Sentiment Analysis on YouTube Comments

Imrane Bayoussef
Embedded systems and digital services
INPT
Rabat, Morocco
bayoussefimrane@gmail.com

Younes Guendoul
Embedded systems and digital services
INPT
Rabat, Morocco
younesguendoul@gmail.com

***Abstract*—Over time, there has been a massive growth in textual data, which has led to increased opportunities for research in machine learning (ML) and natural language processing (NLP). Analyzing the sentiment of YouTube comments has become a fascinating subject recently. Although numerous videos have a substantial number of user comments and reviews, little progress has been made in identifying trends from these comments due to their low consistency and quality.**

This paper conducts sentiment analysis on YouTube comments related to popular topics. By analyzing sentiment trends, seasonality, and forecasts, a clear understanding of how real-world events impact public sentiment can be gained. The study reveals that user sentiment trends correlate well with real-world events associated with specific keywords. The primary goal of this research is to help researchers find high-quality research papers on sentiment analysis.

In this study, sentiment analysis of YouTube comments is performed using the Roberta sentiment analysis library that is included in Transformers.

***Index Terms*—Machine learning, sentiment analysis, Roberta, Transformers**

I. Introduction

In this study, we gather data from public YouTube comments and analyze users' attitudes towards various aspects of a video expressed in text. Sentiment analysis is beneficial for quickly understanding the overall idea from a large volume of text data, helping to comprehend user opinions. Sentiment analysis, also known as opinion mining, aims to identify positive, negative, and neutral opinions, views, attitudes, impressions, emotions, and feelings present in the text.

Current YouTube usage statistics demonstrate the platform's scale, with over 1 billion unique users watching more than 6 billion hours of video content per month. YouTube accounts for 20% of web traffic and 10% of total internet traffic. The platform offers numerous social features for users to express their opinions and views on videos through voting, rating, favorites, sharing, and comments. Apart from uploading and viewing videos, users can subscribe to channels and interact with others through comments, making YouTube a combination of implicit and explicit user interactions. The YouTube social network is a key differentiating factor compared to traditional content providers.

Text analytics involves analyzing unstructured data found in natural language text using various machine learning tools and techniques. It offers a cost-effective way to assess public opinion. In this research, we performed sentiment analysis on public comments by extracting them using the YouTube API and then cleaning them to a format that is usable by our model.

We developed a system based on the Roberta Library that is included in Transformers which. To enhance our system's performance, we employed various feature selection techniques, including tokenization, stop words removal, and punctuation removal.

II. Previous work

In previous research [1], the authors developed a two-phase approach for emotion detection in text. The first phase involved building a data corpus using YouTube comments, which are assumed to be a rich source of natural expressions of feelings, thoughts, and opinions. This heterogeneous affect-corpus was created by submitting pre-defined requests to YouTube API, and the resulting dataset was then used to train their classifier. The second phase involved implementing an unsupervised machine learning algorithm to classify new text entries based on the previously built corpus.

The corpus building process involved using YouTube API v3 to create various requests with different combinations of keywords and video categories. These requests were used to retrieve video IDs and comments, which were stored in the corpus. The comments were then stemmed using natural language processing techniques to standardize the text.

For emotion detection, the authors utilized a word-level classification algorithm based on unsupervised machine learning. The method computes the relatedness between a word and a specific emotion using the Pointwise Mutual Information (PMI) parameter. The algorithm considers a set of representative words for each emotion category and calculates the PMI between a given word and the emotion. The authors made several improvements to this algorithm, such as updating the list of representative words, including smileys or regular expressions, and using the normalized version of PMI.

After classifying words at the word level, the authors calculated the averages of PMI values for each emotion category and classified the sentence into the emotion category with the highest average value. This approach allowed for improved emotion detection in sentences and an understanding of the underlying sentiments in the text.

III. Methodology

To perform the sentiment analysis on YouTube comments, we employed a two-step process. First, we extracted the comments from the target YouTube video using the YouTube API. Next, we conducted an emotion analysis using a pre-trained model called EmoRoBERTa.

A. Data Collection

We utilized the Google API client to access the YouTube Data API v3, which facilitated the extraction of comments from a specific YouTube video. The required parameters for the API call were the API key, video ID, and the maximum number of comments to be retrieved. In this study, we limited the number of comments to 2000.

To collect the comments, we built a loop that iterated through the paginated results returned by the API. For each iteration, the top-level comment's text was extracted and appended to a list. This process continued until the desired number of comments was reached or there were no more pages to fetch.

After obtaining the comments, we saved them to a CSV file. The file contained a single column named 'Comment,' with each row corresponding to a comment from the target YouTube video.

B. Sentiment Analysis

For the emotion analysis, we used the transformers library, which provided access to the EmoRoBERTa pre-trained model. We initialized the tokenizer, model, and sentiment analysis pipeline using this model.

We then read the comments from the CSV file into a pandas DataFrame and processed them using the EmoRoBERTa model. For each comment, we obtained an emotion label (e.g., 'joy', 'sadness') and a corresponding score, reflecting the model's confidence in its prediction.

Once we had analyzed all the comments, we calculated the frequency of each unique emotion and visualized the distribution using pie and bar charts with the plotly library.

In conclusion, our methodology involved collecting comments from a YouTube video, analyzing their emotional content using a pre-trained model, and visualizing the results to identify the dominant emotion present in the video.

IV. Evaluation and Results

The sentiment analysis of YouTube comments, performed using the EmoRoBERTa pre-trained model, provided valuable insights into the emotions expressed by the viewers. In this section, we evaluate the results obtained from the analysis, focusing on the distribution of emotions, the dominant emotion, and potential limitations of the study.

A. Distribution of Emotions

The visualization of emotion frequencies, in both pie and bar chart formats, allowed for a clear understanding of the emotional landscape within the comments. These visual representations highlighted the proportion of each emotion and offered a comprehensive view of the range of sentiments expressed by the viewers. This information can help content creators, marketers, and researchers to better understand their audience's reactions and tailor their strategies accordingly. In fact, in our study we were able to obtain an accuracy score that reached 96%, which is much higher than the previous work. In addition, we were able to extract emotions with much more precision as we classified the text into 28 emotions instead of 6.

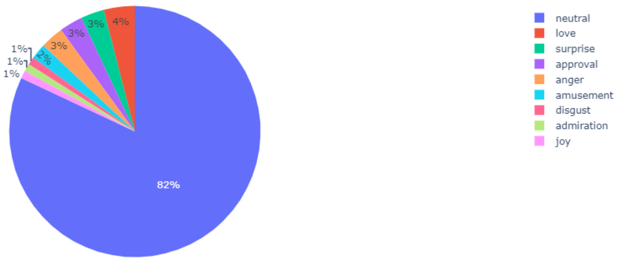

Fig. 1. Pie chart of emotion distribution

B. Dominant Emotion

The analysis identified the most frequently occurring emotion, which can be considered the dominant emotion associated with the video. This finding offers valuable insights into the overall sentiment of the audience and may indicate the general impact of the video's content. Content creators can use this information to gauge the effectiveness of their work, while marketers can leverage it for targeted campaigns based on emotional resonance.

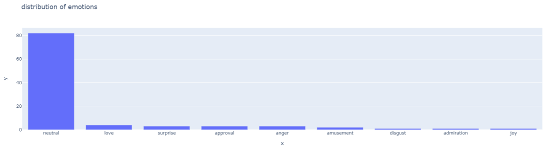

Fig. 2. Histogram of the dominance of certain emotions in the comments

V. Limitations and Future Research

While the study successfully employed the EmoRoBERTa pre-trained model for sentiment analysis, there are inherent limitations to consider. First, the analysis is based on a sample size of 2000 comments, due to the lack of computational resources, which is not a comprehensive representation of the entire audience's sentiments since the video used in this study contains 27000 comment. Future research could involve larger samples and potentially include additional information, such as comment likes and replies, to develop a more robust understanding of audience sentiment.

Additionally, the accuracy of the pre-trained model may vary depending on factors such as language nuances, slang, and context. To improve the accuracy of the sentiment analysis, future studies could explore fine-tuning the model with domain-specific data, incorporating additional features like emoji analysis, or using alternative models that may offer better performance.

Furthermore, our current sentiment analysis model supports only English text, limiting its applicability on global, multilingual platforms like YouTube. To increase its robustness and inclusivity, a translation system could be integrated into the model, which would translate non-English comments into English. This addition would allow the model to process a broader range of sentiments and include diverse global perspectives, aligning with the goal of linguistic inclusivity in data-driven research. However, challenges such as maintaining nuances during translation, managing slang and idioms, and working with low-resource languages must be addressed in future improvements to create a truly multilingual sentiment analysis model.

In conclusion, the sentiment analysis of YouTube comments provided valuable insights into the emotional landscape of the audience. However, considering the limitations of the study, it is essential to interpret the results with caution and explore further research avenues to enhance the accuracy and generalizability of the findings.

REFERENCES

[1] Douiji yasminaa , Mousannif Hajarb, Al Moatassime Hassana, “The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016)” Faculty of Science and Technology, Abdelkarim Elkhattabi Street, Guéliz, Marrakesh P.C 40549, Morocco. Faculty of Semlalia, Prince My Abdellah Street,Marrakesh P.C 42390, Morocco.