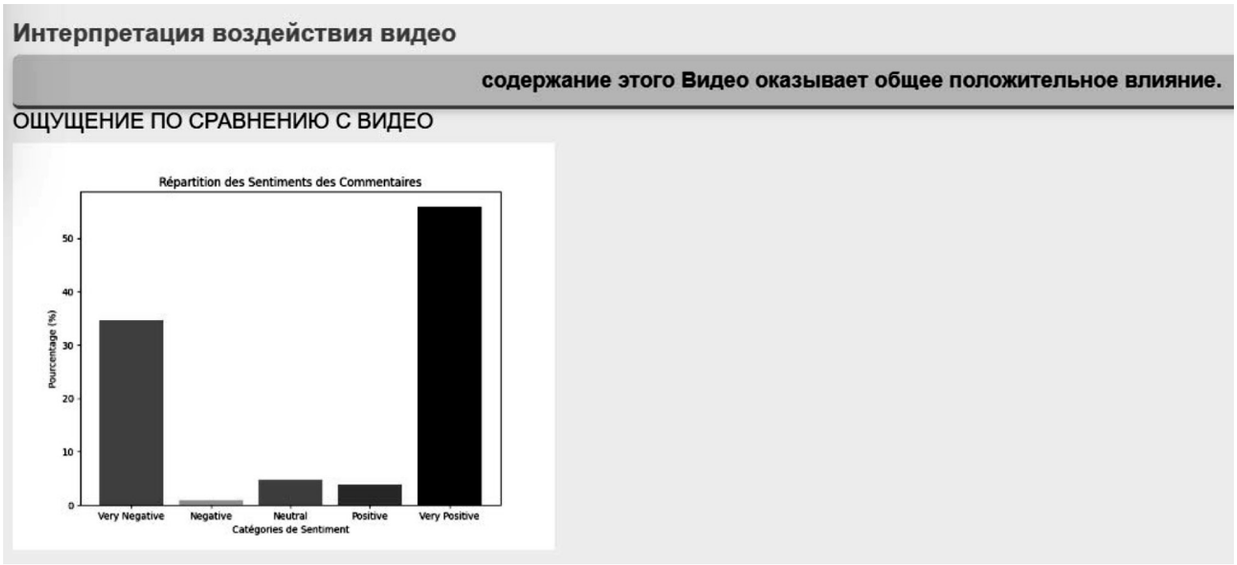

Рисунок 6. Результат воздействия видео с обработанными комментариями
Figure 6. The result of exposure to the video with processed comments

кластеризации и обеспечивает динамическую загрузку данных для удобства работы. Также предусмотрены информативные сообщения об ошибках, помогающие пользователю в случае возникновения проблем.

Экспериментальные данные: В данном исследовании мы сосредоточились на сборе комментариев пользователей к видеоролику на YouTube, в котором Дональд Трамп опубликованном на YouTube-странице Forbes Breaking News, одного из крупнейших американских СМИ, на основе комментариев. Цель – извлечь и структурировать данные для последующего анализа настроений и изучения конкретных ситуаций, а также определить влияние видео. Входные данные; Платформа – YouTube; URL или ID –ID видео; Язык –английский; Алгоритм кластеризации – Kmeans; количество –кластеров:5

В заключение, эта программа предлагает комплексное решение для анализа и визуализации комментариев в социальных сетях, позволяя также оценивать влияние видео на его аудиторию, одновременно отображая обработанные комментарии в таблице. Удобный интерфейс и продвинутые функции анализа данных делают приложение очень полезным для пользователей, желающих понять реакции на их контент. Сочетание моделей, алгоритмов и методов, использованных для создания этой системы, позволило добиться замечательных результатов.

Обсуждение

В данном исследовании мы подчеркнули важность анализа комментариев в социальных сетях и их решающую роль в восприятии контента. Наша программа предлагает сложные инструменты, которые позволяют пользователям извлекать, анализировать и визуализировать комментарии, что обеспечивает более глубокое понимание реакций аудитории. Особенно важен анализ настроений, который помогает уловить эмоциональную тональность отзывов, выходя за рамки простых чисел. Это оказывается особенно полезным для создателей контента и специалистов по маркетингу, которые могут корректировать свои стратегии в зависимости от эмоций и ожиданий своей аудитории, способствуя более искреннему и актуальному взаимодействию.

Кроме того, наша программа включает функцию группировки, которая классифицирует комментарии по темам или эмоциям, позволяя выявлять основные тенденции и проблемы пользователей. Этот метод не только предоставляет ценные insights о реакциях аудитории, но и помогает компаниям и создателям принимать обоснованные стратегические решения. Графическое представление данных делает результаты анализа более доступными и понятными, что облегчает быструю оценку восприятия контента. Интеграция обработанных комментариев в отчеты укрепляет доверие пользователей к полученным результатам, что является ключевым элементом для оптимизации взаимодействия с аудиторией. Этот подход открывает путь к разработке стратегий контента на основе конкретных данных, позволяя глубже понять социальные динамики в онлайн-пространстве.

Заключение

Создание этой интеллектуальной системы для обработки слабо структурированных данных стало значительным шагом вперед в анализе информации, извлекаемой из комментариев к видеороликам в социальных сетях. Интегрируя различные модели и алгоритмы, такие как методы машинного обучения, обработка естественного языка и анализ настроений, наша система эффективно структурирует и анализирует сложные данные. Этот подход позволяет глубже понять мнения и эмоции пользователей, а также выявить ключевые темы и тенденции, возникающие в ходе онлайн-взаимодействий. Применение отраслевой структуризации в сочетании с детальным анализом информации не только улучшает качество собранных данных, но и делает их более актуальными для пользователей. Это способствует более целенаправленному и искреннему взаимодействию с аудиторией, предоставляя ценную информацию для создателей контента и специалистов по маркетингу. Кроме того, интеграция инструментов визуального анализа облегчает быстрое понимание результатов, что особенно важно в постоянно меняющемся цифровом окружении.

В будущем развитие этой платформы может быть дополнено интеграцией новых источников данных и оптимизацией существующих алгоритмов, что еще больше повысит точность и полноту анализа. Таким образом, этот инновационный подход не только способствует вовлечению пользователей, но и создает надежную основу для разработки стратегий контента, основанных на конкретных данных, что позволяет лучше понять социальные динамики в сети и предвосхитить ожидания аудитории.

Литература

1. Кравченко Д.Ю. Модель онтологии знаний для интеллектуальных систем обработки и анализа текстов // Известия ЮФУ. Технические науки. 2024. № 2. С. 38–50.

2. Гулай А.В., Зайцев В.М. Модели знаний как когнитивный компонент системного построения интеллектуальных технологий // Развитие науки и технологий в эпоху глобальной трансформации. Петрозаводск: МЦНП «Новая наука», 2023. С. 158–191.

3. Журавков М.А. Технологии искусственного интеллекта и интеллектуальные системы компьютерного моделирования и инженерных расчетов [Электрон. ресурс]. Минск: Белорусский государственный университет, 2024. 177 с. Режим доступа: https://elib.bsu.by/bitstream/123456789/309072/1/Технологии%20искусственного%20интеллекта%20и%20интеллектуальные%20системы.pdf.

4. Kaplan A.M., Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media // Business Horizons. 2010. Т. 53. № 1. С. 59–68.

5. Cambria E., Schuller B., Xia Y., Havasi C. New avenues in opinion mining and sentiment analysis // IEEE Intelligent Systems. 2017. Т. 28. № 2. С. 15–21.

6. Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019.

7. Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. 2008. Т. 2. № 1–2. С. 1–135.

8. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. Т. 3. С. 993–1022.

9. Chen M., Mao S., Liu Y. Big data: A survey // Mobile Networks and Applications. 2014. Т. 19. № 2. С. 171–209.

10. Gandomi A., Haider M. Beyond the hype: Big data concepts, methods, and analytics // International Journal of Information Management. 2015. Т. 35. № 2. С. 137–144.

11. Smith J., Brown T. Metadata management for large-scale datasets // Journal of Information Systems. 2019. Т. 25. № 3. С. 120–135.

12. Voigt P., Von dem Bussche A. The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer International Publishing, 2017.

13. Ramos Gargantilla J.A., Mora J., Aguado de Cea G. Enhancing the expressiveness of linguistic structures. 2012.

14. Greer K. Concept Trees: Building Dynamic Concepts from Semi-Structured Data using Nature-Inspired Methods. 2014.

15. Galkin M., Mouromtsev D., Auer S. Identifying Web Tables – Supporting a Neglected Type of Content on the Web. 2015.

16. Giunchiglia F., Zamboni A., Bagchi M., Bocca S. Stratified Data Integration. 2021.

17. Tang C., Yuan G., Zheng T. Weakly Supervised Learning Creates a Fusion of Modeling Cultures. 2021.

18. Koo H., Eun Kim T. A Comprehensive Survey on Generative Diffusion Models for Structured Data. 2023.

19. Liu J., Zhao Z., Wu N., Wang X. Research on the structure function recognition of PLOS [Электрон. ресурс]. 2024. Режим доступа: ncbi.nlm.nih.gov.

20. Mittal A., Bheemreddy A., Tao H. Semantic SQL – Combining and optimizing semantic predicates in SQL. 2024.

21. Vanschoren J., Blockeel H., Pfahringer B., Holmes G. Experiment Databases: Creating a New Platform for Meta-Learning Research. 2008.

22. Anstiss S. Understanding data quality issues in dynamic organisational environments – a literature review. 2012.

23. Yadav C., Wang S., Kumar M. Algorithm and approaches to handle large Data – A Survey. 2013.

24. Комарницкая О. Методы автоматизированного семантического анализа естественноязыковой информации. 2018.

25. Leskovec J., Rajaraman A., Ullman J.D. Mining of Massive Datasets. 3rd ed. Cambridge University Press, 2014.

26. Bernstein V., Afanassenkov A. Unsupervised Data Extraction from Computer-generated Documents with Single Line Formatting. 2020.

27. Жоули М., Ганем Р., Аззуза М. Семантический анализ больших данных: вызовы и возможности // Исследования в области больших данных. 2020. Т. 18. № 4. С. 115–130.

28. Чжанг Дж., Ли В., Лю Ц. Обзор алгоритмов машинного обучения для классификации больших данных // Журнал машинного обучения. 2021. Т. 38. № 7. С. 925–940.

29. Нгуен Л., Тран Т., Нгуен Д. Классификация и кластеризация данных: методы и приложения // Журнал вычислительного интеллекта. 2019. Т. 31. № 1. С. 45–58.

30. Чэн Х., Ли Т., Чжан Х. Применение кластеризации данных в здравоохранении и финансах // Журнал научных исследований данных. 2020. Т. 25. № 3. С. 200–214.

31. Ли Д., Парк Дж., Ким С. Роль машинного обучения в маркетинговой аналитике // Научные исследования в области маркетинга. 2022. Т. 39. № 2. С. 189–204.

32. Василенко А., Фролов А., Макаров П. Современные методы обработки неструктурированных данных в новых технологиях // Журнал новых технологий. 2021. Т. 14. № 1. С. 112–124.

33. Chodpathumwan Y. Cost-effective data structural preparation. 2018.

34. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd ed. Elsevier, 2011.

35. Aggarwal C. C., Reddy C. K. Data Clustering: Algorithms and Applications. CRC Press, 2014.

36. Cambria E., Schuller B., Xia Y., Havasi C. New avenues in opinion mining and sentiment analysis // IEEE Intelligent Systems. 2017. T. 28. № 2. C. 15–21.

37. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008.

38. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016.

39. Bishop C. M. Pattern Recognition and Machine Learning. Springer, 2006.

40. Pennington J., Socher R., Manning C.D. GloVe: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014. C. 1532–1543.

References

1. Kravchenko D.Yu. Model of knowledge ontology for intelligent systems of text processing and analysis. Izvestiya YUFU. Tekhnicheskiye nauki = Bulletin of SFedU. Technical sciences. 2024; 2: 38–50. (In Russ.)

2. Gulay A.V., Zaytsev V.M. Knowledge models as a cognitive component of the systemic construction of intelligent technologies. Razvitiye nauki i tekhnologiy v epokhu global'noy transformatsii = Development of science and technology in the era of global transformation. Petrozavodsk: MCNP "New Science"; 2023: 158–191. (In Russ.)

3. Zhuravkov M.A. Tekhnologii iskusstvennogo intellekta i intellektual'nyye sistemy komp'yuternogo modelirovaniya i inzhenernykh raschetov = Artificial intelligence technologies and intelligent systems of computer modeling and engineering calculations [Internet]. Minsk: Belarusian State University; 2024. 177 p. Available from: https://elib.bsu.by/bitstream/123456789/309072/1/Tekhnologii%20iskusstvennogo%20intellekta%20i%20intellektual'nyye%20sistemy.pdf.

4. Kaplan A. M., Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons. 2010; 53; 1: 59–68.

5. Cambria E., Schuller B., Xia Y., Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems. 2017; 28; 2: 15–21.

6. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019.

7. Pang B., Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2008; 1–2: 1–135.

8. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003; 3: 993–1022.

9. Chen M., Mao S., Liu Y. Big data: A survey. Mobile Networks and Applications. 2014; 19; 2: 171–209.

10. Gandomi A., Haider M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. 2015; 35; 2: 137–144.

11. Smith J., Brown T. Metadata management for large-scale datasets. Journal of Information Systems. 2019; 25; 3: 120–135.

12. Voigt P., Von dem Bussche A. The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer International Publishing; 2017.

13. Ramos Gargantilla J.A., Mora J., Aguado de Cea G. Enhancing the expressiveness of linguistic structures. 2012.

14. Greer K. Concept Trees: Building Dynamic Concepts from Semi-Structured Data using Nature-Inspired Methods. 2014.

15. Galkin M., Mouromtsev D., Auer S. Identifying Web Tables – Supporting a Neglected Type of Content on the Web. 2015.

16. Giunchiglia F., Zamboni A., Bagchi M., Bocca S. Stratified Data Integration. 2021.

17. Tang C., Yuan G., Zheng T. Weakly Supervised Learning Creates a Fusion of Modeling Cultures. 2021.

18. Koo H., Eun Kim T. A Comprehensive Survey on Generative Diffusion Models for Structured Data. 2023.

19. Liu J., Zhao Z., Wu N., Wang X. Research on the structure function recognition of PLOS [Internet]. 2024. Available from: ncbi.nlm.nih.gov.

20. Mittal A., Bheemreddy A., Tao H. Semantic SQL – Combining and optimizing semantic predicates in SQL. 2024.

21. Vanschoren J., Blockeel H., Pfahringer B., Holmes G. Experiment Databases: Creating a New Platform for Meta-Learning Research. 2008.

22. Anstiss S. Understanding data quality issues in dynamic organisational environments – a literature review. 2012.

23. Yadav C., Wang S., Kumar M. Algorithm and approaches to handle large Data – A Survey. 2013.

24. Komarnitskaya O. Metody avtomatizirovannogo semanticheskogo analiza yestestvennoyazykovoy informatsii = Methods of automated semantic analysis of natural language information. 2018.

25. Leskovec J., Rajaraman A., Ullman J. D. Mining of Massive Datasets. 3rd ed. Cambridge University Press; 2014.

26. Bernstein V., Afanassenkov A. Unsupervised Data Extraction from Computer-generated Documents with Single Line Formatting. 2020.

27. Zhouli M., Ganem R., Azzuza M. Semantic analysis of big data: Challenges and opportunities. Issledovaniya v oblasti bol'shikh dannykh = Big Data Research. 2020; 18; 4: 115–130.

28. Chzhang Dzh., Li V., Lyu TS. A review of machine learning algorithms for big data classification. Zhurnal mashinnogo obucheniya = Journal of Machine Learning. 2021; 38; 7: 925–940.

29. Nguyen L., Tran T., Nguyen D. Data classification and clustering: Methods and applications. Zhurnal vychislitel'nogo intellekta = Journal of Computational Intelligence. 2019; 31; 1: 45–58.

30. Chen KH., Li T., Chzhan KH. Application of data clustering in healthcare and finance. Zhurnal nauchnykh issledovaniy dannykh = Journal of Scientific Data Research. 2020; 25; 3: 200–214.

31. Li D., Park Dzh., Kim S. The Role of Machine Learning in Marketing Analytics. Nauchnyye issledovaniya v oblasti marketinga = Scientific Research in Marketing. 2022; 39; 2: 189–204.

32. Vasilenko A., Frolov A., Makarov P. Modern Methods of Processing Unstructured Data in New Technologies. Zhurnal novykh tekhnologiy = Journal of New Technologies. 2021; 14; 1: 112–124. (In Russ.)

33. Chodpathumwan Y. Cost-effective data structural preparation. 2018.

34. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd ed. Elsevier; 2011.

35. Aggarwal C. C., Reddy C. K. Data Clustering: Algorithms and Applications. CRC Press; 2014.

36. Cambria E., Schuller B., Xia Y., Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems. 2017; 28; 2: 15–21.

37. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press; 2008.

38. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press; 2016.

39. Bishop C. M. Pattern Recognition and Machine Learning. Springer; 2006.

40. Pennington J., Socher R., Manning C. D. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532–1543.

Сведения об авторах

Хабиб Жан Макс Тапе
Факультет инновационных технологий
Национальный исследовательский Томский государственный университет,
Томск, Россия
Эл. почта: Jeanmax.habib@mail.ru

Алексей Андреевич Погуда
Научный руководитель, к.т.н, доцент,
Факультет инновационных технологий,
Национальный исследовательский Томский государственный университет,
Томск, Россия
Эл. почта: alexsmail@sibmail.com

Information about the authors

Habib Jean Max Tape
Postgraduate student, Faculty of Innovative Technologies
National Research Tomsk State University,
Tomsk, Russia
E-mail: Jeanmax.habib@mail.ru

Alexey A. Poguda
Scientific Supervisor, Candidate of Technical Sciences, Associate Professor, Faculty of Innovative Technologies
National Research Tomsk State University,
Tomsk, Russia
E-mail: alexsmail@sibmail.com