SENTIMENT PATTERNS IN YOUTUBE COMMENTS: A COMPREHENSIVE ANALYSIS

Savitha Shetty [1], Saritha Shetty [1], Priya Kamath B [2], Deepthi Shetty [3]

[1] Department of CSE, NITTE, Bangalore, India
[2] Department of CSE, Manipal Institute of Technology, Manipal, India
[3] Department of CSE, NMIT Bangalore, India

ABSTRACT

The sentiment analysis system is developed for YouTube comments, capable of accurately categorizing comments into positive, negative, or neutral sentiments. The system effectively handles challenges such as informal language, slang, and diverse topics, while being scalable and adaptable to different content genres. The objective is to provide a valuable tool for content creators, marketers, and researchers to gain insights into audience sentiments on YouTube, facilitating data driven decision-making and enhancing user engagement strategies.

KEYWORDS

Youtube comments, sentiment analysis, data processing

1. INTRODUCTION

Sentiment analysis is a method of identifying attitudes in text data about a subject of interest. It involves determining whether the sentiment expressed in a text is positive, negative, or neutral. When applied to YouTube comments, it can provide insights into the public's feelings about the content of the video[1]. VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool widely usedin natural language processing (NLP) tasks. Developed specifically for social media text, VADER assigns polarity scores to individual words in a text, considering their context and intensity, and aggregates these scores to compute the overall sentiment of the text. Its ability to handle emoticons, slang, and emphasis makes it particularly effective for analyzing sentiment in informal and unstructured text data, making it a popular choice for sentiment analysis tasks on platforms like Twitter and Reddit [2].

TextBlob on the other hand, is a simplified and beginner friendly NLP library in Python. It provides a straightforward API for various NLP tasks, including sentiment analysis, part-of-speech tagging, and noun phrase extraction. TextBlob's sentiment analysis module utilizes a pre-trained Naive Bayes classifier to classify text as positive, negative, or neutral based on the presence of certain words and their context. Its ease of use and simplicity make it an excellent choice for quick sentiment analysis tasks, enabling users to analyze the sentiment of text data with minimal setup and coding effort [3].

2. LITERATURE SURVEY

Authors combined the sentiment analysis with graph theory concepts to analyze the user posts and likes/dislikes on a different of social media to suggest recommendations for YouTube videos [4]. Authors utilized Naïve-Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), K Nearest Neighbor (KNN), and Random Forest (RF) techniques but SVM, LR, and RF performed very best with the highest accuracy scores. The system reads the data stored in the file having (Tab Separated Values) format. After reading, pre-processing phase is applied to clean and prepare the data for the use of machine learning algorithms. Directly text data cannot be given to machine learning algorithms, so it is converted into a suitable type. Based on bigram and tri-gram features the system achieved the result of micro-F 85% [5]. The authors proposed that the comments are collected from YouTube videos using YouTube Data API which are preprocessed to filter out the noise. The strategies like Random Forest, Decision Tree, K Nearest Neighbors are utilized and the Decision Tree shows lower performance. The number of classes and sub-classes can be increased in future to represent a more comprehensive comment classification
[6].

3. METHODOLOGY

The methodology included the following steps.

3.1. Data Collection

Utilized the YouTube Data API or web scraping techniques to collect a dataset of YouTube comments related to various videos across different genres. Ensure the dataset is diverse and representative, containing comments expressing a range of sentiments (positive, negative, neutral).

3.2. Data processing

Clean the collected comments by removing any irrelevant characters, URLs, or special symbols. Tokenize the comments into individual words or tokens, considering the context and structure of the text. Normalize the text by converting all characters to lowercase and removing stop words, punctuation, and numbers to focus on meaningful content.

3.3. Sentiment Analysis with VADER

Implemented the VADER sentiment analysis tool, leveraging its pre-trained lexicon and rule-based approach to assign polarity scores to individual words in the preprocessed comments. Aggregate the polarity scores to compute the overall sentiment of each comment as either positive, negative, or neutral. Evaluate the performance of the VADER model using metrics such as accuracy, precision, recall, and F1-score on a validation dataset to assess its effectiveness in sentiment classification.

3.4. Sentiment Analysis with TextBlob

Utilized the TextBlob library to perform sentiment analysis on the preprocessed comments. Employ TextBlob's pretrained Naive Bayes classifier to classify each comment as positive, negative, or neutral based on the presence of certain words and their context. Compare the results obtained from TextBlob with those from VADER to identify any discrepancies or areas of improvement.

3.5. Graphical Model Evaluation

Used a pie chart to represent the percentage distribution of each sentiment category (positive, negative, neutral) among the analysed comments. Used a bar chart to compare the performance metrics of VADER and TextBlob. Calculate the number or percentage of comments falling into each sentiment category using the results obtained from VADER and TextBlob.

4. IMPLEMENTATION

As shown in the figure 1 implementing a sentiment analysis model using VADER and TextBlob for YouTube comments involves several steps. First, the comments data from YouTube must be collected and preprocessed to remove noise, such as irrelevant characters or HTML tags and stored into an '.csv' file. At a time any number of comments can be chosen for the model to work on. Next, the preprocessed data is fed into both the VADER and TextBlob models for sentiment analysis. VADER's strength lies in its ability to handle informal language, slang, and emoticons, making it suitable for parsing the colloquial language often found in YouTube comments. However, it's important to note VADER's limitations in accurately interpreting nuanced or sarcastic comments due to its rule-based nature. TextBlob, leveraging its pre-trained Naive Bayes classifier, offers a straightforward approach to sentiment analysis, albeit with less efficacy in handling slang and informal language compared to VADER. For each model it checks with existing data set and identifies a comment as positive, negative or neutral. TextBlob gives the overall subjectivity and polarity. After sentiment analysis, the results from both models can be compared and analyzed to gain insights into audience sentiments and preferences on the platform. Additionally, fine-tuning the models based on the specific characteristics of YouTube comments can enhance the accuracy of sentiment classification.

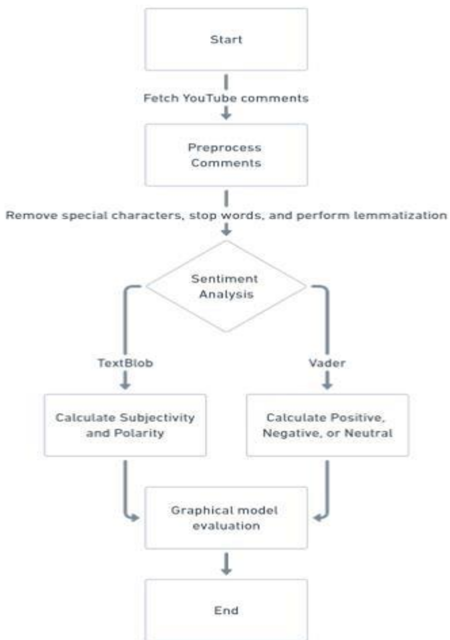

Figure 1. Sentiment analysis model using TextBlob and Vader

5. RESULTS AND DISCUSSION

The sentiment analysis of YouTube comments using VADER and TextBlob offers a comprehensive approach to understanding audience sentiments and preferences on the platform. This project began with the collection of comments from various YouTube videos across different genres, followed by meticulous data preprocessing to clean, tokenize, and filter the comments to focus on meaningful content. Utilizing the VADER and TextBlob sentiment analysis tools, the project successfully classified the preprocessed comments into positive, negative, or neutral sentiments, providing valuable insights into the emotional tone and attitudes of the audience. The followingfigure 2 indicates the output for the following sentence “Sorry not sorry I’m not taking the chance on my 6 year old. He has all his major vaccines but I’m not comfortable until more information is given on this covid vaccine for young children.”

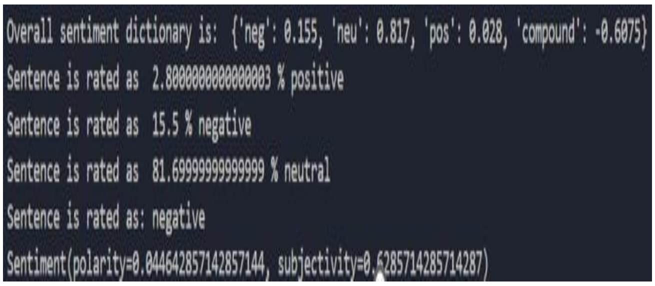

Figure 3. sample output1

The figure 4 indicates the output for the following sentence

“I read about these vary rare cases of myocarditis being reported in Israel about a month ago. Way ahead of you, C DC. 90”.

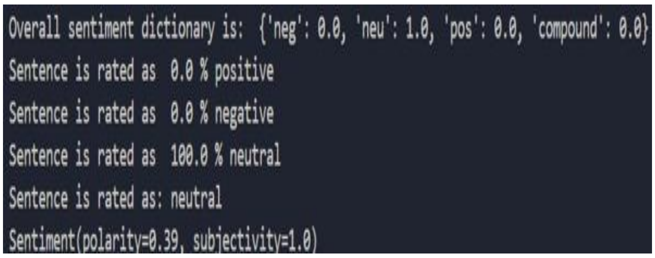

Figure 4. sample output1

Visual representations in the form of pie and bar charts were employed to illustrate the sentiment distribution across the comments, compare the performance metrics of VADER and TextBlob, and visualize the sentiment predictions of both models. The graphical insights revealed the prevalence of different sentiment categories among the comments and highlighted the strengths and weaknesses of each sentiment analysis tool in handling the diverse and informal language commonly found in YouTube comments.

Figure 5 indicates the pie chart to represent the percentage distribution of each sentiment category (positive, negative, neutral) among the analysed comments. Figure 6 indicates the bar chart to compare the performance metrics of VADER and TextBlob.

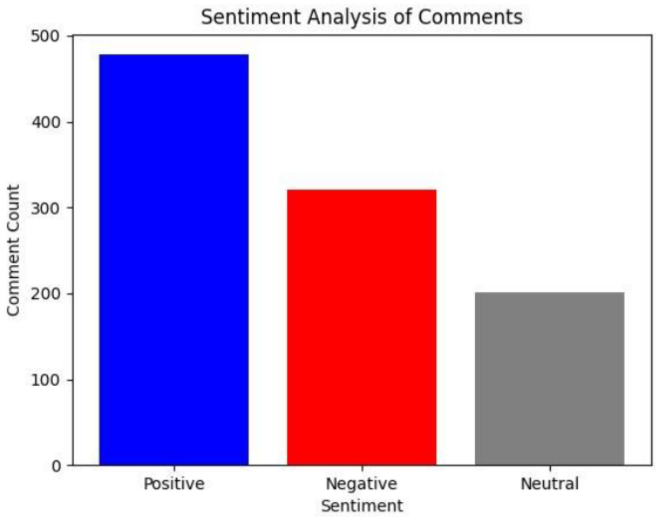

Figure 5. Sentiment analysis of YouTube comments

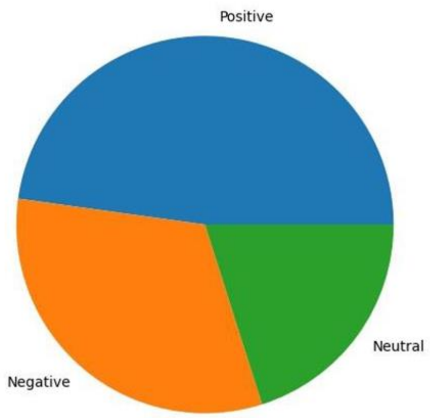

Figure 6. Pie chart representing the Positive, negative and neutral comments

6. CONCLUSION

In conclusion, this project serves as a robust tool for content creators, marketers, and researchers to gain actionable insights into audience feedback, enhance user engagement strategies, and make informed decisions to improve content quality and viewer satisfaction on YouTube. The combination of VADER and TextBlob provides a balanced approach to sentiment analysis, leveraging the rule-based and machine learning-based techniques to capture the nuanced emotions and sentiments expressed by the YouTube community. Further refinements and optimizations can be explored to enhance the accuracy, usability, and scalability of the sentiment analysis system, catering to the evolving needs and preferences of the diverse YouTube audience.

REFERENCES

[1] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), 1093-1113.
[2] Elbagir, S., & Yang, J. (2019, March). Twitter sentiment analysis using natural language toolkit and VADER sentiment. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 122, No. 16).
[3] Biswas, S., Sarkar, I., Das, P., Bose, R., & Roy, S. (2020). Examining the effects of pandemics on stock market trends through sentiment analysis. J Xidian Univ, 14(6), 1163-76.
[4] McGarry, K. (2023). Analyzing Social Media Data using Sentiment Mining and Bigram Analysis for the Recommendation of YouTube Videos. Information, 14(7), 408.
[5] Kavitha, K. M., Shetty, A., Abreo, B., D'Souza, A., & Konkana, A. (2020). Analysis and classification of user comments on YouTube videos. Procedia Computer Science, 177, 593-598.
[6] Pokharel, R., & Bhatta, D. (2021). Classifying YouTube comments based on sentiment and type of sentence. arXiv preprint arXiv:2111.01908.