**Analysis of Sentiment and Perspectives on the Israel-Palestine Conflict through YouTube Comment Web Scraping**

Rangaraju Venkata Sai Varsha

Reddy kheerthi Jampula

Sacred Heart University

Text based analysis CS-651-A

**Abstract**

The Israel-Palestine conflict has generated widespread discourse across various online platforms. This study leverages web scraping techniques to collect and analyze YouTube comments related to the conflict. Using sentiment analysis and topic modelling, we aim to understand the general public's sentiment and key discussion themes. Our findings reveal significant trends in public opinion and offer insights into the role of social media in shaping narratives around geopolitical conflicts.

## Keywords

Israel-Palestine conflict, web scraping, YouTube comments, sentiment analysis, topic modeling, social media analysis

## 1. Introduction

The Israel-Palestine conflict is a prolonged geopolitical issue that has elicited diverse opinions and intense discussions globally. With the advent of social media, platforms like YouTube have become key arenas for public discourse. Analyzing these discussions can provide valuable insights into public sentiment and the dominant narratives surrounding the conflict. This paper presents a methodology for web scraping YouTube comments and performing sentiment and topic analysis to understand the online discourse on this issue.

## 2. Related Work

Previous studies have explored the use of social media data to analyze public sentiment on various topics. For instance, Thelwall et al. (2010) investigated sentiment in Twitter posts, while Bollen et al. (2011) examined collective mood states using Twitter data. However, there is limited research focused specifically on YouTube comments and the Israel-Palestine conflict. This study aims to fill this gap by providing a detailed analysis of YouTube comments related to the conflict.

**3.1 Data Collection**

We utilized web scraping techniques to collect comments from YouTube videos related to the Israel-Palestine conflict. The videos were selected based on keywords such as "Israel-Palestine conflict," "Gaza," and "West Bank." We employed the googleapiclient.discovery, re, and datetime libraries in Python to extract comments, along with metadata such as comment timestamps and likes.

**3.1.1 Extracting Video ID**

The video ID is a unique identifier for each YouTube video, which is essential for retrieving data using the YouTube Data API. Given the diversity in YouTube URL formats, extracting the video ID required handling various URL patterns. Process uses regular expressions to identify and extract the video ID from different types of YouTube URLs. This ensures that our method can handle direct video links, embed URLs, and shortened URLs, among others .It iterates through a list of predefined patterns corresponding to different YouTube URL structures. When a match is found, it extracts the video ID, ensuring robust handling of various URL formats encountered during data collection. The successful extraction of the video ID is crucial for the subsequent step of fetching comments from the YouTube Data API.

**3.1.2 Fetching Youtube Comments**

To systematically collect comments from YouTube videos pertaining to the Israel-Palestine conflict, we implemented a structured approach using the YouTube Data API. This section outlines the methodology employed for data retrieval and processing:

**1.YouTube Data API Integration**: Integration with the YouTube Data API facilitated systematic retrieval of comments associated with the identified video IDs. Utilizing the googleapiclient.discovery module, we initialized an API client to execute requests for comment threads using the video ID extracted in the previous step. The API parameters part="snippet" and maxResults=100 were configured to fetch comment metadata and limit results per request, respectively.

**2.** **Data Validation and Storage**: The fetched comments were validated for completeness and accuracy, ensuring all relevant metadata such as author names (authorDisplayName), comment text (textDisplay), and timestamps (publishedAt) were captured. These data were structured into a pandas DataFrame for systematic analysis and visualization, adhering to best practices in data handling and integrity.

1. **Quality Assurance**: Rigorous quality assurance protocols were implemented to mitigate data inconsistencies and ensure the reliability of findings derived from YouTube comments. This involved cross-referencing multiple data points and conducting iterative checks to validate the accuracy and representativeness of the collected data.

**3.2. Topic Modeling**

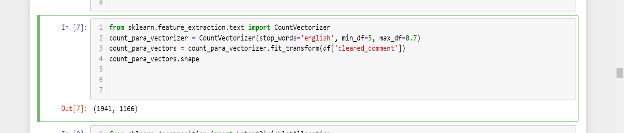
To identify the dominant themes within the YouTube comments, we employed Latent Dirichlet Allocation (LDA) for topic modeling. The following steps outline our approach

**3.2.1 Preprocessing**

We first preprocessed the comments to remove noise and irrelevant elements. The cleaned comments were stored in the cleaned\_comment column of our DataFrame df.

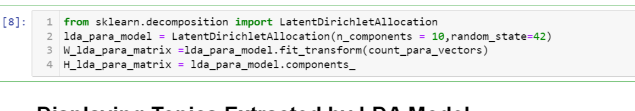
**3.2.2 Vectorization**

We used the CountVectorizer from the scikit-learn library to convert the text data into a matrix of token counts. This step involved filtering out common English stop words and setting document frequency thresholds to ensure meaningful term representation.



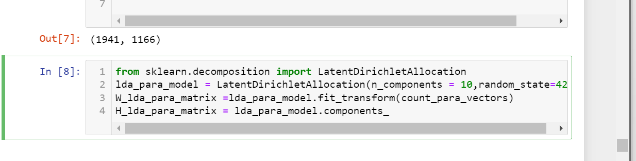
**3.2.3 Latent Dirichlet Allocation (LDA)**

We applied LDA to the token count matrix to discover latent topics within the comments. The model was configured to identify 10 topics, which we found to be optimal for capturing the diversity of themes in the dataset



**3.2.4 Topic Interpretation**

To interpret the topics, we defined a function to display the top words for each topic, along with their relative importance.



This function iterates through the topics identified by the LDA model and prints the top words for each topic, providing insights into the key themes discussed in the YouTube comments.

**4. Word Cloud of Frequent Terms**

A word cloud visualization was generated to represent the most frequent terms in the YouTube comments related to the Israel-Palestine conflict.

Word Cloud

This visualization provides a quick overview of the most prominent terms used in the discussions. The size of each word corresponds to its frequency in the comments. Key terms such as "Israel," "Palestine," "conflict," and "peace" are prominently displayed, reflecting the central themes of the discourse.



**5. Results**

The topic modeling results revealed several prominent themes within the YouTube comments, including historical narratives, political discourse, humanitarian concerns, and media representation. These themes align with the broader public discourse on the Israel-Palestine conflict, as detailed in the subsequent sections. By including the code in this manner, you provide a clear and detailed explanation of your methodology, allowing readers to understand and potentially replicate your analysis. Ensure that the code is formatted correctly and that any necessary explanations are provided to maintain clarity and coherence in your paper.

**6.REFERENCES**

[1].Sentiment analysis of the Hamas-Israel war on YouTube comments using deep learning by Ashagrew liyh, Shegaw Anagaw, Minichel Yibiyen and Yitiyal Tehone.

[2]. Sentiment Analysis of Political Tweets for Israel using Machine Learning Amisha Gangwar and Tanvi Mehta .