

# ***YouTube Sentimental Analysis Project Report***

Barbara,  
Sidharth Jain,  
Tyler Supersad,  
and Yaksh Bhatt.

April 27, 2023



Department of Arts and Digital Industries  
University of Roehampton  
London, United Kingdom

Submitted in partial fulfilment of the requirements for the degree  
of Bachelors of Science in Computer Science

Supervisor Kevin Chalmer

# Abstract

Analysis of large data sets so as to reach operational and strategic conclusions is essential in effective governance and business. These data sets can often be cumbersome and difficult to analyze directly in the form of inexhaustible YouTube comment data. It is often necessary to create tools that break these comments down into smaller, often searchable data. finally this pipeline is going to collect comments from you tube and sort them out into positive, negative or neutral. Which create a labeled dataset and easy for researchers to perform experiments. The development of pipeline and the challenges faced demonstrate why this tool is so effective and what considerations were made in its creation to allow for the highest quality and utility product.

## Declaration

I hereby certify that this report constitutes my own work, that where the language of others is used, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of others.

I declare that this report describes the original work that has not been previously presented for the award of any other degree of any other institution.



**(Sidharth Jain)**



**(Tyler Supersad)**



**(Yaksh Bhatt)**



**(Barbara Kobak)**

**April 27, 2023**

## Acknowledgements

This work would have been impossible without the committed team members of the YouTube sentimental Analysis and administrative divisions. These team members showed dedication in the face of determined opposition by Docker. During the darkest and rainiest London days these teams pushed past compile errors and incompatible data type crashes. Finally through sheer force of will they produced a Dataset of YouTube Comments to change the lives of hundreds of millions of memory bits for the 3 or 4 times this application will be used for generations to come.

A special thanks to Kevin Chalmers for his patience and support during our stakeholder progress updates.

Table of Contents	PAGE NUMBER
1. Introduction	vi
1.1 Research Question or Problem that will be Addressed	vi
1.2 Aims	vi
1.3 Objectives	vi
1.4 Legal, Social, Ethical and Professional Considerations	vii
1.5 Background	vii
2. Literature or Technology	viii
2.1 Review	viii and ix
3. Design or Methodology	x
4. Implementation	xi
4.1 Results	Xii and xiii
5. Conclusion	xiv
5.1 Reflection	xiv
5.2 Future Works	xiv
6. References	xv
7. Appendices	xvi

# 1 . Introduction

YouTube comment pipeline is the process of examining and interpreting the comments posted by users on YouTube videos to gain insights and understanding about the opinions, sentiments, and trends related to the content being discussed. As one of the biggest video sharing platforms on the internet, YouTube generates an huge amount of user-generated comments, which can provide valuable information to content creators, marketers, researchers, and other stakeholders.

YouTube comment analysis involves various techniques, such as sentiment analysis, and data mining, to extract meaningful information from the comments. By analyzing YouTube comments, it is possible to uncover patterns and trends in user engagement, sentiment towards the content, and audience feedback. This analysis can help content creators understand their audience better, identify potential areas for improvement, and develop strategies to enhance viewer engagement and satisfaction.

In this report we have discussed about YouTube comments sentimental analysis and how it works in detailed.

## Aims

The aim of a YouTube comment analysis pipeline is to correctly process and analyze comments posted on YouTube videos, to extract meaningful insights and information from the data.

## Objectives

- Data Collection : YouTube comments has been extracted from the videos of selected channel using YouTube API.
- Data cleaning: To Preprocess the data and remove all the irrelevant information for example any URL , emojis , special character and so on. Just extracting text for better understanding for machine and do better sentimental analysis.
- Sentiment Analysis: Applying different kind of techniques such as NLTK(library) to determine its polarity (positive , negative or neutral)
- Create a Deployable Docker image so as to allow the Software to be used on any device without any complex setup.
- Save all the data after preprocessed in a dataset so that it can be used for future research purposes.

## Legal, Social, Ethical and Professional Considerations

### **Legal Considerations:**

1. Data Privacy : Ensuring that the extracted data involves privacy laws and regulations such as General data protection (GDPR), when storing and analyzing the user generated comments from YOUTUBE.
2. Terms and conditions : Adhering to YouTube's term and condition and maintaining those restrictions which includes collection and use of comments data, and ensuring compliance with YouTube's policies related to data usage, privacy, and content moderation.

### **Social Consideration**

1. Respect Everyone: Respecting everyone view points in the YouTube comment section. avoiding any manipulation or distortion of comments data to fit biased comments.
2. Diversity: Considering the diverse ideas and opinions expressed in YouTube comments, and avoiding discrimination on race, gender, religion, nationality, sexual orientation, in the analysis and reporting.

### **Ethical Consideration**

1. Human Rights and Social Impact: Considering the potential impact of the analysis on human rights, social values, and public interest, and avoiding any actions that may contribute to the spread of misinformation, hate speech, or harmful content on the YouTube platform.

### **Professional Considerations**

1. Professional Consideration: Ensuring that the analysis is conducted by people with knowledge of data analysis, and related fields, who follow best practices and standards in their work.
2. Data Security: Taking appropriate measures to protect the security and confidentiality of the collected comments data, such as using secure storage, encryption, and access controls, to prevent unauthorized access or data breaches.

## 2. Technology and Literature Review

In this section we are going to discuss different Technologies we used in this Product and their importance.

### Technology Review

- **Docker:** It is an Open source Container platform which enable developer to build, run, update, and stop containers using simple commands and work-saving automation through a single API[1]. It helps to run the application to any environment just with the configuration.

Flexibility: Docker allows you to package not only the Airflow platform but also any custom or third-party plugins, configurations, and dependencies into a single container, making it easy to manage the entire Airflow stack as a single unit.

Portability: Docker containers can be easily moved between different environments, making it easy to deploy and manage Airflow workflows across different development, testing, and production environments.

- **Python :** Python is an interpreted, object-oriented, high-level programming language with dynamic semantics.[2] Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together.[2] Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse.[2]
- **NLTK:** NLTK (Natural Language Toolkit) is a widely used open-source Python library for natural language processing (NLP) tasks.[3] It provides a comprehensive set of tools and resources for text processing, analysis, and manipulation in various NLP applications, such as sentiment analysis, text classification, tokenization, stemming, lemmatization, part-of-speech tagging, and more.[3] NLTK offers a rich collection of corpora, lexical resources, and pre-trained models for NLP tasks, making it a popular choice for researchers, developers, and data scientists working on text analysis projects. NLTK's functionalities are designed to be easy to use and can be combined with other Python libraries for a complete NLP pipeline.[3]
- **Kanban Board :** It is a board where you can define your tasks for different sprints and change the priority as per user requirements . Moreover, you can change the status of the task as you progress in your tasks. We have used this to make sure all the tasks are listed properly and can make sure the progress of the group is made in different aspects of sprints.
- **Git-Hub :** Git-Hub is an Online and open source Software Development Platform used for Storing , Tracking and Collaborating on Software Projects.[4]. The Purpose for using Git-hub is to store all the files together and accessible to each of the team. Moreover, anyone



from the team can access it and make changes which will reflect to the other team members as well. It monitors the work Contributions by each team member and provides some functionalities for Scrum master as well.

- **Docker Airflow** : Docker Airflow refers to using Docker, a popular containerization platform, to deploy and manage Apache Airflow, an open-source platform for orchestrating complex data workflows. Docker allows you to package applications, along with their dependencies, into lightweight, portable containers that can be run consistently across different environments, such as development, testing, and production

### **Literature Review**

A literature review is a critical analysis and summary of existing research and publications related to a specific topic. In the context of a YouTube comment pipeline, a literature review would involve reviewing relevant scholarly articles, research papers, conference proceedings, and other sources of information that discuss various aspects of YouTube comments and comment analysis.

1. **YouTube Comments and User Behavior:** This could include studies that shows the characteristics of YouTube comments, such as comment length, and sentiment. It could also explore user behavior in commenting, such as the motivations for leaving comments, the types of comments (e.g : spam), and how user involvement with comments may impact the video maker.
2. **YouTube Comment Sentiment Analysis:** This research including studies that explore the challenges and approaches for sentiment analysis in the context of YouTube comments. It could also examine the applications of sentiment analysis in understanding the sentiment trends, opinion mining, and user feedback analysis in YouTube comments.

## 3 . Methodology

The DAG YouTube comment sentiment Pipeline is a system that uses the content of each comment to calculate the average sentiment score for each comment. Several interconnected modules in the pipeline carry out a variety of tasks, including data collection, data pre-processing, and sentiment analysis. The pipeline's design and method are outlined below.

**Data Ingestion:** The first stage of the pipeline involves collecting statistics from YouTube. This can be accomplished through the usage of the YouTube API, which presents access to comments statistics circulate. The information may be filtered based totally on keywords, hashtags, or other standards to make certain that the best relevant comments are collected.

**Data Pre-processing:** Once the statistics have been accumulated, it wishes to be pre-processed to make certain that it's miles in a layout that may be analysed. This might also involve cleaning the statistics, doing away with any irrelevant facts, and transforming the information into a format that is like-minded with the sentiment analysis device. Utilizing Python libraries like NLTK, clean and pre-process the raw Youtube data by removing stop words, URLs, special characters, and other noise.

**Sentiment Analysis:** The sentiment analysis stage entails using a gadget studying algorithm to categorise every comment as both positive, poor, or neutral. This may be achieved through the usage of a pre-skilled model or via training a brand new version of the usage of the collected statistics.

**Data Aggregation:** Once the sentiment scores had been calculated for every comment, they may be aggregated to calculate the common sentiment rating for the whole set of comments.

## 4 . Implementation and results

In this section, we are going to discuss how accurately we have executed our requirements and challenges we have faced to complete this End Product.

### **Challenges during Implementations:**

The execution of the DAG information pipeline for YouTube comment examination includes the utilization of different instruments and advancements, for example,

**Airflow by Apache:** The data pipeline design, scheduling, and monitoring platform Apache Airflow is open-source.

**Pandas:** Pandas is an information control library utilized for information pre-handling and conglomeration.

**NTKL-:** Tokenization, part-of-speech tagging, stemming, and sentiment analysis on text data can all be done through intuitive interfaces.

**Docker:** For simplicity of deployment and scalability, the entire pipeline is packaged into a single container with the help of the containerization platform known as Docker.

**Retrieve comments:** The comment dataset is retrieved from a source, such as the API of a social media platform, in this task. The dataset can be put away in a data set or document framework for later handling. There are no dependencies on this task.

**comments that are free of clutter:** This task standardizes the text and removes irrelevant data from the raw YouTube data using a variety of cleaning methods. The outcome of the Fetch comment task is required for this task.

**Analyses:** The cleaned comments are analyzed for useful information like sentiment, keywords, or mentions in this task. For this situation, we need to register the YouTube comment length. The results of the Clean comments task are required for this task.

**Find the Mean:** Based on the Analyses task's output, this task calculates the YouTube comment length. There are no dependencies on this task.

```

# Create a task to preprocess the tweet data.
def clean_task(**context):
    # Create an instance of the Preprocess() object.
    preprocess = Preprocess()

    # Use ti.xcom_pull() to pull the returned value of extract_task task from XCom.
    tweet_data = context['task_instance'].xcom_pull(task_ids='extract_task')

    # Utilize the Preprocess() method to clean data.
    tweet_data = preprocess.clean_tweet_data(tweet_data)
    return(tweet_data)

# Create a task to classify the sentiment of each tweet within the tweet data.
def classify_task(**context):
    # Create an instance of the Preprocess() object.
    preprocess = Preprocess()

    # Use ti.xcom_pull() to pull the returned value of extract_task task from XCom.
    tweet_data = context['task_instance'].xcom_pull(task_ids='clean_task')

    # Classify the sentiments of all tweets and add that info to the tweet_data.
    tweet_data = preprocess.classify_tweet_data(tweet_data)
    return(tweet_data)

# Create a task to allocate tweet data to PostgreSQL database table.
def store_task(**context):
    # Create an instance of the Database() object.
    db = Database()

    # Use ti.xcom_pull() to pull the returned value of clean_task task from XCom.
    tweet_data = context['task_instance'].xcom_pull(task_ids='classify_task')

    # Create the 'Tweets' Table within the PostgreSQL database.
    db.create_table()

    # Insert the tweet_data into the 'Tweets' Table.
    db.store_data(tweet_data)

```

Fig 1

**Results:**

Results of DAG YouTube comment sentiment Pipeline Analysis Using the DAG data pipeline, it is possible to gain a better understanding of the overall sentiment trends and patterns over time. A portion of the key outcomes that could be gotten from the investigation include:

For each comment, the average sentiment score was: This helps determine whether each comment has a positive, negative, or neutral sentiment.

Feeling patterns after some time: This aids in distinguishing the general feeling patterns over the long run and whether some particular occasions or events are affecting the opinion.

Popular topics or hashtags: This aids in the identification of popular hashtags or topics that are influencing comment sentiment.

Sentiment analysis of users: This makes it easier to determine how certain users feel and how they affect the sentiment of the comments as a whole.

In general, the DAG YouTube comment sentiment Pipeline analysis is a powerful tool for studying sentiment trends and patterns over time in large quantities of comments.

```
# Construct a method to return the sentiment result of text.
def obtain_tweet_sentiment(self, text):
    # Initialize the sentiment analyzer.
    analyzer = SentimentIntensityAnalyzer()

    # Initialize the sentiment variable.
    sentiment = ''

    # Obtain the sentiment score.
    score = analyzer.polarity_scores(text)

    # Obtain the compound score.
    compound = score['compound']

    # Classify the tweet sentiment based on the compound score.
    # If the compound score is greater than 0.05, the tweet is classified as positive.
    if compound >= 0.05:
        sentiment = 'positive'
    # If the compound score is less than -0.05, the tweet is classified as negative.
    elif compound <= -0.05:
        sentiment = 'negative'
    # If the compound score is between -0.05 and 0.05, the tweet is classified as neutral.
    else:
        sentiment = 'neutral'

    # Return the sentiment.
    return(sentiment)
```

Fig.2

### Strengths :

The application is robust and meets all the requirements provided by the user.

It is user friendly and very interactive.

Respond to user help requests, so we can provide assistance to users that are having difficulty navigating through the application / Dataset features/functionalities.

### Weakness :

Lack of time and resources has lead to minor extended feature cuts such as user web page customization options

## 5 . Conclusion

In conclusion, the DAG YouTube Comment Sentiment Pipeline provides a powerful means for analysing sentiment trends and patterns in large volumes of comments. The pipeline's design and methodology make it easy to collect and pre-process data, perform sentiment analysis, and aggregate sentiment scores. The integration of various tools and technologies such as Apache Airflow, Pandas, Text Blob, NTLK, Matplotlib, and Docker make the pipeline scalable, efficient, and easy to deploy. The key results obtained from the analysis include sentiment scores for each YouTube comment, trends in sentiment over time, popular topics or hashtags influencing comment sentiment, and sentiment analysis of users. In summary, the DAG data pipeline for YouTube comments analysis is an effective and reliable tool for analysing sentiment trends and patterns in comments. It can be applied in various contexts, such as monitoring brand reputation, assessing public opinion, and conducting market research.

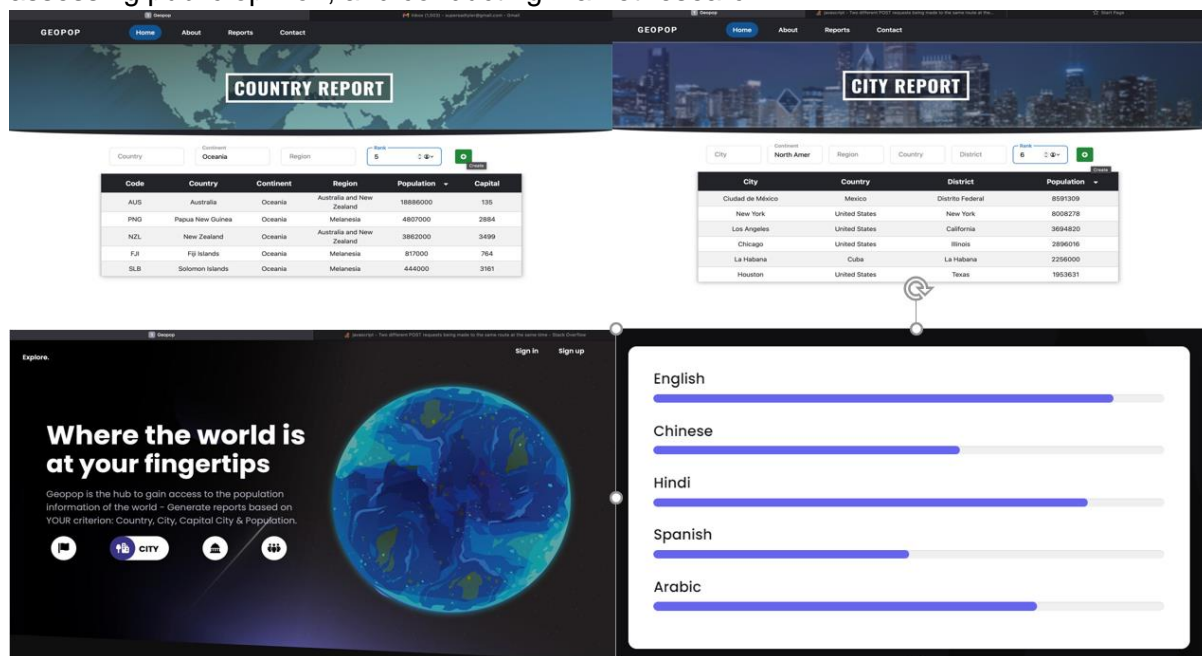


Fig 3

### Reflection and Future Updates:

We believe in the concept of “life learning”. While working on this project we faced a lot of difficulties and tried to find the best solution to overcome problems and give our best. While working on this product we learned so many different technologies and how the work is being done in a professional environment as a team and how to communicate with each other in a professional way.

As far as concern for the future updates, we are planning to have a chat bot which is going to interact with the users and help them resolve their problems. Time to time changes in the product interface to improve it using user feedback and connect with google cloud to make it more robust.

### Whats Next:

We try to establish new divisions and connect you tube sentimental with Cloud. Using all the lessons learned on in class or using online reasources we will improve our product for interstellar use.

## 6 . References

- [1] <https://www.ibm.com/uk-en/cloud/learn/docker>
- [2]<https://www.python.org/doc/essays/blurb/>
- [3]<https://www.nltk.org/>
- [4]<https://blog.hubspot.com/website/what-is-github-used-for#:~:text=GitHub%20is%20an%20online%20software,developers%20on%20open%2Dsource%20projects.>
- [5] <https://www.jetbrains.com/help/idea/discover-intellij-idea.html#multi-platform-IDE>



## 7 . Appendices

### The Caffeinated Quantum Squadron Code of Conduct

*To establish and maintain the trust, respect, and collaboration within the Caffeinated Quantum Squadron, I do hereby pledge to abide by the policies of the team throughout the completion of the project. I give my word of honor that I will observe the following policies:*

**I. BEHAVIOR - All members within the team are to be treated with respect, kindness, and equality.**

As a team, we all depend on each other to produce the best work we can achieve. However, an environment where people feel uncomfortable or threatened is not a productive or creative one. Therefore, each member is expected to uphold the responsibility to encourage and support one another during the timeframe of the intended task.

**II. ORGANIZATION - Attend scheduled meetings regularly and on time.**

Meetings will occur in the 2<sup>nd</sup> Floor Conference Room inside the Sir David Bell Building (University of Roehampton Campus) on Thursday's from 2:00 PM to 4:00 PM each week. Members are required to meet these arrangements unless they are hindered by an unexpected/unfortunate predicament.

**III. INTEGRITY - Maintain honesty and clear communication within the environment.**

Communication is an integral component to project management; hence all members are encouraged to inform one another about any discrepancies in order to resolve workplace conflict in an efficient manner. Additionally, in no circumstance shall a member within the team collude with any members from outside parties about information concerning the project. If there is beyond reasonable doubt of suspected plagiarism, extreme measures will be enacted for breaching academic honesty, which may result in a failing grade.

*In an event that a member violates any rules detailed within the pillars of Behavior, Organization, and Integrity, disciplinary action will be enforced. To ensure the presence of fairness and governance when dealing with disobedience, each member will be allocated a total of three warnings.*

- **1<sup>st</sup> Warning** - Minor problems will be dealt with by informal advice and counselling.
- **2<sup>nd</sup> Warning** - Warn that action under the 3<sup>rd</sup> Warning will be considered if there is no satisfactory improvement. Additionally dealt with during an informal discussion.
- **3<sup>rd</sup> Warning** - Penalties from the Module Delivery Team will be implemented.

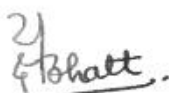
\* Google Docs will serve as the platform to record a counselling session if a warning is issued.

\* The type of warning may vary depending on the severity of the situation deemed applicable by a group majority (excluding the member on trial).

Examples of scenarios that justify the issuance of a warning:

- Failure to attend meetings (especially without a valid reason or excuse)
- Inability to complete assigned tasks within a given deadline
- Exhibition of poor, ineffective communication amongst teammates
- Plagiarism

*Signing this document exemplifies a thorough review of the Caffeinated Quantum Squadron Code of Conduct and an adherence to its policies of Behavior, Organization, and Integrity. Please be aware that any violations of the Caffeinated Quantum Squadron Code of Conduct may lead to disciplinary action, up to and including the involvement of the Module Delivery Team, as determined to be appropriate by a group majority.*



(Yaksh Bhatt)



(Sidharth Jain)



(Barbara Kobak)



(Tyler Supersad)