Design of AI Systems Project

Lee Yu Xuan
Chalmers University of Technology
yuxua@chalmers.se

Zhang Xiaoyang
Chalmers University of Technology
xiazhan@chalmers.se

Abstract

In this project, we have created a system that is designed to collect and analyze comments from Youtube videos through sentiment analysis techniques. With the surge of user-generated content on social media platforms such as Youtube, it is now harder to manually analyze comments for large-scale studies. The system that we have created allows the user to automatically collect the comments from a Youtube video and does sentiment analysis on the comments to get sentiments on the video. This would allow researchers and content creators to gain valuable insights.

1 Introduction

Social media platforms have changed the way individuals interact, share contents and express their opinions online. As such, we have seen many individuals coming onto these social media platforms to share their experiences, express their opinions and interact with their peers and other individuals. Within this dynamic landscape, influencers leveraging social media platforms to cultivate a large, engaged audience have changed the digital marketing and content creation paradigm.

Currently influencers have such a big influence over their audiences preferences, brand perceptions and cultural trending. They have the ability to shape discourse and drive engagement across many different niches and demographics. As such, understanding their audience's feedback, sentiments and interactions within the social media platforms is very crucial for influencers, marketers and researchers alike.

Analyzing user-generated content and interactions on social media platforms can provide valuable insights such as audience's perceptions, preferences and behaviors. This would provide influencers, marketers and researchers a better understanding of the audience's sentiment and engagement. However due to the sheer volume of the content that is generated across social media platforms and an even bigger amount of comments that are generated from these contents poses a significant challenge for manual analysis. As such, there needs to exist a system that is able to extract insights from this set of data automatically and present it to the user in a manner that is easy to understand.

In this report, we propose a system that is capable of extracting all of the comments from a Youtude video and performing sentiment analysis to gain an understanding on whether the audience feels positively or negatively towards the video and determine what the audience is mostly talking about in the comment section.

2 Methodology

Our approach consists of two key steps: finding the best pre-trained model for sentiment analysis and figuring out what the comment section is talking about.

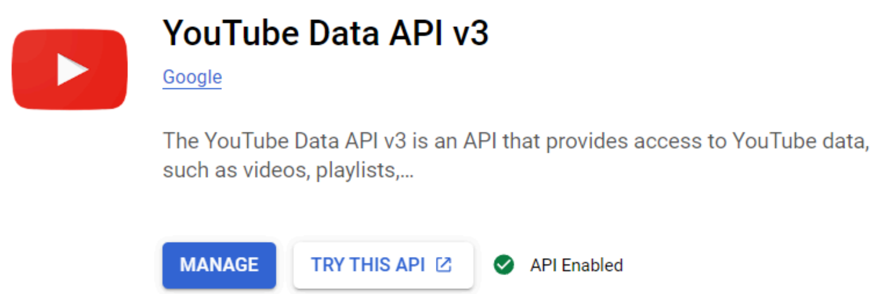
2.1 Finding the best pre-trained model

Due to the restriction of resources, we will be fine-tuning a pre-trained model that is made for doing sentiment analysis on text. However we have a choice in which pre-trained model we want to use. In order to select the most accurate pre-trained model, we selected 2 models and ran it through an annotated dataset to compare their accuracy and performance. The dataset only has 3 annotations : Positive, Negative and Neutral. The dataset that we are using is the ‘Twitter Sentiment Analysis’ from Kaggle containing over 700 tweets.

We select the model that is able to accurately do the sentiment analysis on the tweets.

2.2 Collecting Youtube video comments

Instead of web scraping Youtube to get the comments from a particular video, there is an API call already created, making it much easier to collect all of the comments from a video called “Youtube Data API v3”.



2.3 Figuring out what the comment section is talking about

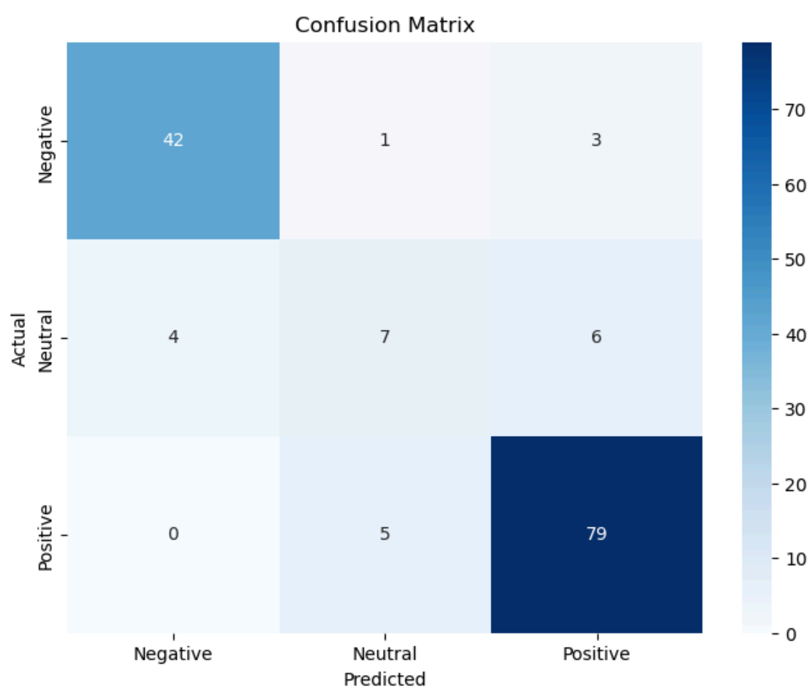
In the system, we have decided to collate the comments and count the number of occurrences of each unique word that appeared in the comment section. The reason why we do this is because something in the video has caught the audience's attention, they are likely to talk about it in the video. As such, by getting the count of each unique word from the comment section, we know what really stood out from the video. We then represent the count of each unique word in a form of a word cloud to visualize which words occur the most in the comment sections.

In addition to that, we choose to focus only on the top 10 words with the highest appearance count. When we are collating the comments, we also choose to remove stopwords, or common words that frequently appear in the English language such as 'the' or 'is' as they will affect the frequency count of unique words appearing and will always be in the top 10.

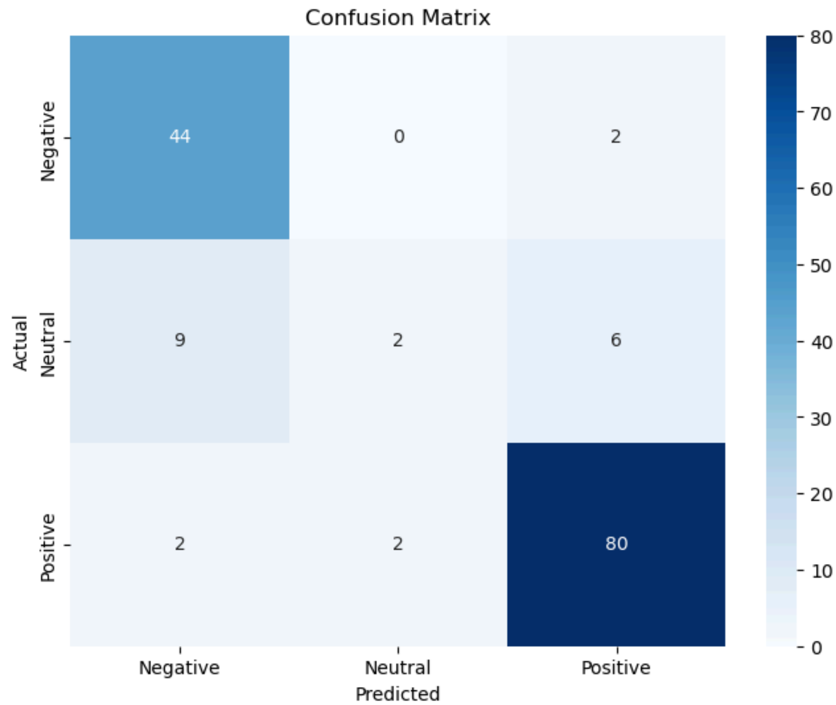
3 Results and Discussion

3.1 Model Performance

Model Number	Accuracy	Precision	Recall	F1 Score
1. BERT	87.07%	0.86	0.87	0.86
2. RoBERTa	85.71%	0.83	0.86	0.83



Confusion Matrix for BERT



Confusion Matrix for RoBERTa

In this study, we evaluated two pre-trained models, BERT and RoBERTa, to analyze the sentiment of the Twitter comments. The models were fine-tuned to this dataset and their performance was assessed based on accuracy, precision, recall, and F1 Score as shown above. While BERT showcased marginally superior overall performance, RoBERTa outperformed BERT in accurately predicting positive and negative sentiments, which are crucial for understanding audience engagement and content reception.

Given the importance of distinguishing between positive and negative sentiments for practical applications, RoBERTa was selected for further analysis of YouTube comments to leverage its strengths in these areas.

3.2 Discussion

Applying the RoBERTa model to analyze sentiments in Youtube comments allowed for a detailed examination of viewer feedback on selected video content. The sentiment analysis revealed a distribution of sentiments that offered insights beyond basic like/dislike metrics, highlighting areas of viewer satisfaction as well as areas of improvement.

Reason for inaccurate prediction of neutral sentiments could be due to the imbalanced dataset. Since our dataset has significantly fewer examples of neutral sentiments compared to positive and negative ones, the model may not learn to recognize neutral sentiments effectively. Moreover, neutral sentiments tend to be more ambiguous, making it hard to predict.

3.3 Results

A youtube video that we utilized that was well liked by viewers was PewDiePie's "How a Reddit Post Changed my Life" Video, whereby he talks about his bouldering journey. Expectedly, the Sentiment Analysis Report we generated highlighted 73.5% Positive Sentiments, 26.3% Negative Sentiments, and 0.2% Neutral Sentiments.

On the flip side, we also ran the model on a not so well like Youtube Video of Logan Paul finding a Dead Body in Japan which resulted in 23.95% Positive Sentiments, 75.94% Negative Sentiments, and 0.11% Neutral Sentiments.

Such sentiment analysis can benefit content creators as they can use the findings to enhance viewer satisfaction by focusing on elements that receive positive feedback, address and mitigate aspects that lead to negative sentiment, overall engaging the audience and fostering a community driven content creation process.

Word Cloud for PewDiePie's video

Top 10 most common words: [('climbing', 201), ('magnus', 96), ('like', 82), ('bouldering', 81), ('pewds', 79), ('love', 77), ('v', 76), ('video', 73), ('climb', 67), ('really', 61)]

In addition, we also incorporated a word cloud and top 10 most frequently used words. Prominent terms such as "Climbing", "bouldering" and "love" not only underscores the context of the video but also positive reception of the viewers. The use of visual aids can quickly grasp the predominant themes that resonate with the audience.

3.4 Limitations and Future Directions

Due to the limitations of our hardware, we were unable to further fine-tune the pre-trained model to make its sentiment analysis even more accurate. If we were given more resources and time, we could have further fine-tuned the model. Another limitation we had was the fact that our API for collecting the Youtube video comment was limited everyday as we did not pay for the premium service, as such we could only do analysis for videos that do not exceed over a 100,000 comments. In the future, we could further expand this sentiment analysis on social media platforms by targeting other social medical platforms such as Facebook, Twitters and even Instagram. This system could be further expanded on by looking at news articles.

Selecting the top 10 words from the unique word count might not give us the general idea of the audience's view on the particular video. As such this could be further expanded by integrating a model that is able to summarize text. There was an attempt to integrate this feature into the system, however given the time frame, we were unable to fix any bugs and integrate it into the system.

4 Conclusion

In conclusion, creating a system that is able to get the sentiments of any social media post is essential as the ways of digital marketing have changed so much that influencers and companies need a more efficient method to determine whether their marketing campaigns are efficient. By leveraging current technologies and automation, we are able to convert sentiment analysis based on comments from a manual job into an automated job. This allows for much quicker reaction and valuable insights into the audiences sentiments and engagement across any social media platforms,

Bibliography

“nlptown/bert-base-multilingual-uncased-sentiment · Hugging Face.” *Hugging Face*, https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment. Accessed 12 March 2024.

“YouTube Data API v3.” *YouTube: Home*, 3 June 2009, https://console.cloud.google.com/marketplace/product/google/youtube.googleapis.com?q=search&referrer=search&project=graphical-bus-361912. Accessed 12 March 2024.