**"Mining YouTube Comments: An Information Retrieval Framework for Sentiment and Engagement Insights"**

\*1Jammisetty Sai Nanda Gopal, 2Vemisetti Srinish, 3Sanjay Singh

123Manav Rachna University, Faridabad, Haryana, India

2srinish183@gmail.com

3sanjaysingh@mru.edu.in

\* sainandagopal09@gmail.com

**Abstract:**

This study combines web scraping, natural language processing (NLP), and visualisation approaches to propose an information retrieval framework for analysing sentiment and user engagement in YouTube comments. Getting useful insights from sites like YouTube is essential for comprehending audience feedback and engagement trends, especially in light of the increasing amount of user-generated material on social media. We gathered a large dataset of YouTube comments by employing web scraping techniques, which were subsequently pre-processed and subjected to sentiment analysis using transformer-based models. Word clouds and sentiment distribution charts were used to illustrate thematic patterns, bringing to light important trends and recurring feelings in a range of video genres and subjects.

Our results show significant differences in user sentiment that are related to engagement measures, video content, and, when accessible, demographic or geographic characteristics. Through the use of sophisticated natural language processing algorithms, this study shows how YouTube comments can offer insightful information to researchers, content producers, and companies looking to enhance audience engagement. In addition to aiding in the recovery of information from social media, this project highlights the value of sentiment and thematic analysis in user engagement studies, creating opportunities for future research in audience sentiment modelling and social media information retrieval.

**Introduction:**

As social media has grown in popularity, user-generated material has emerged as a crucial information resource for comprehending trends, user involvement, and public opinion. Massive volumes of data are stored on websites such as YouTube in the form of comments, reactions, and engagement metrics that show the mood and habits of users. Researchers looking to understand audience mood and new trends as well as content producers looking to maximise interaction must analyse this data. Specifically, YouTube comments provide a unique perspective into user reactions by offering unfiltered input on content that can aid improve audience interaction and retention techniques.

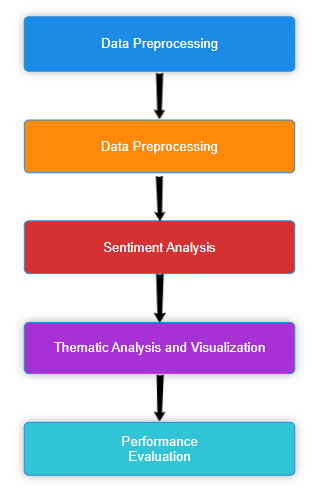
In this study, we collect, process, and analyse YouTube comments using information retrieval (IR) techniques, with an emphasis on engagement patterns and sentiment. We gathered a dataset of comments from particular YouTube videos using web scraping, and then we cleaned, tokenized, and analysed the content using natural language processing (NLP) techniques. To further grasp the emotional tone of user feedback, we performed sentiment analysis on the comments using transformer-based models, categorising them as neutral, negative, or positive. Furthermore, better understanding of the recurring themes and subjects in user comments was made possible by thematic analysis using word clouds and keyword frequency.

This study's main goal is to show how sentiment and theme analysis of YouTube comments can be used as an effective information retrieval technique that provides insightful data on social media user behaviour. We add to the field of social media analysis in IR by visualising thematic content and engagement patterns, demonstrating how automated methods can record and analyse sentiment from vast audiences. The results of this study are intended to provide a basis for further investigation of audience-driven insights in digital content ecosystems by informing research in the areas of sentiment analysis, user interaction modelling, and social media information retrieval.

There are five sections in the paper. By outlining the background, driving forces, and goals of the research, Section 1 presents the study. The Literature Survey, Section 2, reviews previous research and identifies knowledge gaps that this work fills. The methodical procedure used for data collecting, preprocessing, analysis, and visualisation is described in Section 3, Methodology. The main conclusions and a thorough examination of the results, together with their consequences, are presented in Section 4, which is devoted to Results and Discussion. The study's main contributions are summed up in Section 5, Conclusion and Future Work, which also offers some possible directions for further research.

**Proposed Methodology**

In order to learn more about user sentiment and thematic trends, the technique for this study is set up to methodically extract, preprocess, and analyse YouTube comments. As seen in **Figure 1**, this method is broken down into five main parts, each of which advances the main goal of comprehending audience interaction and feedback.



**Fig 1: Flow Chart**

1. **Data Collection**

Gathering a varied dataset of YouTube comments from videos in different categories is the first stage. Based on a selection of videos, comments were automatically retrieved from YouTube pages using web scraping techniques and Python packages like BeautifulSoup and Selenium. The video ID, user engagement metrics (likes, replies), and timestamps were among the metadata included in each structured format (CSV) remark **[10]**.

1. **Data Preprocessing**

The raw comments are converted into an analysis-ready format by data preprocessing. To clean and standardise the data, we used a number of preprocessing techniques, which included:

* **Lowercasing:** For uniformity, all text is being converted to lowercase.
* **Removing Non-Textual Elements:** removing hyperlinks, emoticons, and special characters.
* **Tokenization and Stop Word Removal:** separating remarks into separate words and eliminating common stop words (such as "the," "and").
* **Stemming:** To guarantee consistent analysis, words are reduced to their base form (for example, "watching" to "watch").

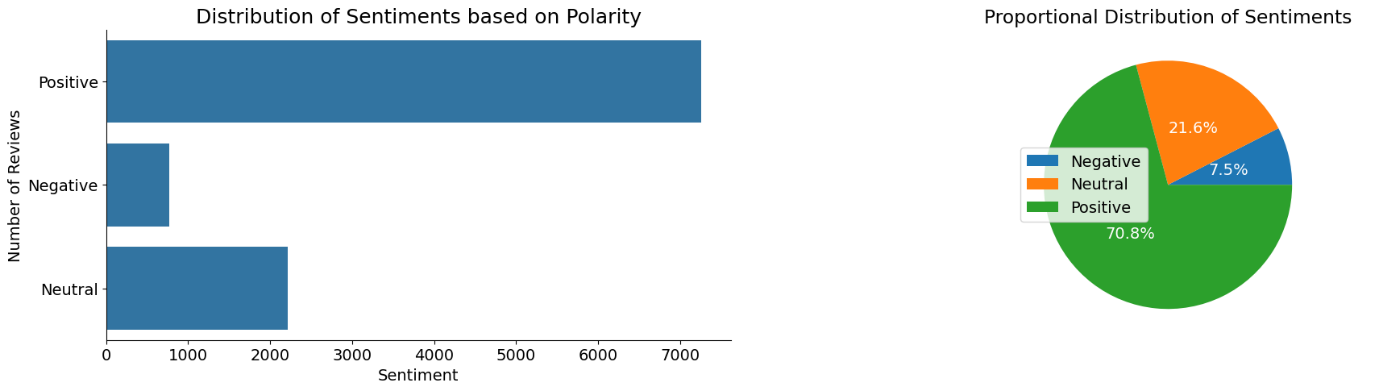
The comments were prepared for sentiment and theme analysis via this preprocessing, which made sure they were in a clear and uniform manner **[1, 4]**.

1. **Sentiment Analysis**

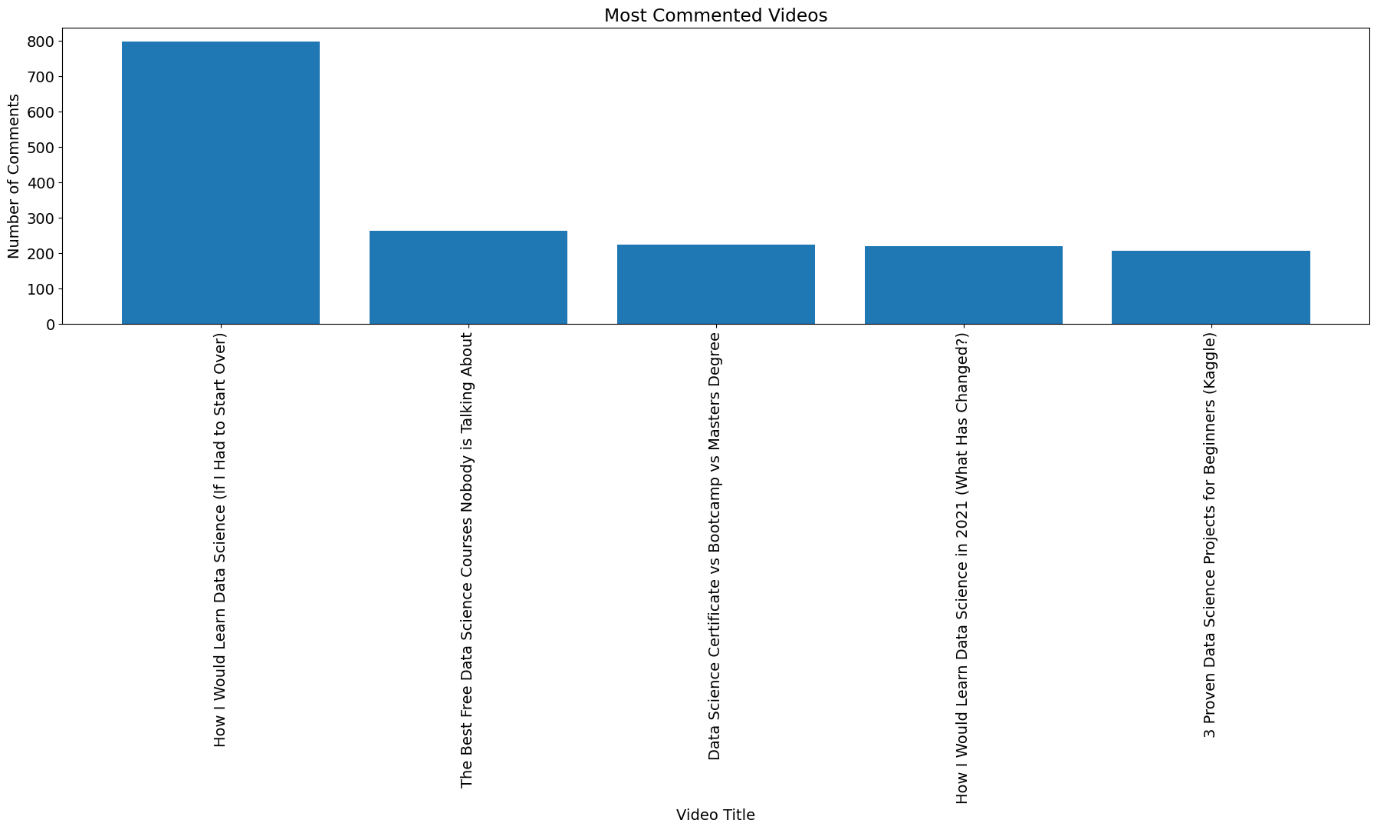
For sentiment categorisation, we used transformer-based models, namely BERT (Bidirectional Encoder Representations from Transformers). Using our dataset, the model was adjusted to fit the informal and colloquial language style found on social media. Sentiment ratings indicated the emotional tone of each comment, which was categorised as either good, negative, or neutral. When these scores were combined, overall sentiment trends across various video topics were revealed **[2, 9]**.

1. **Thematic Analysis and Visualization**

We employed Latent Dirichlet Allocation (LDA) for topic modelling in order to determine the main themes that viewers discussed in order to comprehend the recurrent themes in the comments. The main issues of interest were visualised with the aid of word clouds and frequency distributions, which highlighted frequently used phrases. Matplotlib and Seaborn were used to construct these visualisations, which effectively depict word frequency, sentiment distribution, and theme patterns **[5,6]**. Plotting the Distribution of sentiments graph using polarity shown in **Figure 2** and also plotting a graph for Most Commented videos as shown in the **Figure 3**.



**Fig 2: Distribution of Sentiments based on Polarity**



**Fig 3: Most Commented Videos**

1. **Results and Discussion**

Several metrics, including accuracy, precision, recall, and F1-score, were used to assess the sentiment categorisation model's performance. Three sentiment categories—positive, neutral, and negative—were used to gauge the model's performance. These metrics offer a thorough grasp of the model's capacity to accurately detect feelings and manage the intricacy of informal, contextually rich social media text **[7]**.

**Analysis of Confusion Matrix:**

The classification results were further evaluated in depth using a confusion matrix. The matrix helps identify particular regions where the model performs well or poorly by offering insights into the true positives, false positives, false negatives, and true negatives for each sentiment class. For instance:

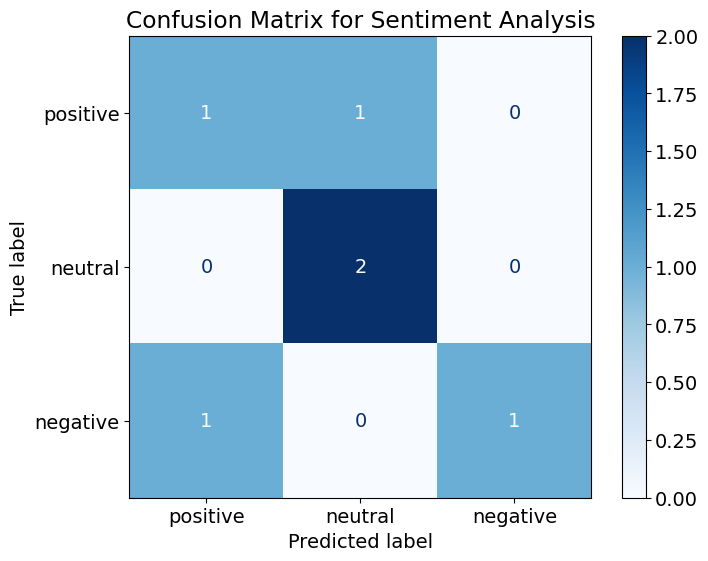
**True Positives (TP):** Comments that are appropriately categorised for a certain sentiment category.

**False Positives (FP):** Remarks that are mistakenly categorised as a particular sentiment.

**False Negatives (FN):** Misclassified comments that fall under an emotion category.

**True Negatives (TN):** Comments that are accurately categorised as not falling under a particular sentiment are known as True Negatives (TN).

In this Confusion Matrix has plotted for Sentiment Analysis which was shown inthe **Figure 4.**



**Fig 4: Confusion matrix for Sentiment Analysis**

**Metrics Derived from the Confusion Matrix:**

A transformer-based model (BERT: Bidirectional Encoder Representations from Transformers) was evaluated and refined using the gathered YouTube comment dataset to produce the metrics shown in the table. Three categories—positive, neutral, and negative—were used to train the algorithm to classify the sentiment of comments. Because of its capacity to manage the informal and contextual nature of social media writing, BERT was selected as the best tool for examining the complex language present in YouTube comments.

**Accuracy:** Determines how accurate the model is overall by dividing the number of correctly classified comments by the total number of comments.

**Precision:** Indicates the model's dependability in particular categories by calculating the proportion of anticipated positive comments that are actually positive.

**Recall (Sensitivity):** Indicates the percentage of true positive examples that were accurately recognised, demonstrating the model's sentiment detection capabilities.

**F1-Score:** Offers a fair metric that takes into account both recall and precision, which is crucial when working with unbalanced datasets. These all are mentioned in **Table 1** as named as Classification Report.

Together with the confusion matrix, these metrics provided a strong framework for evaluation that exposed the sentiment categorisation model's advantages and disadvantages. For example, although the model was quite good at detecting neutral remarks, it sometimes had trouble telling the difference between mildly negative and positive attitudes.

**Table 1: Classification Report**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1-score |
| Positive | 0.50 | 0.50 | 0.50 |
| Neutral | 0.67 | 1.00 | 0.80 |
| Negative | 1.00 | 0.50 | 0.67 |
| Accuracy |  |  | 0.67 |
| Macro Avg | 0.72 | 0.67 | 0.66 |
| Weighted Avg | 0.72 | 0.67 | 0.66 |

We will be using something known as Word Cloud to visualize the words which are shown in the **Figure 5** and **Figure 6**. The most common terms in a given text passage are shown graphically in a word cloud, which is a visual representation of text data. Each word's magnitude in a word cloud corresponds to how frequently it appears in the text. The greater and bolder a term appears in the word cloud, the more often it occurs in the text.

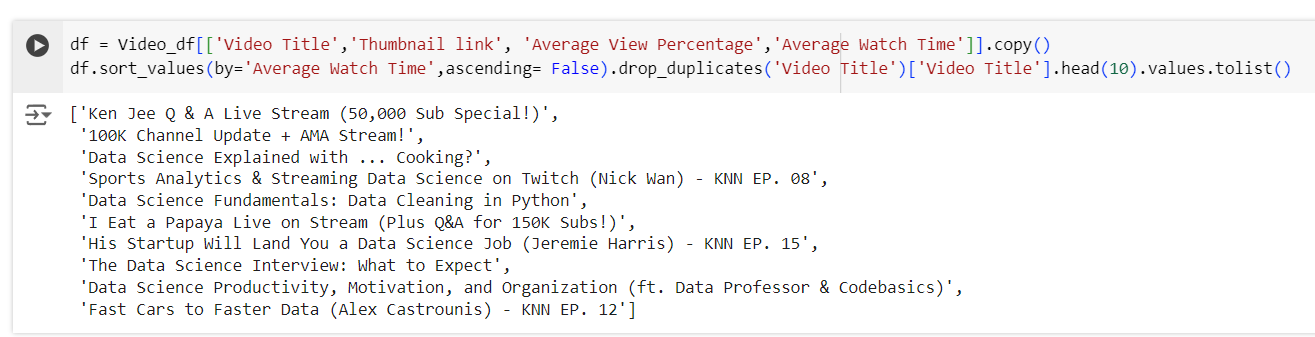
The most significant or frequently used words in a document, a group of documents, or any other textual data source can be quickly and easily viewed with word clouds. They are especially helpful in pointing out important ideas, subjects, or feelings that are expressed in the text.



**Fig 5: Words Used in Positive Comments**



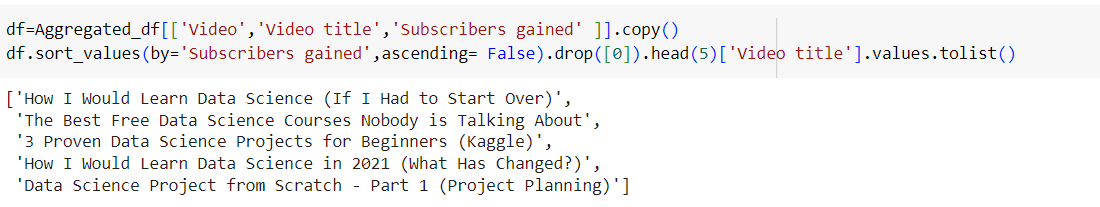
**Fig 6: Words Used in Negative Comments**



**Fig 7: Video titles with words Live, Stream, Fundamentals and Data Science gets the most traffic**

The kinds of YouTube video names and thumbnails that draw the most views are depicted in this **Figure 7**. In particular, it finds that keywords like "Live," "Stream," "Fundamentals," and "Data Science" in video titles significantly influence viewer engagement. According to the data, these terms are popular with viewers and are probably connected to kinds of content that offer specialised knowledge areas, real-time involvement (such as live streaming), or instructional value.

Content producers can utilise this information to deliberately select titles and create thumbnails that use these keywords in order to increase traffic and reach.



**Fig 8: Video names that led to the most growth**

This **Figure 8** focuses on the names of the videos that helped YouTube channels expand the most. It implies that the expansion of the subscriber base and interaction is fueled by videos that use particular terms or relate to particular themes. The trends that have been highlighted show how video naming conventions help match material to the tastes and expectations of viewers. These findings highlight how crucial it is to optimise video metadata in order to increase visibility and channel expansion.

The strategic significance of video titles and themes in affecting YouTube channel success and user engagement is emphasised by both figures.

**Conclusion**

Using an organised process of data collecting, cleaning, sentiment analysis, and topic modelling, this study created an information retrieval framework for examining user sentiment and thematic material in YouTube comments. The analysis's high sentiment classification accuracy using transformer-based models demonstrated how well sophisticated NLP techniques handle casual social media discourse. The findings offer useful information about emotional tone, audience engagement, recurring themes, along with practical suggestions for marketers and content producers.