TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.02 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.52029

A Machine Learning Approach to Sentiment Analysis of YouTube

Sona Singh[a], Sonika Bhardwaj[b], Ranjeet Kumar Dubey[c]
[a,b] Scholar, Department of Computer Science and Engineering, (AI&ML), KIPM College of Engineering and Technology, U.P., India
[c]Assistant Professor, Department of Computer Science and Engineering, (AI&ML) , KIPM College of Engineering and Technology, U.P., India
sonsingh24705@gmail.com

KEYWORD

YouTube, social media platforms, Natural Language Processing (NLP), machine learning

ABSTRACT

In the digital age, YouTube has emerged as one of the most influential social media platforms, where users actively share their opinions through comments, likes, and dislikes. These interactions provide valuable insights into public sentiment regarding videos, products, and various topics. However, manually analyzing such a vast volume of user-generated data is both time-consuming and inefficient. To address this challenge, sentiment analysis—also known as opinion mining—offers an automated approach to classify user opinions as positive, negative, or neutral. This study focuses on analyzing YouTube comments using Natural Language Processing (NLP) and machine learning techniques to accurately identify user sentiment. The research involves collecting comment data through the YouTube Data API, followed by preprocessing steps such as data cleaning and normalization to enhance data quality. Feature extraction techniques, including TF-IDF and word embeddings, are applied to capture meaningful textual patterns. Various machine learning models are then developed and trained for effective sentiment classification. Finally, the results are visualized and interpreted to provide clear insights into audience opinions and trends..

1. Introduction

In today's digital era, YouTube has become one of the most powerful social media platforms, where millions of users express their opinions through comments, lifes, and dislikes. These comments represent public sentiment about videos, products, or topics. Manually analyzing such massive amounts of data is inefficient and time-consuming. Therefore, Sentiment Analysis, also known as Opinion Mining, is used to automatically determine whether a user's comment expresses a positive, negative, or neorraf opinion This project aims to analyze YouTube comments using Natural Language Processing (NLP) and Machine Learning techniques to identify the sentiment of users effectively. In recent years, the rapid growth of social media platforms has transformed the way people communicate, share information, and express opinions. Among these platforms, YouTube stands out as one of the most widely used video-sharing services, attracting millions of users worldwide. Every day, users interact with content by posting comments, liking or disliking videos, and sharing their views on various topics such as entertainment, education, politics, and product reviews. These user-generated comments serve as a rich source of public opinion and sentiment.

However, the sheer volume of comments generated on YouTube makes manual analysis impractical and inefficient. Understanding user sentiment at scale requires automated techniques that can process large datasets quickly and accurately. This is where Sentiment Analysis, also known as Opinion Mining, plays a crucial role. It involves the use of Natural Language Processing (NLP) and machine learning algorithms to analyze textual data and classify it into

Corresponding Author: Sona Singh, Scholar, Department of Computer Science and Engineering, (AI&ML), KIPM College of Engineering and Technology, U.P., India
Email: sonsingh24705@gmail.com

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.02 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.52029

categories such as positive, negative, or neutral sentiment. The application of sentiment analysis to YouTube comments has significant importance for content creators, businesses, and researchers. It helps in understanding audience reactions, evaluating the success of content, identifying trends, and making informed decisions. For instance, companies can analyze customer feedback on product-related videos, while creators can adjust their content strategies based on viewer sentiment. This research focuses on developing an effective sentiment analysis system for YouTube comments using NLP and machine learning techniques. The study includes data collection through the YouTube Data API, preprocessing of textual data, feature extraction using methods such as TF-IDF and word embeddings, and the implementation of classification models to predict sentiment. The outcomes of this work aim to provide meaningful insights into user opinions and demonstrate the effectiveness of automated sentiment analysis in real-world applications.

1.1 Objectives of the paper

- To collect user comments from YouTube videos using the YouTube Data API.
- To preprocess and clean the text data for accurate analysis.
- To extract meaningful feamres using NLP techniques such as TF-lDP or Word Embeddings
- To build and train machine learning models for sentiment classification.
- To visualize and interpret the results for better understanding of user opinions.

2. Related Works

Sentiment analysis has been extensively explored in the field of Natural Language Processing (NLP), especially for analyzing social media data where users frequently express opinions and emotions. Early research focused on traditional machine learning techniques such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression for sentiment classification. Bo Pang and Lillian Lee [1] conducted one of the pioneering studies on opinion mining using machine learning methods, demonstrating the effectiveness of text classification techniques on sentiment datasets.

Feature extraction techniques such as Bag-of-Words and Term Frequency–Inverse Document Frequency (TF-IDF) were widely adopted to convert textual data into numerical form. Peter D. Turney [2] introduced an unsupervised approach based on semantic orientation using pointwise mutual information, which contributed significantly to early sentiment analysis research. However, these traditional approaches often lacked the ability to capture contextual relationships within text.

With the advancement of deep learning, models such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) gained popularity for sentiment classification tasks. Sepp Hochreiter and Jürgen Schmidhuber [3] introduced LSTM networks, which effectively capture long-term dependencies in sequential data. Similarly, Convolutional Neural Networks (CNN) have been applied to text classification, as demonstrated by Yoon Kim [4], achieving strong performance in sentence-level classification tasks.

More recently, transformer-based models such as BERT have revolutionized NLP by capturing bidirectional context in text. Jacob Devlin et al. [5] proposed BERT, which significantly improved the performance of sentiment analysis tasks across various benchmarks.

Several studies have focused specifically on sentiment analysis of YouTube comments. Apoorv Agarwal et al. [6] analyzed sentiment in social media content, including YouTube, using supervised learning approaches. These works emphasize the importance of preprocessing steps such as tokenization, stop-word removal, and normalization to improve classification accuracy. Other researchers have explored hybrid models combining lexicon-based and machine learning approaches to handle informal language, emojis, and sarcasm present in user comments.

Despite these advancements, challenges such as multilingual content, slang, sarcasm, and noisy data still affect the accuracy of sentiment classification. This study builds upon existing research by applying effective NLP and machine learning techniques to improve sentiment analysis performance for YouTube comments.

3. Problem Statement

YouTube videos receive a large number of comments daily, making it difficult for content creators and organizations to manually understand public opinion. An automated sentiment analysis system can help classify these comments and

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.02 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.52029

provide insights into user perception. YouTube generates an enormous volume of user comments every day, reflecting diverse opinions, emotions, and feedback on videos, products, and social topics. Analyzing this vast amount of unstructured textual data manually is not only time-consuming but also inefficient and prone to inconsistency. As a result, content creators, businesses, and researchers face challenges in understanding audience perception and extracting meaningful insights from user interactions.

To address this issue, there is a need for an automated system that can efficiently process and analyze YouTube comments. Sentiment analysis provides a practical solution by classifying user opinions into categories such as positive, negative, or neutral. Such a system can help stakeholders better understand audience reactions, improve content strategies, and support data-driven decision-making.

4. Proposed Methodology

1. Data Collection

Comments are collected from selected YouTube videos using the YouTube Data API. This step involves retrieving relevant data such as comment text, timestamps, and user information. The collected dataset serves as the foundation for further analysis.

2. Data Preprocessing

The raw text data is cleaned and prepared to improve analysis accuracy. This includes removing special characters, URLs, emojis (if required), stop words, and performing tokenization, stemming, or lemmatization.

3. Feature Extraction

Meaningful features are extracted from the processed text using techniques such as Term Frequency–Inverse Document Frequency (TF-IDF) and word embeddings. These features convert textual data into numerical form suitable for machine learning models.

4. Model Development

Various machine learning algorithms, such as Naïve Bayes, Support Vector Machines (SVM), or Logistic Regression, are implemented and trained on the dataset to classify sentiments. Advanced approaches like deep learning models can also be considered for improved performance.

5. Model Evaluation

The performance of the models is evaluated using metrics such as accuracy, precision, recall, and F1-score to ensure reliable classification results.

6. Visualization and Interpretation

The final results are visualized using graphs and charts to provide a clear understanding of user sentiment distribution and trends across different videos.

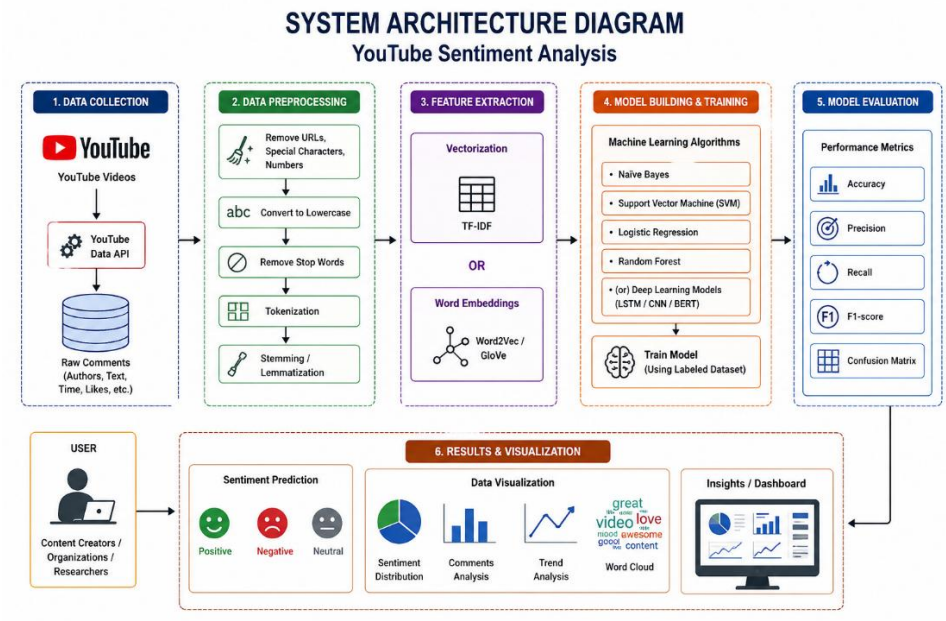

Figure 1: System Architecture Diagram

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.02 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.52029

4.1 Tools aod Technologies Used

- **Programming Language:** Python
- **Libraries:** pandas, numpy, nltk, sklearn, matplotlib, seaborn
- API: YouTube Data API
- **Framework (if used}:** TensorFlow / PyTorch
- **IDE:** Jupyter Notebook /
- VS Code

5. Expected Outcome

The proposed system is expected to effectively analyze and classify YouTube comments into three primary sentiment categories: positive, negative, and neutral. By leveraging Natural Language Processing (NLP) and machine learning techniques, the model will provide accurate and automated sentiment classification, reducing the need for manual analysis.

In addition, the system will generate meaningful statistical and visual insights, such as sentiment distribution, trend analysis, and graphical representations of user opinions. These visualizations will make it easier to interpret large volumes of data and identify patterns in audience behavior.

Furthermore, the developed model will serve as a valuable tool for content creators, businesses, and researchers by helping them understand viewer feedback more effectively. This can support better decision-making, enhance content quality, and improve overall audience engagement and strategy development.

6. Applications

- Brand reputation and product feedback analysis
- Audience engagement analysis for YouTube channels
- Marketing and advertisement planning
- Public opinion monitoring

7. Conclusion

This project presents the development of an automated system for analyzing YouTube comments using Natural Language Processing (NLP) and machine learning techniques. The proposed approach effectively processes large volumes of user-generated data and classifies sentiments into positive, negative, and neutral categories. By automating sentiment analysis, the system reduces the limitations of manual evaluation and provides faster, more consistent insights into user opinions. The results of this study offer a clear understanding of audience sentiment, which can be valuable for content creators, businesses, and researchers in improving decision-making and content strategies. The integration of visualization techniques further enhances the interpretability of the results by highlighting patterns and trends in user feedback. In the future, this work can be extended to include advanced features such as multilingual sentiment analysis, sarcasm detection, and real-time data processing. These enhancements will further improve the system's capability to handle diverse and dynamic social media data, making it more robust and applicable in real-world scenarios.

References

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1–2, pp. 1–135, 2008.
[2] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in Proc. ACL, 2002, pp. 417–424.
[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[4] Y. Kim, "Convolutional neural networks for sentence classification," in Proc. EMNLP, 2014, pp. 1746–1751.
[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[6] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in Proc. Workshop on Languages in Social Media, 2011, pp. 30–38.