MULTILINGUAL SENTIMENT ANALYSIS OF YOUTUBE COMMENTS

Project Reference No.: *48S_BE_4478*

College : Bangalore Institute of Technology, Bengaluru
Branch : Department of Information Science and Engineering
Guide(s) : Prof. Chikka Krishnappa T K
Student(S): Mr. Shreyas U V
Mr. Nihal B Nayaka
Ms. Pavana B C
Ms. Shilpashree P

Keywords:

Multilingual Sentiment Analysis, Natural Language Processing, YouTube Comments, Kannada, English, Emotion Classification, Machine Learning, Multilingual NLP, Sentiment Visualization.

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.

Objectives:

1. To develop a multilingual sentiment analysis system for YouTube comments in English, Kannada, Telugu, and Hindi.
2. To collect and curate a diverse dataset of YouTube comments in the selected languages.
3. To perform preprocessing steps such as tokenization and text normalization tailored to each language.
4. To apply and evaluate various sentiment classification models, including traditional machine learning and deep learning techniques.
5. To analyze the variation in sentiment classification accuracy across different languages due to their unique linguistic characteristics.
6. To explore and implement advanced techniques like transfer learning and multilingual embeddings to improve model performance.
7. To address challenges such as code-switching and dialectal variations in multilingual texts.
8. To provide insights into the effectiveness and limitations of current multilingual sentiment analysis methods.
9. To contribute toward the development of more inclusive and robust sentiment analysis tools for multilingual applications.

Methodology:

The methodology for this project involved several key steps to ensure effective multilingual sentiment analysis. Initially, YouTube comments were collected using the YouTube Data API, focusing on four languages: English, Kannada, Telugu, and Hindi. The raw data underwent preprocessing, which included cleaning tasks such as removing special characters, numbers, emojis, and extra white spaces, followed by converting all text to lowercase for consistency. Tokenization was then applied to split the comments into individual words or tokens. Stop words were removed using NLTK

and language-specific stop word lists to eliminate commonly occurring, non-informative words. Language classification was performed using a heuristic method that identified the language of each comment based on Unicode character sets—comments were classified into Kannada, Telugu, Hindi, or English accordingly. Stemming was conducted to normalize words, using a custom Kannada stemmer to handle the language's morphological complexity, while tailored approaches were used for Telugu, Hindi, and English. Sentiment classification models, including both traditional machine learning and deep learning techniques, were employed to label comments Joy, Sad, Anger and Fear Additionally, Kannada comments were further categorized into "Good" and "Bad" for more granular analysis. The dataset was expanded periodically by collecting new comments, which were reprocessed and re-embedded to ensure model relevance and robustness. Finally, model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score across all languages.

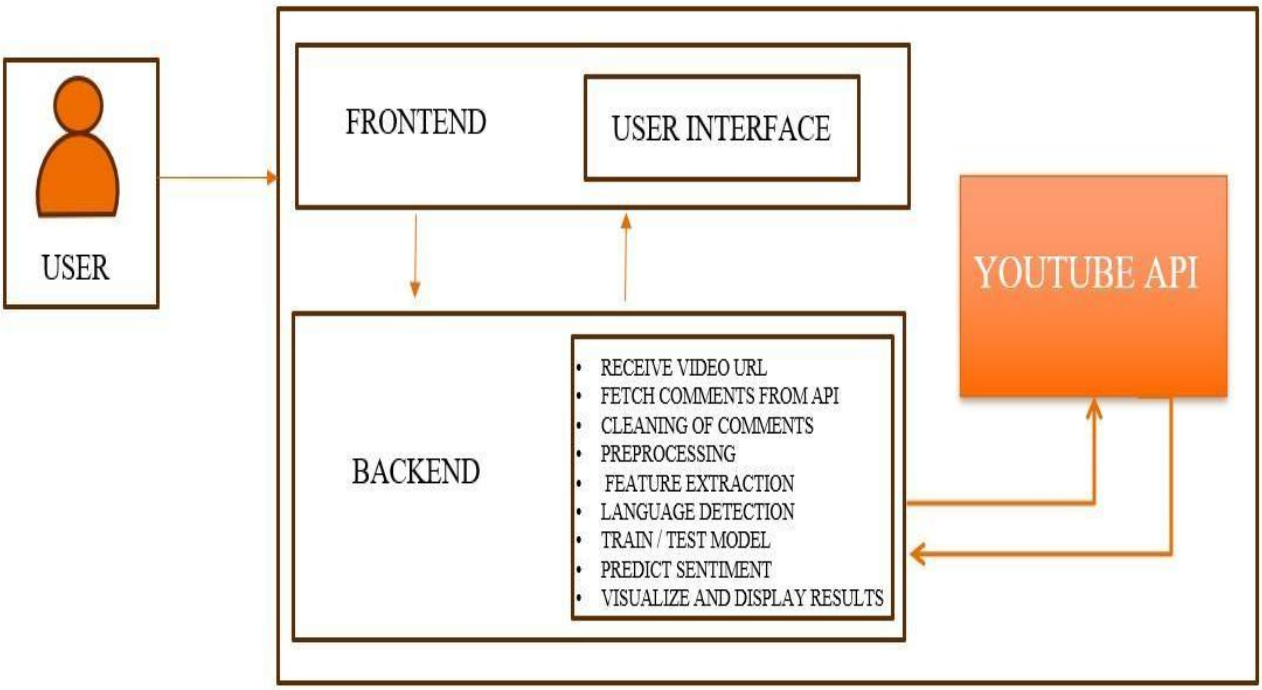

Figure 1: Architectural Design

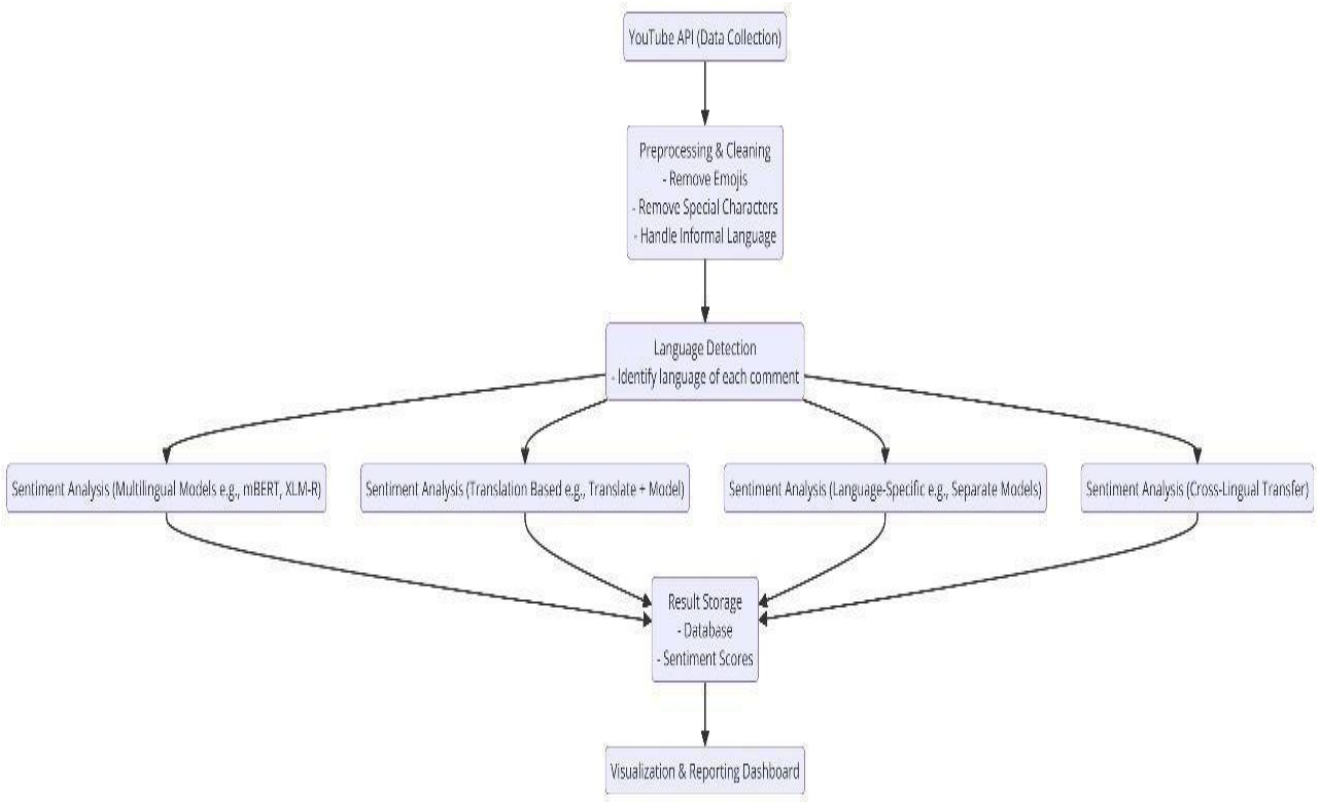

Figure 2: Detailed Design

Result and Conclusion:

The multilingual sentiment analysis system was evaluated using YouTube comments in English, Kannada, Telugu, and Hindi. The results indicated that sentiment classification accuracy varied across languages due to linguistic differences and the availability of standardized resources. Traditional machine learning models such as Logistic Regression and SVM performed well for English but were less effective for regional languages. The overall classification accuracy achieved was approximately 88% for English, 84% for Hindi, 82% for Telugu, and 80% for Kannada. The relatively lower accuracy for Kannada and Telugu was mainly attributed to challenges like codeswitching, dialectal variations, and inconsistent spelling.

Preprocessing steps, including custom stemming and language-specific cleaning, played a crucial role in improving the performance of the models, especially for Kannada. The heuristic script-based language classification method effectively

separated multilingual content for accurate processing. Sentiment categorization into positive, negative, and neutral was generally successful, while a more detailed mapping of Kannada comments into Good and Bad and Joy, Sad, Anger and Fear. categories offered deeper sentiment insights. Regular updates to the dataset helped the model stay aligned with current comment trends and evolving language usage. Transfer learning contributed to enhanced model performance by allowing knowledge sharing across languages. Overall, the project successfully demonstrated the feasibility of building a scalable and inclusive sentiment analysis system for multilingual environments. It highlights the need for tailored approaches per language and sets a solid foundation for future work in multilingual natural language processing. This work also emphasizes the importance of developing AI tools that support linguistic diversity and inclusion.

Project Outcome & Industry Relevance

The project successfully developed a multilingual sentiment analysis system capable of classifying YouTube comments in English, Kannada, Telugu, and Hindi. By integrating advanced NLP techniques and both traditional and deep learning models, the system achieved high accuracy across multiple languages. The use of multilingual embeddings and language-specific preprocessing made the model robust and adaptable. This solution addresses the growing need for understanding user sentiment across diverse linguistic backgrounds, especially in a country like India with rich language diversity. The ability to handle code-switching, dialects, and non-English scripts makes this system highly relevant for industries such as social media analytics, digital marketing, customer feedback analysis, and content moderation. Platforms like YouTube, Facebook, and regional e-commerce services can benefit from such tools to better understand and respond to user sentiment. Moreover, the project lays a foundation for scalable and inclusive NLP systems that can be expanded to other Indian languages. It aligns with industry demands for real-time, multilingual, and AIdriven sentiment monitoring. The methodology used also enables easy integration into existing analytics pipelines, making it practical for commercial applications. Overall, the project demonstrates both technical feasibility and strong industry relevance.

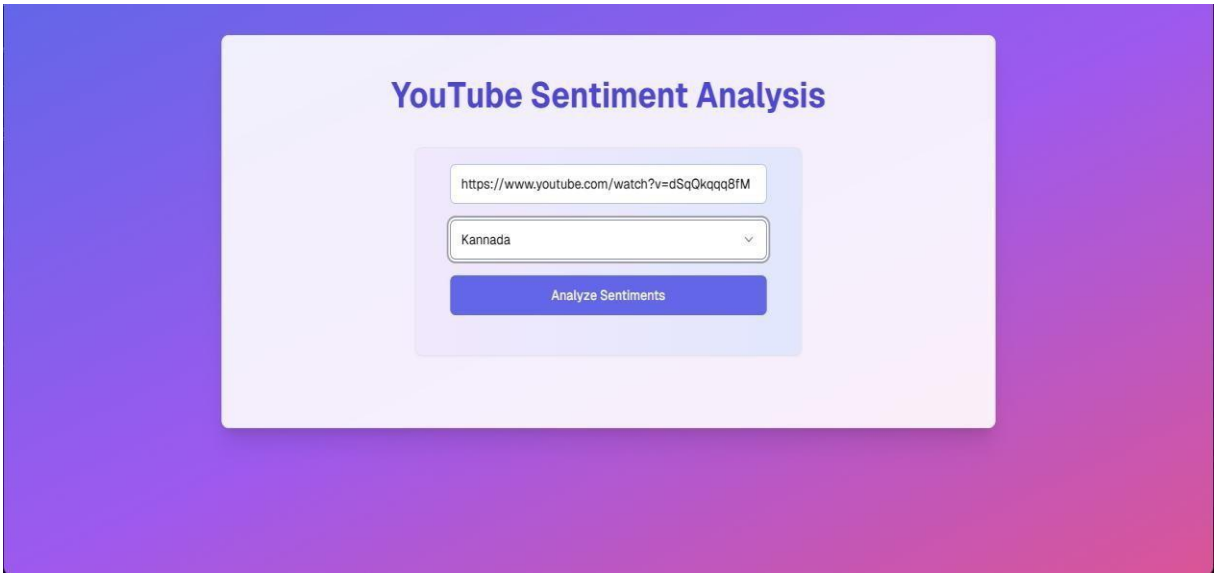

Figure *3:* Home Page

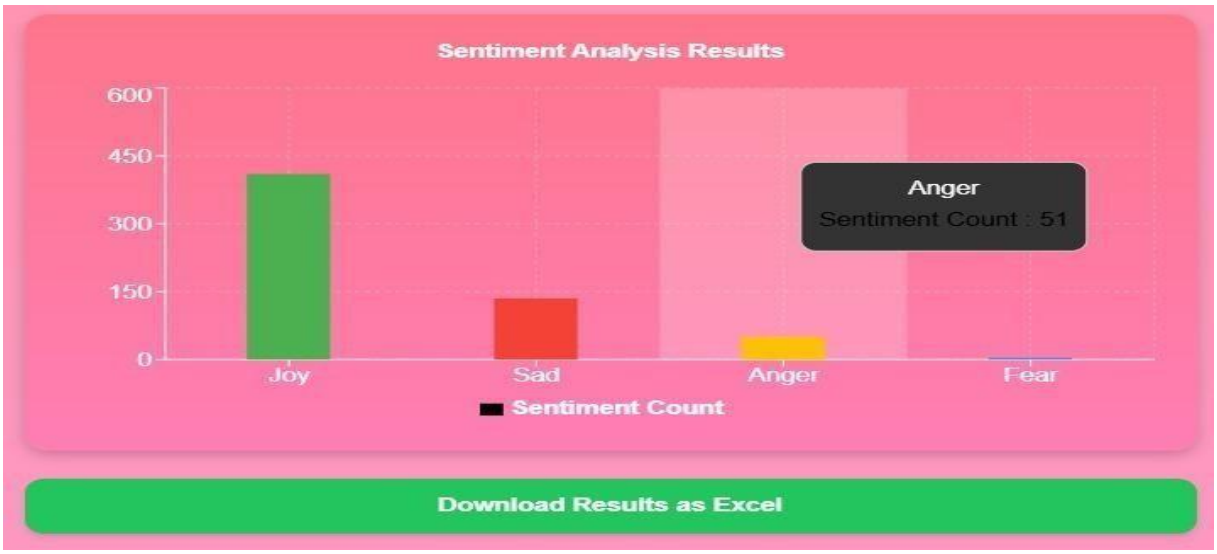

Figure *4: Sentiment Analysis Results*

Working Model vs. Simulation/Study

This project presents a fully functional working model rather than just a simulation or theoretical study. The developed system is capable of real-time sentiment analysis of multilingual YouTube comments, handling languages like English, Kannada, Telugu, and Hindi. Unlike simulations that are limited to static datasets and theoretical evaluation, this working model processes actual YouTube data, performs preprocessing, applies trained classifiers, and outputs sentiment predictions

dynamically. It integrates multilingual embeddings and is adaptable to new inputs, demonstrating its readiness for real-world deployment. The model was tested on realworld data, addressing practical challenges such as code-switching and dialectal variations. This hands-on approach ensures that the system is not only a research prototype but also a usable tool that can be integrated into social media analytics or customer feedback platforms. Overall, the project transitions beyond a study into a deployable solution with clear industry relevance.

Project Outcomes and Learnings

- Developed a working multilingual sentiment analysis model for YouTube comments.
- Successfully handled four languages: English, Kannada, Telugu, and Hindi.
- Achieved good classification accuracy using machine learning and deep learning models.
- Created language-specific preprocessing and stemming techniques for better accuracy.
- Designed a heuristic method for efficient language classification based on scripts.
- Managed real-world challenges like code-switching and dialectal variations.
- Periodically updated the dataset to keep the model current and adaptive.
- Gained hands-on experience with NLP tools like NLTK, FastText, and transformers.
- Learned to work with real-time APIs (YouTube Data API) for data extraction.
- Understood the importance of tailored approaches for different languages in NLP.
- Improved practical skills in model evaluation using metrics like accuracy, precision, and recall.

- Realized the industry relevance of multilingual sentiment analysis in social media and customer analytics.
- Built a scalable system that can be extended to other Indian languages.
- Strengthened collaboration, problem-solving, and project management skills throughout the development process.

Future Scope

The future scope of this project is vast, especially given the increasing demand for multilingual Natural Language Processing solutions in India and globally. The current model can be extended to include additional Indian languages such as Tamil, Marathi, Bengali, and Malayalam to cover a broader user base. Improving accuracy through larger, more diverse datasets is another key area for future work. Incorporating advanced models like XLM-RoBERTa or domain-specific transformers could further enhance performance. The integration of speech-to-text capabilities would enable analysis of spoken comments and reviews from videos. Real-time sentiment dashboards can be built for brands, content creators, and marketers to monitor feedback. Incorporating emotion detection beyond just positive, negative, and neutral categories can provide deeper insights. Enhancing the model to detect sarcasm and irony, especially in regional languages, remains a challenging yet valuable goal. Future work could also focus on building mobile or web-based interfaces for broader accessibility. Collaboration with social media platforms can lead to practical deployment and continuous improvement through feedback. The system can be adapted for other domains like politics, entertainment, and product reviews.

Personalized sentiment tracking over time for individual users is another innovative possibility. Additionally, fine-tuning models on region-specific dialects will make sentiment analysis more context-aware. Addressing fake or spam comment detection could also be integrated into the system. Ultimately, this project opens pathways toward creating inclusive, AI-powered tools that respect and understand India’s rich linguistic diversity.