**A red and black logo

Description automatically generated**

**Comment Sentiment Analysis**

**Lecturer:**

Dr. Chen Hajag

**Group Members:**

Einav Diar 319010807

Eden Mengistu 209218890

**GitHub Link:**

<https://github.com/einavdiar/YouTube-Comment-Sentiment-Analysis.git>

**Abstract**

This project investigates the Natural Language Processing (NLP) for sentiment analysis on YouTube, utilizing comments to gauge viewer perception of video content.

We used the YouTube data API to collect comments, video titles, and IDs from 50 videos and hundred comments for each video, for getting a substantial dataset for analysis.

we extracted comments related to the search term "Politics".

This report offering insights into the sentiments expressed in YouTube comments and presenting a novel approach to content categorization based on viewer feedback.

**Introduction**

The proliferation of user-generated content on platforms like YouTube has accentuated the need for effective sentiment analysis tools. Our project addresses this need by focusing on YouTube comments, a rich source of viewer feedback. By analyzing sentiment, we aim to provide content creators and platform moderators with insights into public perception of their content. The objective was to develop an NLP-based model capable of classifying videos into sentiment categories based on comment analysis, thereby facilitating a better understanding of audience response.

**Dataset and Features**

**Data Collection:**

We employed the YouTube API to collect comments, video titles, and IDs from 50 videos, resulting in a dataset encompassing approximately 100 comments per video. This extensive collection process ensured a diverse and representative sample of viewer interactions.

**Preprocessing Steps:**

Our preprocessing pipeline included cleaning text data by removing special characters and irrelevant information, such as URLs and user mentions, punctuation, symbols and stop words (common words like "the", "a") to focus on sentiment-carrying words.

This preprocessing was critical for ensuring data quality and relevance was critical for ensuring data quality and relevance.

**Feature Selection:**

Our sentiment analysis project used three features extracted from the YouTube dataset:

Comments: The main focus of our analysis, comments offer direct insights into viewers’ sentiments towards video content. Through NLP, we categorized these sentiments as positive, neutral, or negative.

Video ID: This unique identifier for each video was crucial for organizing comments and associating them with the correct video content, ensuring accurate sentiment analysis.

Video Title: Titles provided contextual information, aiding in the interpretation of comments’ sentiment in relation to the video’s theme or content.

**Methodology**

We run an NLP model, TextBlob, to indicate initial sentiment.

The TextBlob model was chosen for its simplicity and efficiency in generating initial sentiment scores. TextBlob's functionality allowed us to quickly assess the emotional tone of comments, and served as a foundational step in our analysis.

The distribution of original comment sentiment scores helps by providing insights into the overall sentiment bias in the dataset.

we use the elbow method to find the best balance between the number of clusters and the model's accuracy, while silhouette analysis ensures these clusters are meaningful and well-separated.

we employed Principal Component Analysis (PCA) for dimensionality reduction, facilitating a more focused analysis of sentiment patterns.

Next, we developed a custom machine learning model that classifies videos into positive, neutral, and negative sentiment rates.

Finally, we used a GUI that providing a user-friendly interface for real-time analysis, visualizing results for easier interpretation, streamlining the analysis workflow, and making the tool accessible to a wider, non-technical audience

**Experiments/Results/Discussion**

**Experiments**

Our experimental setup involved splitting the data into training and testing sets, ensuring a robust evaluation of the model's performance. We employed cross-validation to assess the model's consistency across different subsets of the data.

**Parameter Choices**

We utilized the YouTube API to fetch comments, using parameters that target specific videos to ensure a focused dataset relevant to our study.

The TextBlob library was our primary tool for sentiment analysis, given its efficiency in quickly evaluating the emotional tone of text data without requiring additional parameter tuning.

After the sentiment analysis stage we used PCA.

This step enhanced our ability to discern meaningful insights from the YouTube comments, contributing to a deeper understanding of public sentiment towards political content.

By incorporating PCA into our methodology, we were able to streamline our analysis pipeline and extract more meaningful information from the YouTube comment dataset, ultimately enhancing the quality and depth of our findings.

**Evaluation Metrics**

Because our project is unsupervised project, and we're dealing with finding patterns without pre-labeled data, traditional supervised learning metrics such as accuracy, precision, recall aren't applicable. Instead, we rely on metrics designed to evaluate the model's performance based on the structure discovered in the data.

For determining the optimal number of clusters for our K-means clustering algorithm, we employed the Elbow Method. This technique involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the "elbow" point, where the rate of decrease in WCSS sharply decreases.

By visually inspecting the plot generated by the Elbow Method, we were able to identify the most suitable number of clusters for our dataset, optimizing the clustering solution for subsequent analysis.

In addition, we used Silhouette Score to measures how similar an object is to its own cluster compared to other clusters. The value ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

With visual Inspection through seaborn and matplotlib, we conducted qualitative assessments of our clustering, observing the distribution and separation of sentiment scores post-normalization.

**Quantitative and Qualitative Results**

Our analysis revealed insights into common sentiment trends across different types of content, highlighting the model's utility in capturing viewer sentiment.

Qualitative Observations: Visual analysis revealed clear separations between positive, neutral, and negative sentiment clusters, affirming the effectiveness of our preprocessing and the potential of our model for categorizing sentiments.

By combining quantitative assessments of silhouette scores with qualitative observations through visual analysis, we gained comprehensive insights into the performance and effectiveness of our sentiment analysis model. These findings not only validate the utility of our methodology but also provide a solid foundation for understanding and interpreting sentiment patterns in political discourse on YouTube.

**Algorithm Performance**

TextBlob provided a robust foundation for preliminary sentiment analysis, with its simplicity and effectiveness making it a valuable tool for assessing comment sentiment.

Our custom machine learning model for categorizing sentiment showed promising results in preliminary tests.

While further optimization and fine-tuning are necessary to maximize performance, the preliminary results suggest that our custom model holds promise for effectively categorizing sentiment in YouTube comments related to political content.

**Conclusion and Future Work**

Conclusion and Future Work

Our project contributes to the growing field of sentiment analysis by demonstrating a successful application of NLP techniques to YouTube comments.

Future directions include expanding the dataset, exploring more sophisticated models, and integrating additional features such as emoji analysis to enhance sentiment detection.

Based on the robust findings of the Silhouette analysis, it becomes apparent that Natural Language Processing (NLP) stands out as a reliable and effective method for classifying videos based on the sentiments expressed within their comments.

**Contributions**

The team collaborated effectively, with roles encompassing data collection, preprocessing, model development, and analysis. Each member's contribution was vital to the project's success, demonstrating the value of teamwork in tackling complex data science challenges.

At the beginning of the project, we thought of splitting the work and working at the same time.

Following experience from the assignment we received last year, Einav is the one who pulled the data with the help of an API after thinking together which columns are of interest to him.

In the next step, Eden started preparing the data, the step included checking that there are no missing values, cleaning stop words and unusual values.

After that, we decided to work together and combine our abilities until the end of writing the project.