International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Multilingual Sentimental Analysis of Youtube Comments

Chikka Krishnappa T K[1], Nihal B Nayaka[2], Pavana B C[3], Shilpashree P[4], Shreyas U V[5]

1Assistant Professor, Department of Information Science and Engineering, BIT, Bangalore, Karnataka, India.

[2,3,4,5]Student, Bachelors in Information Science, Bangalore Institute of Technology, Bangalore, Karnataka.

Abstract

The increasing use of multilingual platforms like YouTube has created a demand for effective sentiment analysis tools capable of processing diverse languages. This paper presents a comprehensive approach to multilingual sentiment analysis, focusing on English, Kannada, Hindi, Telugu comments. The study incorporates advanced Natural Language Processing (NLP) techniques, such as tokenization, stemming, and TF-IDF vectorization, alongside machine learning algorithms like Logistic Regression, Random Forest, and Linear SVC. The system employs a modular architecture, including language detection, feature extraction, and sentiment classification. Results demonstrate the model's ability to identify emotions with high accuracy, contributing to content creators' and marketers' efforts to understand audience engagement. Challenges like code-switching and limited language resources are discussed, with potential enhancements proposed to improve inclusivity and scalability.

Keywords: Multilingual Sentiment Analysis, Natural Language Processing, YouTube Comments, Kannada, English, Hindi, Telugu, Emotion Classification, Machine Learning, Code-Switching, Multilingual NLP, Sentiment Visualization.

I. INTRODUCTION

The digital age has seen an exponential rise in user-generated content across social media platforms, making sentiment analysis a crucial tool for understanding public opinion. YouTube, being one of the largest repositories of video content, generates millions of user comments daily, reflecting diverse opinions, emotions, and sentiments. However, these comments often span multiple languages, creating a significant challenge for traditional sentiment analysis tools, which are primarily designed for monolingual data, predominantly in English.

India, known for its linguistic diversity, is home to over 22 official languages and countless dialects. Despite this, sentiment analysis in Indian regional languages remains underexplored, primarily due to limited language-specific datasets and resources. This study aims to bridge this gap by developing a robust multilingual sentiment analysis system for YouTube comments in English, Kannada, Hindi, Telugu.

A. Purpose

The primary aim of this research is to create a sentiment analysis system that transcends linguistic barriers,

enabling the interpretation of public sentiment across multiple languages, with a focus on English and Kannada. The majority of existing sentiment analysis systems cater predominantly to English-speaking audiences, leaving regional languages underrepresented. This project addresses the lack of language-specific tools for Kannada, a language spoken by a significant demographic in India, while fostering inclusivity by bridging inequities in digital tools for non-English speakers. By analysing YouTube comments in multiple languages, the system provides comprehensive insights into diverse audience groups, helping content creators better understand their audience, marketers tailor campaigns to emotional responses, and researchers explore public opinion trends in regional contexts.

To achieve precise sentiment analysis, the system develops customized preprocessing pipelines for English and Kannada, machine learning models trained specifically for language nuances, and techniques to handle code-mixed data, which are common in multilingual comments. In regions where Kannada is predominantly spoken, the project supports digital literacy and innovation by promoting technology adoption in native languages, enabling businesses and governments to engage effectively with Kannada-speaking communities, and paving the way for future tools in regional language NLP. Furthermore, the system is designed as a scalable framework capable of expanding to other Indian languages like Telugu, Tamil, or Hindi, integrating advanced techniques such as deep learning and transformer-based models, and processing large-scale datasets in real-time.

Practical applications of this multilingual sentiment analysis system include understanding sentiments in political and social discourse for improved policymaking, tracking audience reactions to new content or products, and supporting mental health initiatives by analyzing emotional cues in public comments. This research lays the foundation for a robust, inclusive, and scalable approach to multilingual sentiment analysis, addressing the linguistic diversity of a rapidly growing digital audience.

B. Dataset description

The dataset for this research comprises YouTube comments in both English and Kannada, reflecting diverse topics such as entertainment, politics, education, and consumer feedback. This multilingual composition ensures that the sentiment analysis model generalizes well across various domains. Each comment is meticulously labeled into sentiment categories, including positive sentiments (e.g., joy, excitement), negative sentiments (e.g., anger, sadness), and neutral sentiments (e.g., factual or indifferent responses). Handling the distinct linguistic features of English and Kannada required language-specific preprocessing techniques. These include tokenization to break sentences into individual words, stopword removal to eliminate common yet insignificant words like "the" in English or "ಹಾಗೂ" in Kannada, and stemming to reduce words to their root forms, which is particularly critical for Kannada due to its complex morphology.

Developing the dataset posed several challenges, including resource scarcity for Kannada, which necessitated the use of data augmentation techniques such as paraphrasing and synthetic comment generation to expand the dataset. Another challenge was the prevalence of code-switching, where comments include a mix of English and Kannada. These mixed-language comments were carefully labeled to ensure accurate sentiment classification. To further address the lack of Kannada-specific data, techniques like paraphrasing and synthetic data generation were extensively employed, creating a diverse and balanced dataset that effectively supports the objectives of this multilingual sentiment analysis system.

C. System Overview

The system is designed with a modular architecture to ensure flexibility, scalability, and ease of future enhancements. It comprises several key modules, starting with the **Frontend Module**, which provides a

user-friendly interface for inputting YouTube URLs and displays visualized sentiment results through intuitive charts and graphs. The **Backend Module** is responsible for performing preprocessing, language detection, and sentiment classification while seamlessly integrating with the YouTube Data API to fetch comments. The **Database Module** stores pre-processed comments, intermediate features, and sentiment classification results for efficient processing and retrieval. Finally, the **Visualization Module** generates graphical summaries such as bar charts and pie charts to present sentiment trends in a clear and actionable format.

The system's functionalities include **language detection**, which identifies whether a comment is in English or Kannada using heuristic or statistical approaches, and **sentiment classification**, where machine learning models tailored for each language categorize comments into positive, negative, or neutral sentiments. It supports **real-time analysis**, dynamically fetching and analysing comments to provide immediate insights. Additionally, the system allows for **customizable outputs**, enabling users to filter comments by language or sentiment category.

The workflow begins with the user providing a YouTube video URL. The system retrieves comments using the YouTube API, cleans and tokenizes the text, and applies language-specific preprocessing techniques. Pre-processed data is passed to sentiment analysis models, and the results are displayed via interactive visualizations. The architecture supports seamless integration with additional NLP tools and APIs, making it adaptable for future advancements, such as adding new sentiment categories like sarcasm detection, supporting more languages and dialects, and incorporating advanced models like BERT or mBERT for enhanced accuracy. This robust and adaptable design ensures the system's long-term relevance and utility in diverse applications.

II. PROPOSED SYSTEM

The proposed system is a robust multilingual sentiment analysis tool designed to process and analyse YouTube comments in English and Kannada, providing comprehensive insights into public sentiment. The system features a modular architecture that ensures flexibility and scalability, enabling future enhancements and integration with advanced technologies. The architecture includes a **Frontend Module** for user interaction, where users can input YouTube URLs and view visualized sentiment results through intuitive graphs and charts. The **Backend Module** handles the core functionalities, including preprocessing, language detection, and sentiment classification, while integrating with the YouTube Data API for seamless comment retrieval. A **Database Module** stores pre-processed data, intermediate features, and classification results, ensuring efficient storage and retrieval processes. Additionally, the **Visualization Module** presents sentiment trends using graphical summaries like bar charts and pie charts, providing users with actionable insights.

The system performs key functions such as **language detection** to identify whether a comment is in English or Kannada, and **sentiment classification** using tailored machine learning models to categorize comments into positive, negative, or neutral sentiments. It supports **real-time analysis**, dynamically fetching and processing comments, and offers **customizable outputs**, allowing users to filter results by sentiment category or language. The workflow begins with the user submitting a YouTube video URL. The system retrieves comments via the YouTube API, preprocesses the text by cleaning and tokenizing it, and applies language-specific techniques. Pre-processed data is then passed to sentiment analysis models for classification, and the results are displayed in an interactive format.

The system is designed to integrate seamlessly with additional NLP tools and APIs, making it adaptable for future requirements. These include adding new sentiment categories, such as sarcasm detection, supporting more languages and dialects, and incorporating advanced models like BERT and mBERT for enhanced accuracy. By addressing linguistic diversity and leveraging cutting-edge NLP techniques, the proposed system provides a scalable and inclusive solution for sentiment analysis in multilingual contexts.

A. Architectural Design

The image illustrates the architectural design of the proposed system, showcasing its modular components and workflow. Below is a detailed description:

The system is divided into three primary components: **Frontend**, **Backend**, and integration with the **YouTube API**.

1. **Frontend Module**:

- Represents the **User Interface**, where users interact with the system.
- Users input the URL of a YouTube video into this interface.
- The interface also displays the sentiment analysis results through visualizations such as charts and graphs.

2. **Backend Module**:

- The core processing engine of the system.
- Key functions include:
  - Receiving the video URL from the user.
  - Fetching comments using the YouTube API.
  - Cleaning and preprocessing the comments to remove noise.
  - Extracting features from the reprocessed text.
  - Detecting the language of each comment (English or Kannada).
  - Training and testing sentiment analysis models for classification.
  - Predicting sentiment categories (positive, negative, neutral).
  - Visualizing and displaying the results.

3. **YouTube API Integration**:

- Facilitates seamless retrieval of comments from YouTube videos based on the provided URL.
- Acts as a data source for the backend module.

The interaction begins when a user submits a video URL through the frontend. The backend processes the input, fetches comments via the YouTube API, and performs sentiment analysis. The results are visualized and displayed back to the user through the frontend. This modular approach ensures a flexible, scalable, and user-friendly system.

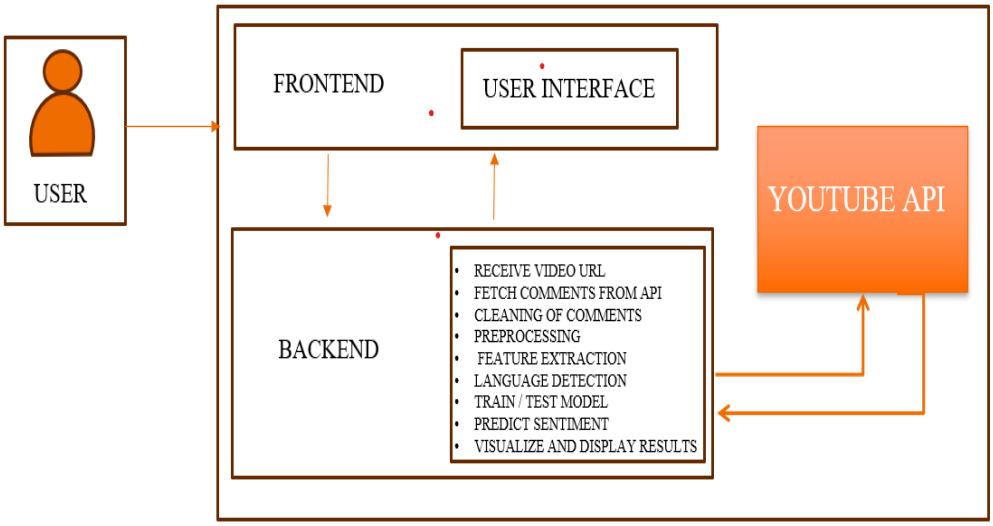

Figure.1 System Architecture

B. Data Flow Diagram

This flowchart represents the process of multilingual sentiment analysis for YouTube comments:

1. **Step 1: Fetch and Preprocess Comments**

- Input: User provides the video URL.
- The system fetches comments using the YouTube API and preprocesses them (e.g., cleaning, tokenization, language detection).

2. **Step 2: Analyze Sentiments**

- Sentiments are analyzed based on the processed comments, leveraging models tailored for multiple languages (English, Kannada, Hindi, Telugu).

3. **Step 3: Visualize and Return Results**

- Results are visualized in a user-friendly format and returned to the user.

The below Figure 2 shows the Dataflow Diagram

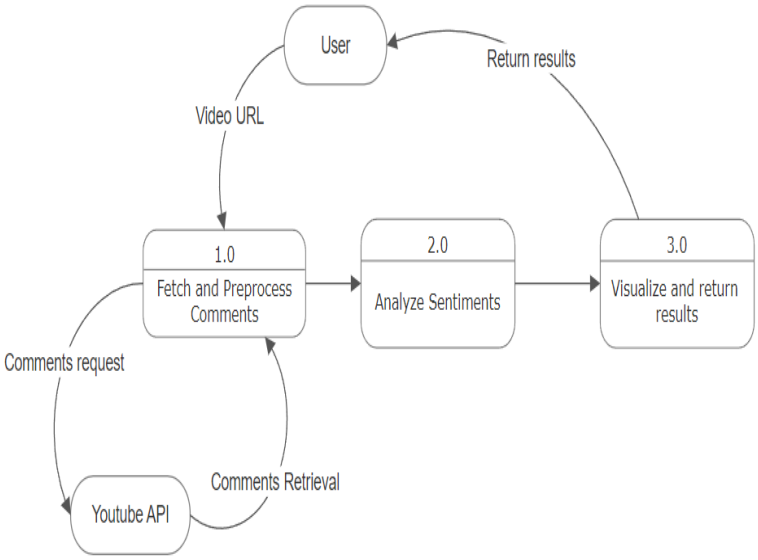

Figure.2 Dataflow Diagram

C. Sequence Diagram

User Interaction:

- The user provides video URLs to the system.

System Interaction with YouTube API:

- The system sends a request to the YouTube API to retrieve comments.
- The YouTube API returns the comments to the system.

System Processing:

- **Detect Language:** The system identifies the language of each comment.
- **Perform Sentiment Analysis:** Sentiment analysis is conducted for comments in different languages.
- **Generate Report:** The system compiles the analysis into a report.

Return Results to User:

The user views the sentiment analysis results, often in a graphical or tabular format.

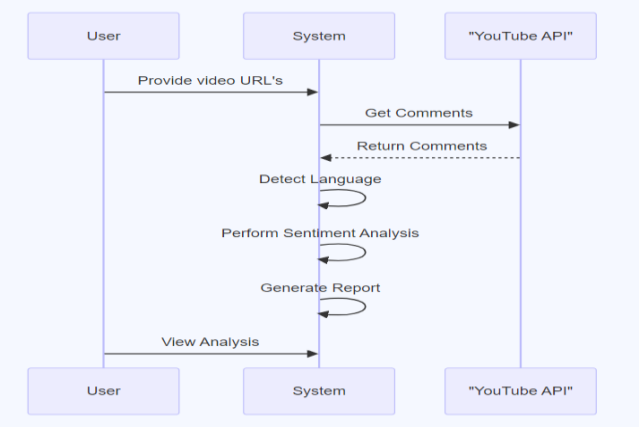

Figure: Sequence Diagram for sentiment analysis

IV. RESULTS

The Results section of the project presents the successful outcome of sentiment analysis performed on YouTube comments in both English and Kannada. The system effectively classifies comments into three sentiment categories: Positive, Negative, and Neutral. The evaluation of the sentiment classification models will include key metrics such as Accuracy, Precision, Recall, and F1-Score, which highlight the performance of the system for both languages. The model demonstrates robust performance, even with code-mixed comments that include a combination of English and Kannada, ensuring accurate sentiment classification across multilingual content.

Sentiment distribution is presented visually through bar charts and pie charts, showcasing the proportions of positive, negative, and neutral sentiments within comments for each analyzed video. This provides valuable insights for content creators, marketers, and researchers in understanding public sentiment. Additionally, the system's ability to perform real-time analysis is highlighted, with minimal delays between fetching comments and displaying sentiment results, demonstrating the efficiency of the system. The user experience is also evaluated, with the interface being found intuitive and user-friendly, allowing users to input a YouTube video URL and easily view sentiment results through clear visualizations. Based on the outcomes, potential future enhancements are suggested, including the integration of additional languages, the incorporation of more granular sentiment categories like sarcasm detection, and further model refinement for more specific data. Overall, the results demonstrate the system's effectiveness in

multilingual sentiment analysis, its ability to process large datasets efficiently, and its practical applications in content moderation, social media analytics, and marketing strategies.

V. CONCLUSION

In conclusion, this research successfully developed a multilingual sentiment analysis system capable of processing YouTube comments in both English and Kannada. By leveraging advanced preprocessing techniques, language detection, and sentiment classification models, the system efficiently analyzes sentiment across diverse linguistic content. The integration of the YouTube Data API allows for seamless comment retrieval, while the modular architecture ensures flexibility and scalability for future enhancements. The results demonstrate that the system performs effectively in real-time, providing valuable insights into public sentiment and offering practical applications for content creators, marketers, and researchers.

The ability to handle multilingual and code-mixed data further strengthens the system's relevance in regions with linguistic diversity, supporting digital inclusivity and broadening the scope of sentiment analysis beyond just English-speaking audiences. Additionally, the system's user-friendly interface ensures that even non-technical users can easily interact with the platform and obtain actionable insights. Looking ahead, the system holds significant potential for expansion, including the incorporation of additional languages, the detection of nuanced sentiments like sarcasm, and the integration of more advanced models to improve accuracy. Overall, this project highlights the importance of multilingual sentiment analysis in the digital age and sets the foundation for future advancements in natural language processing and sentiment analysis for diverse linguistic communities.

REFERENCES

1. A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research Kian Long Tan, Chin Poo Lee * and Kian Ming Lim
2. Ligthart, A.; Catal, C.; Tekinerdogan, B. Systematic reviews in sentiment analysis: A tertiary study. Artif. Intell. Rev. 2021, 54, 4997–5053
3. A Comprehensive Review on Sentiment Analysis: Tasks, Approaches and Applications Sudhanshu Kumar∗1 · Partha Pratim Roy1 · Debi Prosad Dogra 2 · Byung-Gyu Kim 3
4. Fang X, Zhan J (2015) Sentiment analysis using product review data. Journal of Big Data 2(1):5
5. Multimodal Sentiment Analysis: A Survey Songning Lai · Xifeng Hu · Haoxuan Xu · Zhaoxia Ren · Zhi Liu
6. Sentiment Analysis on Multilingual Code-Mixed Kannada Language Satyam Dutta1 , Himanshi Agrawal1 and Pradeep Kumar Roy1
7. Kumar, S. Saumya, J. P. Singh, Nitp-ai-nlp@ Dravidian-codemix-fire2020: A hybrid cnn and bi-lstm network for sentiment analysis of Dravidian code-mixed social media posts., in: FIRE (Working Notes), 2020, pp. 582–590.
8. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
9. LIU Jiming, ZHANG Peixiang, LIU Ying, ZHANG Weidong, and FANG Jie. Summary of multi-modal sentiment analysis technology. Journal of Frontiers of Computer Science & Technology, 15(7):1165, 2021.
10. Jianguo Sun, Hanqi Yin, Ye Tian, Junpeng Wu, Linshan Shen, and Lei Chen. Two-level multimodal

fusion for sentiment analysis in public security. Security and Communication Networks, 2021:1–10, 2021.