IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 3, March 2025

Decoding YouTube Comments: A Sentiment Analysis Approach

Radhika Bansode[1], Saloni Shivangekar[2], Sanchita Patil[3], Janhavi Patankar[4], Priyanka Javkar[5]
Students, Computer Technology[1-4]
Lecturer, Computer Technology[5]
Sou. Venutai Chavan Polytechnic, Pune, India

Abstract: *This study introduces an advanced sentiment analysis system for YouTube comments, focusing on sentiment classification, sentiment distribution visualization, temporal sentiment analysis, and automated email summaries. The system uses Random Forest and Naive Bayes classifiers to classify comments as neutral, negative, or favorable. It also provides a visualization function for trend analysis and temporal analysis. The automated email feature simplifies sending sentiment analysis reports, ensuring stakeholders receive quick, actionable insights. This comprehensive approach improves stakeholder reactions to viewer feedback.*

Keywords: Sentiment analysis, analysis automated, analysis visualization, automated, email, comments, distribution

I. INTRODUCTION

This research aims to introduce a comprehensive sentiment analysis framework for YouTube comments, addressing the unique challenges posed by these comments. The framework comprises four interlinked components:

1. sentiment classification
2. visualization of sentiment distribution
3. temporal sentiment analysis
4. automated email summaries

The Random Forest algorithm is used for sentiment classification, achieving a nuanced categorization of comments into three distinct categories: positive, negative, and neutral. The Naive Bayes algorithm is used for sentiment classification, achieving a nuanced categorization of comments into positive, negative, and neutral. The framework provides users with intuitive visual representations of sentiment distribution, enabling them to easily grasp prevailing sentiment trends. It also conducts a temporal analysis to discern how sentiments evolve in the changing time, providing stakeholders with critical insights into viewer sentiment dynamics. The system considers factors such as video upload date and comment submission date, enhancing the depth of understanding regarding audience feedback trends. The integration of an automated email functionality ensures that stakeholders receive timely sentiment analysis summaries, enhancing their capacity to respond effectively to viewer feedback. This streamlined mechanism not only facilitates accessibility to sentiment insights but also serves as a testament to the system's user-centric design.

II. LITERATURE REVIEW

The growth of user-generated material on internet platforms has led to notable developments in sentiment analysis in recent years. Numerous methods for sentiment categorization have been investigated in this field of study, including machine learning algorithms such as Random Forest, Naive Bayes, Logistic Regression, and Support Vector Machines (SVM). While Logistic Regression models yield findings that can be understood, Naive Bayes classifiers are well-liked for their ease of use and effectiveness while processing high-dimensional data. In high-dimensional spaces, support vector machines are well-known for their efficiency. They work especially well when paired with kernel functions to manage non-linear interactions. A flexible ensemble learning method, Random Forest is renowned for its accuracy, capacity to manage a high volume of information, and resilience to overfitting

DOI: 10.48175/568

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 3, March 2025

Charts and graphs are examples of visual representations that offer intuitive insights about sentiment distributions. Important insights into how sentiment changes over time can be gained through temporal analysis. Including automated email summaries improves user response and convenience. Sentiment analysis's future depends on combining multilingual support, real-time tracking, and sophisticated Natural Language Processing (NLP). The development of sentiment analysis algorithms for online platforms such as YouTube will be heavily influenced by accessibility, performance optimization, and ethical considerations.

III. METHODOLOGY

A. Research Stage

The research stages of this study involve systematically collecting YouTube comment data, preprocessing the extracted comments for enhanced clarity, analyzing the sentiments to derive meaningful insights, and ensuring transparency and reproducibility of the results to maintain scientific rigor.

Data Collection

- YouTube video URLs were collected from various sources, including popular content categories, trending videos, and user-specified queries.
- Web scraping techniques were employed using Python's Selenium library to extract a large volume of comment data efficiently while ensuring adherence to data collection best practices.
- The collected data was stored securely in a structured format, facilitating easier access and manipulation during analysis.

Data Collection

- YouTube video URLs were collected from various sources, including popular content categories, trending videos, and user-specified queries.
- Web scraping techniques were employed using Python's Selenium library to extract a large volume of comment data efficiently while ensuring adherence to data collection best practices.
- The collected data was stored securely in a structured format, facilitating easier access and manipulation during analysis.

Data Preprocessing

- Extracted comments underwent multiple preprocessing steps to remove inconsistencies and noise that might affect sentiment analysis.
- Stop words, punctuation marks, and special characters were removed to improve dataset quality and standardization.
- Words were stemmed or lemmatized to ensure consistency in analysis, reducing words to their base or root form for better classification accuracy.
- Duplicate comments and spam messages were identified and filtered out to maintain data integrity.

Sentiment Analysis

- The scikit-learn library in Python was utilized to implement the Random Forest algorithm for sentiment classification, chosen for its high accuracy and robustness.
- Features were extracted from preprocessed comments using bag-of-words representation and TF-IDF (Term Frequency-Inverse Document Frequency) to improve classification precision.
- The trained model was rigorously evaluated using cross-validation techniques to assess its classification performance across different datasets and ensure generalizability.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 3, March 2025

Ethical Considerations

- Approval from the Institutional Review Board (IRB) was obtained to guarantee that the research complies with ethical standards and regulatory guidelines.
- Written consent was obtained from human participants whenever necessary, confirming voluntary participation.
- Data privacy and confidentiality were maintained by anonymizing sensitive information in accordance with best practices for data protection.
- Principles of responsible AI were taken into account to reduce algorithmic bias and ensure equitable sentiment classification.

Data Visualization

- Detailed visualizations of sentiment analysis data were created using the Python matplotlib module, which offers graphical representations like bar charts and line graphs for improved interpretability.
- Word clouds representing commonly used terms in the dataset were produced using the wordcloud library, providing information on the most talked-about subjects and sentiment-heavy keywords. To comprehend changes in user attitudes and engagement patterns, sentiment trends across time were compared.

Statistical Analysis & Reproducibility

- Summary statistics, such as mean sentiment scores and sentiment distributions, were computed to provide a quantitative overview of the dataset.
- The reproducibility of the research was ensured by maintaining a version-controlled repository of all code related to data collection, preprocessing, model training, and visualization, available upon request.
- Sensitivity analysis was performed to evaluate the robustness of the sentiment model against variations in input data.

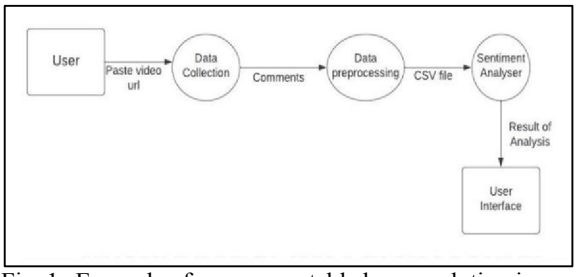

Fig. 1. Example of an unacceptable low-resolution image

B. System Design

The architecture and essential elements of the sentiment analysis system created for efficient YouTube comment analysis are described in the system design.

Architecture of the System

- User Interface: Made to be simple and easy to use, this interface enables users to enter their email addresses and YouTube video URLs with ease.
- Data Collection Module: Ensures efficiency and data integrity by implementing automatic scraping technologies to extract comments related to the supplied YouTube video URLs.
- Data Preprocessing Module: This module improves the quality of the sentiment analysis process by cleaning and preprocessing the extracted comments using linguistic normalization techniques.
- Sentiment Analysis Module: Ensures precise and significant insights by using the Random Forest method to categorize comments into positive, negative, or neutral attitudes.
- Result Generation and Visualization Module: Using the sentiment analysis module's predictions as a guide, this module creates and displays sentiment analysis results as tables, charts, and visual summaries.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 3, March 2025

- Email Sending Module: Enhances usability and accessibility by automating the creation and delivery of an email with four sentiment analysis results to the user's supplied email address.
- Integration Module: Assures effective operation and reduced latency by facilitating smooth data processing and flow between various system components.

C. Development Model

The development model adopted for the implementation of the sentiment analysis system is based on an iterative and incremental approach, specifically the Agile methodology. Agile methodologies emphasize flexibility, collaboration, and rapid feedback, making them well-suited for dynamic and evolving projects such as sentiment analysis.

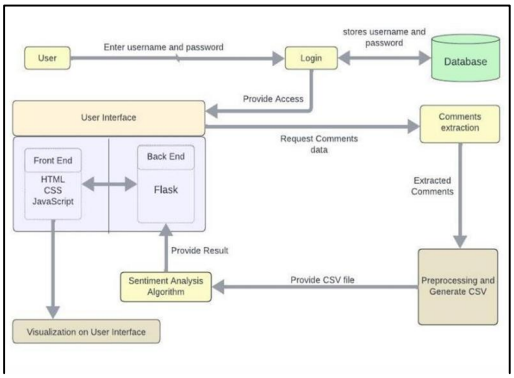

Fig. 2. Development model

D. Agile Development Process

- The development process was organized into short iterations (sprints), generally lasting from one to four weeks, making development modular and efficient.
- Each sprint was dedicated to delivering a specific set of features or functionalities, prioritized according to their importance to the end user and the objectives of the system.
- Continuous feedback mechanisms were established to support enhancements, making sure that every iteration improved existing functionalities and introduced new enhancements based on user preferences.
- Routine testing and validation were performed to guarantee system stability, accuracy, and usability.
- Collaboration among multidisciplinary teams, including data scientists, developers, and UX designers, was encouraged to speed up development and improve system performance.

By implementing the Agile development approach, the sentiment analysis system for YouTube comments gained from improved collaboration, quick iterations, and adaptability to user needs. The iterative aspect of Agile ensured that the system progressed gradually, providing valuable insights to users with each new sprint while smoothly incorporating new requirements.

IV. FUTURE SCOPE

The Comprehensive Sentiment Analysis System for YouTube Comments is a powerful tool for understanding and leveraging audience sentiment. Its potential for expansion and refinement is significant, with potential for multilingual support, fine-grained sentiment analysis, and advanced Natural Language Processing techniques. Real-time sentiment tracking offers immediate feedback for content creators allowing them to adapt their strategies dynamically. Contextual analysis, considering factors like reply chains and video content, can refine sentiment classification accuracy. The system's capabilities to process comments in real-time during live streams provide instantaneous insights, enhancing viewer interaction. Integrating sentiment analysis results with content recommendation systems allows content offerings to align with viewer sentiments and preferences. Accessibility and inclusivity features are crucial for a more inclusive user experience. Ethical considerations and bias mitigation strategies are essential for delivering fair and unbiased sentiment analysis results. Scalability and performance optimization are essential for handling increasing comment

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 3, March 2025

volumes while maintaining optimal responsiveness. These enhancements highlight the system's dynamic potential and promise more robust tools for content creators to engage effectively on the YouTube platform.

V. CONCLUSION

The research presents a Sentiment Analysis System for YouTube comments, which uses advanced techniques, user-centric features, and a Random Forest algorithm to categorize comments into positive, negative, and neutral sentiments. The system offers intuitive visual representations, temporal analysis, and an automated email module for stakeholders to receive timely summaries of sentiment analysis results. The potential for enriching this system is boundless, with multilingual support and fine-grained sentiment analysis. The Integration of state-of-the-art Natural Language Processing (NLP) techniques and real-time sentiment tracking promises more immediate and nuanced insights. Ethical considerations, accessibility, and bias mitigation strategies guide the pursuit of a fair and inclusive sentiment analysis system. Scalability and performance optimization are paramount as the system adapts to the growing demands of an ever- expanding user base. The findings underscore the significance of sentiment analysis in understanding user sentiment and enhancing user experience on online platforms like YouTube. Future research and development include exploring the integration of advanced machine learning techniques and further investigation into the impact of sentiment analysis on user engagement and content performance metrics.

ACKNOWLEDGMENT

We would like to sincerely thank Sou. Venutai Chavan Polytechnic, Pune, our institution, for giving us the chance and means to carry out this study. We would like to express our sincere gratitude to Priyanka Javkar, our mentor, for her essential assistance, encouragement, and direction during this study. Additionally, we are grateful to our friends and peers for their helpful criticism and enlightening conversations, which influenced the course of this study. Last but not least, we are grateful to the International Journal of Advanced Research in Science, Communication, and Technology (IJARSCT) for giving us a venue to publish our work and support the scholarly community.

REFERENCES

[1]. N. Anggraini and M. J. Tursina, "Sentiment Analysis of School Zoning System On Youtube Social Media Using The K-Nearest Neighbor With Levenshtein Distance Algorithm" 2019 7th International Conference on Cyber and IT Service Management (CITSM), Jakarta, Indonesia, pp. 1-4

[2]. M. Alkaff, A. Rizky Baskara and Y. Hendro Wicaksono, "Sentiment Analysis of Indonesian Movie Trailer on YouTube Using Delta TFIDF and SVM", 2020 Fifth International Conference on Informatics and Computing (ICIC), Gorontalo, Indonesia, 2020, pp. 1-5

[3]. K. Goel and K. Batra, "A Deep Learning Classification Approach for Short Messages Sentiment Analysis"2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, pp. 1-3

[4]. J. Li and L. Qiu, "A Sentiment Analysis Method of Short Texts in Microblog," 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), Guangzhou, 2017, pp.776-779, doi:10.1109/CSE EUC.2017.153.

[5]. 2019 7th International Conference on Cyber and IT Service Management (CITSM), Jakarta, Indonesia, 2019, pp. 1-4, doi: 10.1109/CITSM47753.2019.8965407. N. Anggraini and M. J. Tursina, "Sentiment Analysis of School Zoning System On Youtube Social Media Using The KNearest Neighbor With Levenshtein Distance Algorithm."

[6]. The paper "Sentiment Analysis of Indonesian Movie Trailer on YouTube Using Delta TFIDF and SVM" was presented at the Fifth International Conference on Informatics and Computing (ICIC) in 2020 in Gorontalo, Indonesia. It was published on page 1–5, with a doi: 10.1109/ICIC50835.2020.9288579.