АНАЛИЗ ДАННЫХ YOUTUBE

А. Р. Файзрахманов,[1] Н. И. Мордовин,[1] А. В. Шварц[1]

[1] *Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, 194064 Россия*

Платформа youtube занимается распространением создаваемого контента в формате видеороликов пользователями (создателями контента). Каждый видеоролик имеет дополнительную информацию в виде обратного отклика от зрителей: количество лайков, количество просмотров, коментарии и т.д. Для понимания какой контент выпускать пользователю необходимо вручную анализировать данную информацию. Например, владелец обучающего канала хотел бы иметь представление о том, как видео донесло предлагаемую информацию до конечного потребителя. Помимо этого, пользователю для развития канала было бы полезно знать общую статистику всех видео на канале.

Ключевые слова: youtube аналитика, эмоциональный анализ, тональный анализ, машинное обучение, bert, распознование языков

1. ВВЕДЕНИЕ

В современном информационном обществе, где цифровые технологии становятся все более важными, изучение данных, полученных из различных онлайн-платформ, приобретает особенное значение. Одной из таких платформ, на которой ежедневно создается огромное количество контента, является YouTube. В данной работе проводится анализ данных, полученных с YouTube, с целью выявления закономерностей, трендов и понимания влияния различных факторов на восприятие и популярность видеоконтента. В данном исследовании мы обращаем внимание на ключевые параметры, такие как число просмотров, лайков, а также взаимодействие в комментариях, чтобы раскрыть основные тенденции и динамику изменения интересов аудитории.

Мы предлагаем создать приложение, которое не только предоставит пользователям доступ к нашим результатам, но и даст им собственные инструменты анализа. Это приложение, снабженное интуитивно понятным интерфейсом и функциональностью, поможет каждому пользователю провести собственный анализ данных YouTube, адаптированный к его индивидуальным интересам и предпочтениям. Таким образом, наше исследование не только предоставит новые знания в области цифрового контента, но и внесет практический вклад в разработку пользовательских инструментов для анализа данных в современной цифровой среде.

Анализ данных с YouTube представляет собой не только возможность более глубокого понимания пользовательского поведения, но и важный шаг в направлении разработки стратегий контент-маркетинга, а также повышения качества и релевантности предлагаемого видеоконтента. Надеемся, что результаты нашего исследования помогут расширить знания в области цифрового медиа и внести вклад в более эффективное использование платформы YouTube в современном информационном пространстве.

2. ОБЗОР ЛИТЕРАТУРЫ

Исследования в области анализа популярности видео на YouTube с применением эмоционального и тонального анализа комментариев активно развиваются в связи с растущей значимостью социальных медиа и онлайн-платформ. Тональный анализ направлен на определение отношения зрителей к контенту видео. Исследования в этой области пытаются определить, является ли общее мнение положительным, отрицательным или нейтральным. Эмоциональный анализ подразумевает решение задачи классификации эмоций, представленных в комментарии. Анализ эмоций и тональности комментариев может предоставить более глубокое понимание воздействия контента на зрителей и динамику

взаимодействия в сообществах **?**, поэтому мы решили сосредаточиться на данном направлении для нахождения метрик оценки популярности видео. Естественным выбором для проведения эмоционального и тонального анализа являются модели машинного обучения **?**. Существует множество моделей обученных для различных задач , таких как например эмациональный анализ постов на платформк Twitter **?**Ориентируясь на результаты, достигнутые **?** мы выбрали предобученные модели машинного обучения, основаные на архитерктуре BERT, такие как описанные в **?** для тонального анализа и в **?** для эмоционального. Также, для определения языка на котором написан комментарий была выбрана модель, описанная в Unsupervised Cross-lingual Representation Learning at Scale(2019) и дообученная на Language Recognition датасете.

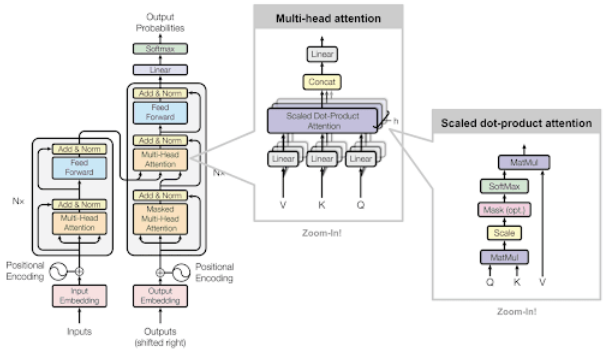

Рис. 1. Базовая архитектура моделей

3. ОПИСАНИЕ И ХАРАКТЕРИСТИКА МОДЕЛИ ДАННЫХ

3.1. Источник данных

В качестве источника данных использовалась Youtube API. Данный API предоставляет доступ к метаданным о видео, комментариях, плейлистах и каналах. YouTube API, несмотря на свою функциональность, имеет определенные ограничения, установленные Google. Одно из ключевых ограничений включает ограниченное на количество запросов в день для каждого API-ключа, что ограничивает интенсивность скачивания данных. Всего в день для одного ключа доступно 10000 единиц квоты. Стоимость запросов на поиск элементов составляет 100 единиц квоты, а для получения данных стоимость составляет 1 единицу квоты. Также еще одним ограничением является то, что за один запрос можно получить только 100 комментариев. Для обхода данных ограничений было сгенерировано несколько аккаунтов, каждый из которых имеет свой api ключ.

3.2. Характеристики датасета

Поскольку Youtube API возвращает данные в виде json, для реализации модели данных была выбрана NoSQL база данных MongoDB, которая имеет гибкую схему, масштабируемость, высокую производительность и документоориентированность (данные храняться в документах в BSON формате).

Всего было создано 7 коллекций (Рис. 2):

1. **analysis** - данные о построенной аналитике
2. **api_keys** - ключи, который используются для получения доступа к данным youtube
3. **channels** - данные о каналах, полученных с youtube
4. **comments** - данные о комментариях, полученных с youtube
5. **requests** - созданные с помощью пользовательского интерфейса запросы на скачивание данных с youtube либо на построение аналитики
6. **video_categories** - категории видео, по которым производился поиск каналов и видео
7. **videos** - данные о видео, полученных с youtube

Основными являются коллекции с данными о каналах, видео и комментариях, так как именно они подвергаются анализу. В итоге было получено более 11 гигабайт данных, более 14 миллионов комментариев под более чем 5000 видео.

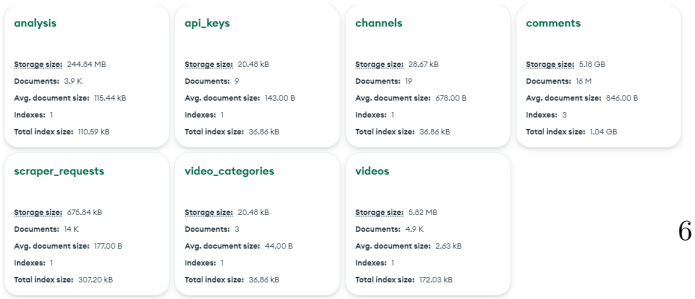

Рис. 2. Характеристика датасета

4. АРХИТЕКТУРА СИСТЕМЫ

Архитектура проекта состоит из 6 основных модулей:

1. Пользовательского интерфейса, с помощью которого пользователь может взимодествовать с системой (просматривать выгруженные данные, просматривать построенный анализ видео, создавать запросы на скачивание данных и на анализ данных).

2. Сервера приложения, в котором запущен FastAPI сервис. Данный сервис позволяет пользователю взаимодействовать с системой за счет RestAPI запросов, создаваемых пользовательским интерфейсом.

3. Сервера базы данных, в котором развернута MongoDB. Подключение к данному серверу другими модулями реализовано с помощью SSH.

4. Сервиса скачивания данных, который обращается с помощью YoutubeAPI к youtube с целью получить данные. Для этого данный сервис просматривает очередь запросов и исходя из типа запроса выкачивает определенные данные и сохраняет их в бд. Экземпляров данного сервиса может быть много, что позволит распараллелить скачивание данных и увеличить производительность.

5. Сервиса анализа данных, который анализирует данные, по поступающим в очередь запросам. После анализа данный сервис сохраняет результаты в базе данных. Данный сервис также может существовать в нескольких экземплярах.

6. Youtube - это сторонний сервис, который с помощью своего API предоставляет доступ к данным.

Логическая(Рис. 3) и физическая(Рис. 4) архитектура проекта представлены в виде диаграммы компонентов и диаграммы развертывания соответсвенно.

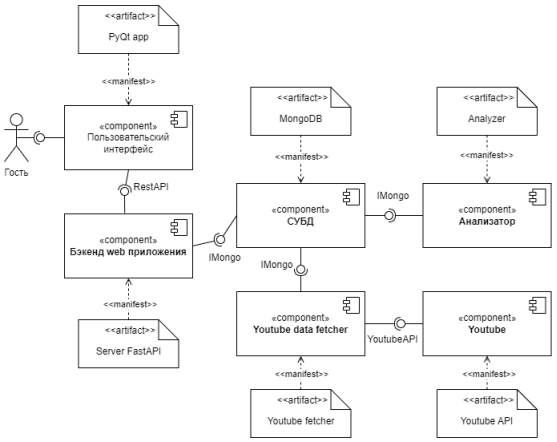

Рис. 3. Логическая архитектура системы

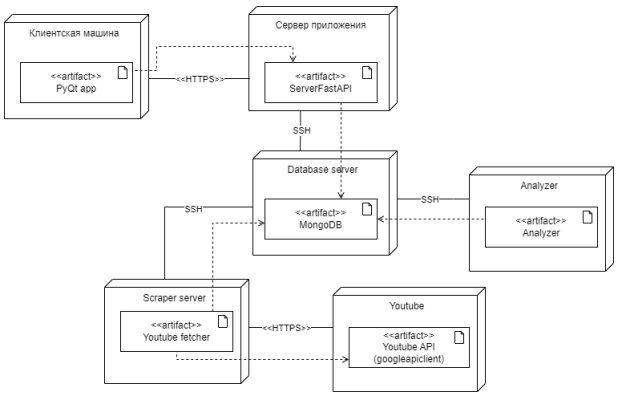

Рис. 4. Физическая архитектура системы

Архитектура была спроектированы таким образом, чтобы поддерживать горизонтальное масштабирование. Это достигается за счет модульности системы и использования распределенной базы данных MongoDB.

5. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

Важным аспектом программной реализации нашего проекта является удобство работы с кодом. Все микросервисы были разработаны с использованием модульной архитектуры, что позволяет легко масштабировать и поддерживать приложение. Все сервисы были написаны на языке программирования Python версии 3.10

База данных MongoDB используется для хранения и управления данными в нашем приложении. MongoDB является документоориентированной NoSQL базой данных, которая обеспечивает гибкую схему и высокую производительность. Python предоставляет удобную интеграцию с MongoDB, что делает его идеальным выбором для реализации хранения данных.

Для удобства развертывания и управления всеми компонентами нашего приложения мы использовали Docker контейнеры. Docker предоставляет среду для запуска приложения и его зависимостей в изолированной среде. Это позволяет с легкостью масштабировать и управлять компонентами приложения.

Backend нашего приложения был разработан с использованием фреймворка FastAPI. FastAPI является современным и эффективным фреймворком для создания веб-сервисов на Python. Он обеспечивает высокую производительность и поддерживает автоматическую генерацию документации API.

6. РЕЗУЛЬТАТЫ АНАЛИЗА

В качестве формулы для численной метрики оценки популярности видео была предложена :

$$P = F_{time} + 25N_{favorite} + F_{comments} * F_{emotion}$$

Поскольку действительное количество комментариев N_{real} под многими видео отличается от заявленного платформой Youtube $N_{Youtube}$, был сделан вывод о том, что часть комментариев удаляется владельцами видео или самой платфомой в связи со спамом или иными причинами. В связи с этим, для прогнозирования реального количества комментариев и была предложена следующая компонента, причем $C(t)$ – дискретная функция, описывающая количество комментариев от времени:

$$F_{comments} = F_{spam} * max(1, \frac{dC}{dt}) \ \text{,где}$$

$$F_{spam} = \frac{N_{Youtube}}{N_{real}}$$

Основными параметрами обратного отклика, на которые ориентируются пользователи, являются количества лайков N_{likes} и просмотров N_{views}. Также, несоменнно, на популярности должно сказываться количество времени $T_{now} - T_{published}$ в минутах, прошедшее с момента публикации:

$$F_{time} = \frac{N_{likes} + 0.25N_{views}}{T_{now} - T_{published}}$$

Как видно из графика популярное видео сделать сложно – чем больше метрика, тем меньше меньше количество видео.

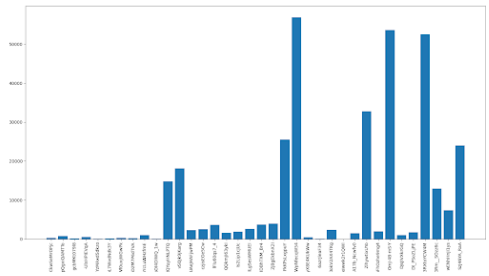

Рис. 5. Метрика популярности различных видео с разных каналов

Еще одной составляющей популярности является то, насколько комментарии под видео эмоциональны. Для вычисления этого используются значения $N_{positive}$ и $N_{Negative}$, полученные с помощью моделей машинного обучения, являющиеся количеством поизитивных и негативных комментариев соответственно.

$$F_{emotion} = \frac{N_{positive} + N_{Negative}}{N_{real}}$$

В качестве второго значения, ползволяющего конечному пользователю оценивать удачность

видео была предложена метрика эмоционального отклика:

$$E = N_{joy} + N_{surprise} - (N_{fear} + \\ + N_{disgust} + N_{sadness} + N_{anger})$$

где все значения получены с помощью соответственной модели и показывают количество комментариев с распознанной эмоцией. Большое положительное значение данной мертики означает наличие позитивного отклика, большое отрицательное – негативное. Как видно из графика видео с политическим контентом (пиковое отрицательное значение) могут вызывают много негативных эмоций, а познавательное видео (пиковое положительное значение) могут вызвать большое количество положительных комментариев.

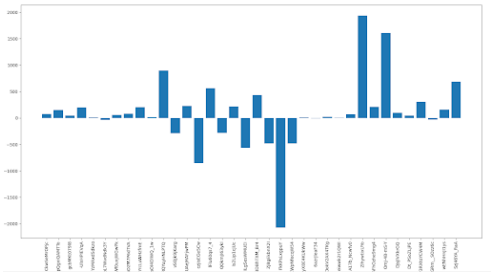

Рис. 6. Метрика эмоционального отклика различных видео с разных каналов

Из графика ниже видно, что канал развивался преимущественно с 2014 года, причем основной рост пришелся на 2014-2022 годы.

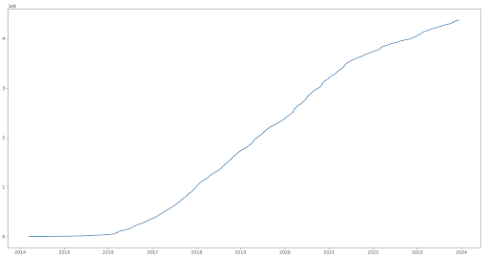

Рис. 7. Зависимость количества комментариев от времени для канала Veritasium

Как видно из нижеприведенных графиков, для канала Veritasium среднее значение длины комментария (красная линия на графике) состовляет около 50 символов, однако, существуют выбросы, такие как коментарии длиной более 17000 символов. На гистограмме видно, что количество ответов убывает соответственно длине.

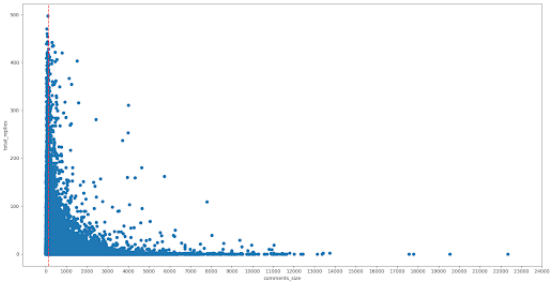

Рис. 8. Зависимость длины комментариев от количества ответов под ними для видео канала Veritasium

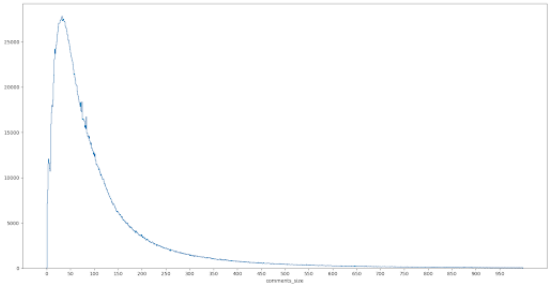

Рис. 9. Гистограмма распределения размера комментариев для канала Veritasium

В качестве одного из результатов были построены следующие круговые диаграммы для видео, имеющих большое по модулю значение метрики эмоционального отклика. Так, для политического видео 10, доля негативных комментариев составляет более 60 приоцента, а для образовательного контента 11 положительные комментарии составляют более половины.

Помомо графиков, для визуализации были составлены карты слов для негативных и позитивных коментариев, на изображении ниже приведен пример для позитивной карты слов.

7. ЗАКЛЮЧЕНИЕ

В рамках данного проекта была успешно разработана и реализована система анализа

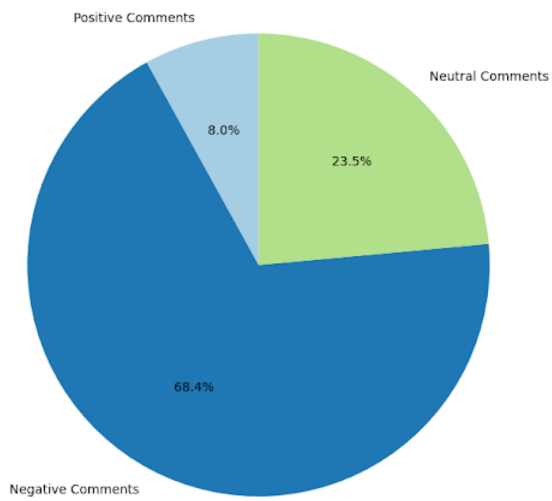

Рис. 10. Видео: Harvard President Refuses To Discuss Disciplinary Actions For Antisemitic Incidents

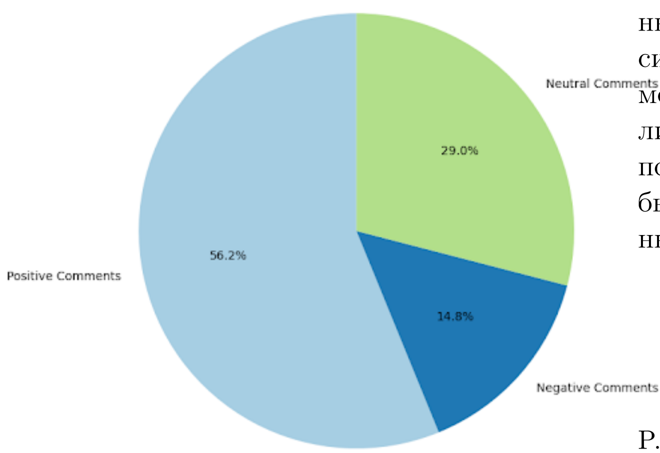

Рис. 11. Видео: The Surprising Genius of Sewing Machines

данных мета данных видео контента на основе информации, получаемой из Youtube. Разработанные микросервисы демонстрируют эффективность в сборе, обработке данных. Процесс сбора данных был осложнен ограничениями API на размер квоты и количеству данных полчаемых за один запрос. Однако, благодаря стратегии использования нескольких API ключей удалось обеспечить регулярное и непрерывное обновление данных. Для анализа полученных данных использовались модели машинного обучения, с помощью кото-

Рис. 12. Карта слов для позитивных комментариев для видео “The OTHER AI Alignment Problem: Mesa-Optimizers and Inner Alignment”

рых проводился эмоциональный и тональных анализ комментариев. Полученные значения использовались для посторения метрик популярности и эмоционального отклика от видео. Были построены различные графики, показывающие значения данных метрик для различных видео, а также описывающие такие зависимости как количество комментариев от времени публикации, размера комментария от количества ответов под ним и другие. Для дополнительной визуализации работы моделей были построены карты слов для положительных и отрицательных комментариев.

СПИСОК ЛИТЕРАТУРЫ

P. Sobhani, S. Mohammad, and S. Kiritchenko, Detecting stance in tweets and analyzing its interaction with sentiment, in *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, edited by C. Gardent, R. Bernardi, and I. Titov (Association for Computational Linguistics, Berlin, Germany, 2016) pp. 159–169.

R. Socher, Deep learning for sentiment analysis - invited talk, in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, edited by A. Balahur, E. van der Goot, P. Vossen, and A. Montoyo (Association for Computational Linguistics, San Diego, California, 2016) p. 36.

J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, and M. V. Martínez, pysentimiento: A python toolkit for opinion mining and social nlp tasks (2023), arXiv:2106.09462 [cs.CL].

D. Q. Nguyen, T. Vu, and A. T. Nguyen, Bertweet: A pre-trained language model for english tweets, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (2020) pp. 9–14.

P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, Semeval-2016 task 4: Sentiment analysis in twitter, arXiv preprint arXiv:1912.01973 (2019).

YOUTUBE DATA ANALYSIS

A. R. Faizrahmanov[1], N. I. Mordovin[2], and A. V. Schwartz[1]

[1]St. Petersburg Polytechnic University Peter the Great, St. Petersburg, 194064 Russia

The youtube platform distributes created content in the video format by users (content creators). Each video has additional information in the form of feedback from viewers: the number of likes, the number of views, comments, etc. To understand what content to release , the user must manually analyze this information. For example, the owner of a training channel would like to have an idea of how the video conveyed the proposed information to the end user. In addition, for the development of the channel, it would be useful for the user to know the general statistics of all videos on the channel

Keywords: *youtube analytics, emotional analysis, tonal analysis, machine learning, bert, language recognition*