Enhancing YouTube Comment Insights: A Machine Learning Approach to Sentiment Analysis

Dipika Raigar[a], Poonam Chapke[b], Pranav Dangare[c], Pratik Dhagude[d] and Varsha Malode[e]
Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune-411043, Maharashtra, India

Keywords: YouTube Comments, Machine Learning, Natural Language Processing, Supervised Learning, Sentence Classification.

Abstract: As YouTube channels grow, they receive large volumes of comments that provide valuable feedback, crucial for understanding audience sentiment and improving engagement. However, existing sentiment analysis approaches focus only on single-language positive/negative classification, and struggle with the informal language in YouTube comments. This project addresses these limitations by building a system to classify comments based on sentiment (positive, negative, neutral) and sentence types, using advanced NLP techniques and machine learning models. Our system will support multilingual analysis, increasing accessibility, and its performance will be evaluated using cross-validation and F1 scores to help creators improve their contents.

1 INTRODUCTION

The rise of social media platforms like Facebook, Twitter, and YouTube has changed how people interact with content. These platforms allow users to communicate, share ideas, and express opinions openly. Among them, YouTube stands out as one of the most popular video sharing platforms. It not only gives users a chance to watch videos but also to share their thoughts through comments. These comments serve as feedback for the video creators, offering them a way to understand how their audience feels about their content. YouTube comments are essential because they often reflect the public's overall opinion or mood. This feedback can help content creators know what their audience likes or dislikes, allowing them to improve their future videos and create content that resonates better with viewers. However, with the massive number of comments posted daily, it becomes difficult to analyse and understand all of them manually. This is where technology, specifically Sentiment Analysis (SA), becomes very useful. Sentiment Analysis is a branch of Natural Language Processing (NLP) that focuses on identifying the emotions or opinions expressed in a piece of text. For instance, it can determine whether a comment is positive, negative, or neutral. However, analysing YouTube comments comes with its own set of challenges. Unlike formal writing, YouTube comments are usually unstructured. People often use slang, abbreviations, emojis, or even misspelled words, making it harder to understand the real meaning behind the text. Also, sentence structures can vary greatly from one comment to another. Some comments might be long and detailed, while others are short and vague. These factors make it tough to process the comments automatically using traditional tools. The goal of this paper is to create a sentiment analysis system that can tackle these challenges effectively. By categorizing YouTube comments based on their sentiment, such as positive or negative, and identifying different types of sentences, the system can offer content creators more detailed and useful

[a] https://orcid.org/0000-0002-9906-9459
[b] https://orcid.org/0009-0009-3682-0952
[c] https://orcid.org/0009-0009-1325-3115
[d] https://orcid.org/0009-0007-2724-3829
[e] https://orcid.org/0009-0000-1744-6410

Raigar, D., Chapke, P., Dangare, P., Dhagude, P. and Malode, V.
Enhancing YouTube Comment Insights: A Machine Learning Approach to Sentiment Analysis.
DOI: 10.5220/0013735100004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 3*, pages 930-933
ISBN: 978-989-758-763-4

feedback. With this information, creators can adapt their content, better engage their audience, and improve the overall quality of their videos. Ultimately, this system aims to make it easier for video creators to understand their audience's emotions and reactions, helping them stay relevant and successful on the platform.

2 RELATED WORK

Pokharel, Rhitabrat, and Dixit Bhatta (2021)." Classifying YouTube comments based on sentiment and type of sentence." This study explores emotion classification of Indonesian YouTube comments using various word embeddings and machine learning models. The CNN method showed the best performance, with an accuracy of 76.2% (Wei and Zhang, 2024). Wei, Zhongliang, and Shunxiang Zhang (2024)." A structured sentiment analysis dataset based on public comments from various domains." This paper introduces a dataset offering diversity and quality, enabling targeted analysis of sentiment classification models, especially in domain-specific contexts (Kumari, Anupriya, et al. , 2024). Ruchita Kumari, et al. (2024)." Comment Analyzer for Sentiment Analysis in Social Media and E-Commerce Platforms." The authors focus on sentiment analysis in social media and e-commerce platforms. Their model effectively classifies comments by sentiment using NLP and machine learning algorithms (Neve, Pachpute, et al. , 2024). Sainath Pichad, et al. (2023)." Analysing Sentiments for YouTube Comments using Machine Learning." This study uses Naive Bayes and SVM to classify sentiments in YouTube comments (Cunha, Costa, et al. , 2019). Singh, R., & Tiwari, A. (2021)." Sentiment Analysis of YouTube Comments." This paper presents a sentiment analysis model that classifies YouTube comments using SVM and Decision Trees. Aditya Neve, Kalpesh Pachpute, Bhimashankar Mathapati, Prerana Thorve (2024)." YouTube Comment Sentiment Analysis." This study explores sentiment analysis for YouTube comments using machine learning techniques, improving engagement through classification methods (Muhammad, Bukhori, et al. , 2019. Al Hujaili, Rawan Fahad, and Wael MS Yafooz (2021)." Sentiment analysis for YouTube videos with user comments." This paper presents a sentiment analysis method for YouTube videos, focusing on user feedback and employing deep learning algorithms for analysis (Alberto, Lochter, et al., 2021). Muhammad, Abbi Nizar, Saiful Bukhori, and Priza Pandunata (2019). " Sentiment analysis of positive and negative YouTube comments using NB-SVM classifier." The paper discusses the hybrid Naive Bayes-SVM classifier for sentiment classification, yielding improved performance for polarity-based classification (AlHujaili, Yafooz, et al. , 2021). Bhuiyan, Hanif, et al. (2017)." Retrieving YouTube video by sentiment analysis on user comment." (Bhuiyan, et al. , 2017), this study focuses on retrieving YouTube videos based on sentiment classification of user comments, emphasizing machine learning models for extracting valuable content (Sungheetha and Sharma, 2020).

3 RELATED GAP ANALYSIS

Several Gaps in Current Research Remain Unaddressed

1. **Domain-Specific Sentiment Analysis** Existing sentiment lexicons often struggle with domain-specific language (Pichad, Kamble, et al. , 2023).
2. **Handling Informal Language and Nuances** Informal language and slang hinder accurate sentiment classification (Pokharel, Bhatta, et al. , 2021).
3. **Big Data Challenges** The growing volume of online data presents challenges in storage and analysis (Pichad, Kamble, et al. , 2023).
4. **Multimodal Sentiment Analysis** Sentiment can be expressed through text, emojis, images, and videos(Alberto, Lochter, et al. , 2021).
5. **Real-Time Sentiment Analysis** Research into efficient and scalable real-time sentiment analysis techniques is essential (Singh and Tiwari, 2021).

4 MODELS

1) Rule-Based VADER

2) Classical ML Naive Bayes, SVM, Logistic Regression, Random Forest

3) Deep Learning CNN, RNN, LSTM, BERT

5 METHODOLOGIES

Our Approach Involves the Following Steps

5.1 Data Collection

YouTube comments are collected using YouTube's API.

5.2 Data Preprocessing

Tokenization, stop-word removal, stemming, and lemmatization are applied.

5.3 Feature Extraction

Word embeddings (Word2Vec or GloVe) and TF-IDF are used.

5.4 Modelling

Machine learning models (Naive Bayes, SVM, CNN) classify comments by sentiment and sentence type.

5.5 Evaluation

Cross-validation and F1 score are used to evaluate model performance.

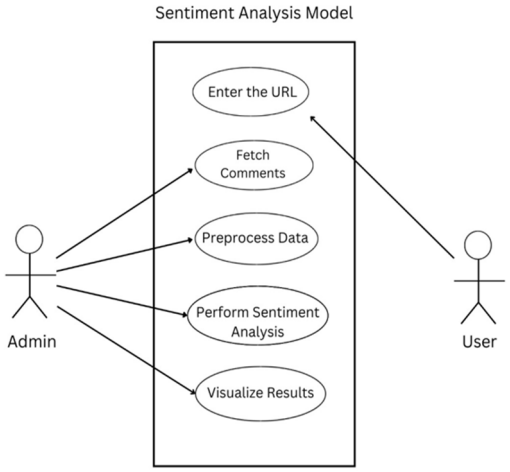

Figure 1. The process of collecting data, preprocessing it, and utilizing classifier models

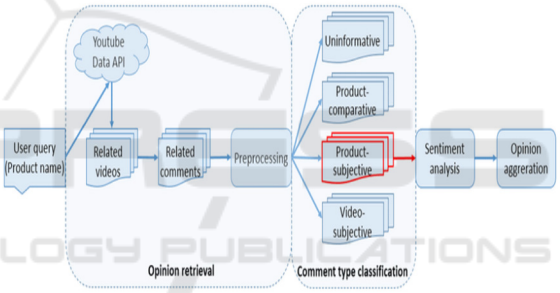

Figure 2. System Architecture

6 APPLICATIONS

6.1 Content Creation

Helps YouTubers identify valuable feedback

6.2 Marketing and Analytics

Assists brands in gauging public sentiment

6.3 Audience Engagement

Enables more meaningful viewer engagement

6.4 Social Media Analysis

Provides insights into trends and opinions

6.5 Spam Detection

Enhances spam filtering capabilities

7 CONCLUSIONS

As YouTube channels grow, they receive large volumes of comments that provide valuable feedback, crucial for understanding audience sentiment and improving engagement. However, existing sentiment analysis approaches focus only on single-language positive/negative classification, and struggle with the informal language in YouTube comments. This project addresses these limitations by building a system to classify comments based on sentiment (positive, negative, neutral) and sentence types, using advanced NLP techniques and machine learning models. Our system will support multilingual analysis, increasing accessibility, and its performance will be evaluated using cross-validation and F1 scores to help creators improve their content.

REFERENCES

Alberto, T. C., Lochter, J. V., & Almeida, T. A. (2021). Tubespam: Comment spam filtering on YouTube. In *2021 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE.

AlHujaili, R. F., & Yafooz, W. M. S. (2021). Sentiment analysis for YouTube videos with user comments. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*. IEEE.

Bhuiyan, H., et al. (2017). Retrieving YouTube video by sentiment analysis on user comment. In *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE.

Cunha, A., Costa, M., & Pacheco, M. (2019). Sentiment analysis of YouTube video comments using deep neural networks. *Springer*, 11508. https://doi.org/10.1007/978-3-030-20912-451

Kumari R., Anupriya, Singh, S., & Shukla, H. (2024). Comment analyzer for sentiment analysis in social media and e-commerce platforms. *International Journal of Innovative Research in Computer Science Technology (IJIRCST), 12*(Special Issue-1), March 2024.

Muhammad, A. N., Bukhori, S., & Pandunata, P. (2019). Sentiment analysis of positive and negative YouTube comments using Na"{i}ve Bayes–support vector machine (NB-SVM) classifier. In *2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*. IEEE.

Neve, A., Pachpute, K., Mathapati, B., & Thorve, P. (2024). YouTube comment sentiment analysis. *International Journal of Creative Research Thoughts (IJCRT), 12*(6), June 2024.

Nawaz, S., Rizwan, M., & Rafiq, M. (2019). Recommendation of effectiveness of YouTube video contents by qualitative sentiment analysis of its comments and replies. *Pakistan Journal of Science, 71*(4 Suppl.), 91-97.

Pichad, S., Kamble, S., Kalamb, R., & Chavan, S. (2023). Analysing sentiments for YouTube comments using machine learning. *International Journal for Research in Applied Science and Engineering Technology (IJRASET), 11*(V), May 2023.

Pokharel, G., & Bhatta, D. (2021). Classifying YouTube comments based on sentiment and type of sentence. *arXiv preprint arXiv:2111.01908*.

Savigny, J., & Purwarianti, A. (2017). Emotion classification on YouTube comments using word embedding. In *2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*. IEEE.

Singh, R., & Tiwari, A. (2021). Sentiment analysis of YouTube comments. *International Journal of Scientific Research in Engineering and Management (IJSREM), 5*(5), 1-5.

Sungheetha, A., & Sharma, R. (2020). TransCapsule model for sentiment classification. *Journal of Artificial Intelligence, 2*(3), 163-169.

Timani, H., Shah, P., & Joshi, M. (2019). Predicting success of a movie from YouTube trailer comments using sentiment analysis. In *Proceedings of the 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom 2019)* (pp. 584-586).

Wei, Z., & Zhang, S. (2024). A structured sentiment analysis dataset based on public comments from various domains. *Data in Brief, 53*, 110232.