**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.

[1] <https://www.ibm.com/uk-en/cloud/learn/docker>

[2]<https://www.python.org/doc/essays/blurb/>

[3]https://www.nltk.org/