YouTube Comment Analysis Using Lexicon Based Techniques

Mohan Sai Dinesh Boddapati, Madhavi Sai Chatradi, Sridevi Bonthu(✉), and Abhinav Dayal

Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India
{20PA1A0521,20PA1A0526,sridevi.b,abhinav.dayal}@vishnu.edu.in

Abstract. YouTube is used to watch music videos, comedy shows, how-to guides, recipes, hacks, and more. As of February 2020, more than 500 h of video were uploaded to YouTube every minute. This equates to approximately 30,000 h of newly uploaded content per hour. The amount of content on YouTube has increased dramatically as consumers' appetites for online video have grown. Indeed, the number of video hours uploaded every 60 s increased by roughly 40% between 2014 and 2020. The direct means of user review for this content is the comment section. To survive this cut-through competition, the content creators should constantly check up on their viewers' opinions, their reviews, and their sentiments toward the video. Although comments provide a direct means of feedback, the YouTuber cannot actually read all those comments. There may be times when he wants to know the drawbacks of the video. This paper proposes a dashboard that assists the content creators in actually looking at the positivity and negativity they have gained through the video. We opted for lexicon-based techniques over traditional classification to classify the comments into various categories. This project is a boon to the content creator's ability to increase his viewership.

Keywords: Summarization · Comment classification · Natural language processing · Polarity · YouTube comments

1 Introduction

1.1 Content Creation

With 2 billion monthly active users, YouTube is the second most popular social media platform based on total visitors and page views. It has gained huge popularity among content creators. A large number of content creators upload their video content on this platform. These videos get tonnes of views and comments. Content creators need to continuously work on maintaining the quality and quantity of their content. To do so, they must collect feedback from their viewers. This feedback lets them understand the influence of their creations. In addition to improving audience engagement, feedback also provides information on the aspects of the content that need improvement. The Youtube comment section is a direct means of feedback from the users.

N. Gupta et al. (Eds.): IC4S 2022, LNICST 472, pp. 76–85, 2023.
https://doi.org/10.1007/978-3-031-28975-0_7

1.2 YouTube Revolution

YouTube comments are an opportunity for the site's 1.7 billion unique monthly visitors to share what they love, hate, or simply must troll. Comments can also be a powerful opportunity for positive community building and, at the same time, can be a place for negativity.

It is part of many people's social media strategy to convince the audience that they want to make the most of their presence. Managing comments effectively (with moderation, replies, and analysis) is critical. The comment section booms within seconds for most of the popular channels, and it is a difficult job for the content creator to manually read all those comments. We created a dashboard where the content creators get a glance at the positivity and negativity of their content.

1.3 Contributions

The main contributions of this work are

- The classification of raw comments into five categories, viz., highly-positive, positive, neutral, negative, and highly-negative comments, and later summarises them as positive, negative, and neutral comments. This approach helps the content creators to increase their viewership and analyse the opinions of the content viewers.
- Analyze the video based on computing rank by considering the intensity of positive and negativeness in those raw comments. So that it helps the viewers decide to watch the video or not based on the rating computed.
- Creation of a user-friendly dashboard that helps YouTubers to know about information related to that video, classified comments, summarization of comments, and the rank of that video.

YouTube videos could be a powerful medium to spread information because the audience can access them easily. Also, it could describe the public opinion that could affect national unity. YouTube helps in various sectors like education(E-learning), entertainment, etc.

2 Related Work

There has been much sentiment analysis on YouTube. These works involve collecting and analysing YouTube comments from various areas in order to gain significant and interesting insights.Currently, YouTube has over 122 million active users on a daily basis. 1 billion hours of content are watched across the world every day 62.

There are many ways to classify raw comments: machine learning models and NLP-based algorithms are used for sentiment analysis.

2.1 Machine Learning Models

Many researchers build machine learning models for the classification of raw comments by using supervised and unsupervised techniques. Singh et al. built a classification model for classifying raw comments by using SVM, Naive Bayes, and KNN algorithms [1]. They analysed their model with various algorithms for better classification of comments and for good accuracy. Hajar et al. [2] built a text-based emotion detection system, based on an unsupervised machine learning algorithm by using YouTube comments as a data corpus. They achieved an average precision of 92.75%, and a 68.82% average accuracy by using SVM as a machine learning algorithm. Wadhwani et al. [3] built a machine learning model for sentiment analysis of YouTube comments. They achieved an accuracy of 81% but proposed an accuracy of 89.3% by using performance metrics.

2.2 NLP Based Algorithms

Nowadays, NLP is a growing technology, frequently used for sentiment analysis, language modeling, speech recognition, etc. Many researchers are using NLP for sentiment analysis. Asghar et al. [4] classified raw comments based on polarity-based techniques. They used several categories of videos, like news, comedy, shows, etc. They evaluated their model by using performance metrics and used SVM for classification. Olga Uryupina et al. [5] classified comments based on a corpus containing labels. It is an annotation project for text categorization and targeted opinion mining of user-generated comments on YouTube videos. Potthast et al. [6] worked on summarizing and visualizing the opinions expressed in the form of Web comments. Poche et al. [7] worked on Analyzing user comments on YouTube coding tutorial videos and summarizing those comments by using a dataset of 6000 comments and building an SVM model with an average accuracy of 77%. Kalra et al. [8] worked on YouTube comments by scraping the YouTube comments using Selenium, requests, and Beautiful Soup and using various evaluation metrics for Random Forest Classifiers. We analyzed 300 videos collected from various sectors to classify comments, video ratings, and topic-based recommended videos. Many researchers show various ways to analyze the comment classification of raw comments by building prediction models on top of them. We used NLP-based techniques to build a model.

3 Data Collection

The required data is collected by using the following two techniques.

- YouTube API and
- Webscraping

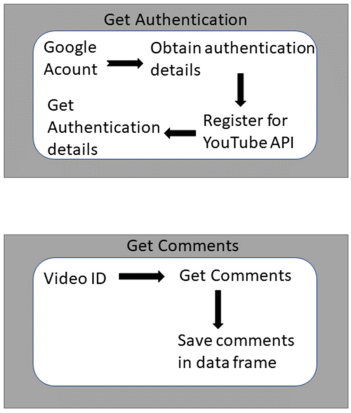

Fig. 1. Working of YouTube API

3.1 YouTube API

Extracting information from the web is pretty cool, and it becomes more professional when you use APIs provided by an application. YouTube is the most popular video-sharing platform, and it is owned by the world's biggest tech giant, Google. So, extracting comments would be easier. There are several ways to extract comments from YouTube. Firstly, set up the YouTube API and get the credentials to obtain the API key. With that key, we can extract comments by using YouTube ID. The process of working of YouTube API is clearly described in Fig. 1. The major drawback is that we can only extract 100 comments by using this technique.

3.2 Web Scrapping

3.3 YouTube API

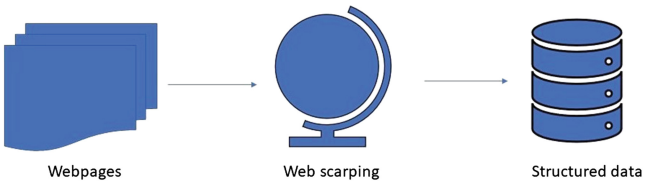

Fig. 2. Outline of web scrapping procedure

To overcome the problem faced during the extraction of comments using the YouTube API, we used the scraping technique. As shown in Fig. 2, web scrapping algorithms take web pages as input and produces structured output. Gen-

erally, the Python Selenium[1] and BeautifulSoup[2] libraries are used to scrape the comments manually by scrolling down the pages. Although extracting data from YouTube can be done with the help of web scraping, it is slow, and most of the big web applications do not promote web scraping.

4 Methodology

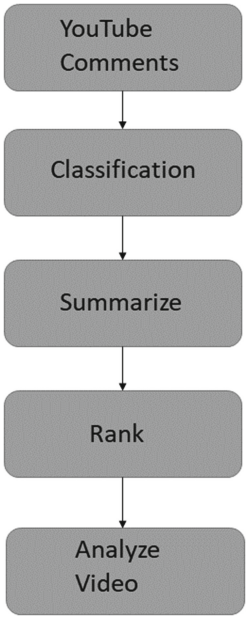

Fig. 3. Architecture of the proposed work.

The architecture of the proposed work is shown in Fig. 3.

4.1 Extraction of Comments

We extracted raw comments by using the YouTube API and scraping techniques. Fig. 1 focuses on the extraction of raw comments from the YouTube API by authorising the account and registering for the YouTube key to get those raw comments. The main drawback of this technique is that we can extract only 100 raw comments based on either relevance or the newest comments.

To overcome that problem, we used the scraping technique by using Selenium and BeautifulSoup libraries in Python so that the comments are extracted manually and stored in a data frame. The drawback of this technique is that it takes more time to scrap all those raw comments.

[1] https://www.selenium.dev/.
[2] https://pypi.org/project/beautifulsoup4/.

4.2 Classification

Those raw comments are further classified into five categories. Let the training data $D = \{X, y\}$ where X is a set of raw comments, y is a class label $\in$ {Highly-Positive, Positive, Neutral, Negative, and Highly-Negative} for each raw comment, x_i in X that maps to the corresponding class label in y. A sample set of training examples are shown in the Table 1.

Table 1. Sample Training examples after scrapping.

Comment	Label
Mam, your presentation is too excellently good and your smile is too really good mam .....!!!!	Highly positive
It's a good and nice video to watch	Positive
Ma'am please make a fast video I req please daily Make 3 videos per day	Neutral
Mind blowing eyes	Positive
Ur are saying something wrong in java course	Negative
Tq mam for your kind-hearted patience and we love you so much for your excellent teaching skills your way to speaking helps a lot in campus placements	Highly positive
You are looking nice mam and your speaking is good	Positive
Whatever your teaching is not understood by me and your way of speaking is too worst and don't make any videos right now ?	Highly negative
It's really a wrong question at 20:03 s don't say wrong to students	Negative

There are different approaches to classifying the sentiment of the text. We used lexicon-based modelling in NLP for the classification of raw comments. With lexicon-based sentiment analysis, words in texts are labelled as positive or negative (and sometimes as neutral) with the help of a so-called valence dictionary. Consider the phrase, "Good people sometimes have bad days." The word "good" is labelled as positive, the word "bad" as negative, and possibly the other words as neutral. The TextBlob[3] library in Python helps to compute the polarity of a sentence or word that determines the sentiment of that sentence or word and helps in the classification of raw comments. The polarity ranges from [−1, 1], where −1 indicates negative polarity, 1 indicates positive polarity, and 0 indicates neutral polarity. The polarity values for a given set of sample sentences is tabulated in Table 2. Based on a range of polarity, we classified those raw comments.

[3] https://textblob.readthedocs.io/en/dev/.

Table 2. Polarity values of some sentences.

Sentence	Polarity
I love Monday but I hate Monday	−0.15
Miss Universe is beautiful	0.85
I love California	0.5
The covid pandemic is really terrific	−0.9
I'm travelling	0.0
We are having fun there	0.3
We hate you	−0.8

4.3 Summarization

Let X be a set of classified comments *for eg. positive or negative or neutral).* This work used Sumy library of python for doing summarization. Now X can be summarized into W where $(W << X)$ so that it helps the YouTubers to look at the overview of comments.

4.4 Dashboard

Creating a user-friendly dashboard helps both content creators and users. For content creators, it helps to look at the view counts, classified comments, summarization of classified comments, and analyze the video. For users it helps to watch a video and analyze the video to watch or not based on rank. We have created the dashboard by using Python Streamlit[4] library.

4.5 Extensions

It's difficult to analyze the video that provides users with the best content based on the YouTube recommendation system, to overcome this problem we computed the rank for the top 5 recommended videos by YouTube so that it helps users to watch the best video based on their search as shown in Fig. 4.

4.6 Rank

To analyse the video's best or not views and likes count doesn't play a major role because for a channel having more number of subscribers will easily get more views and likes therefore based on intensity of comments it's easy to analyse video. So to analyse video by computing rank determines the video is best or not. Rank Algorithm helps to compute rank by considering the emotional intensity of positive and negative comments along with their count.Based on rank value a star rating is given to each video that helps for YouTube users (Figs. 5, 6, 7, 8 and 9).

[4] https://streamlit.io/.

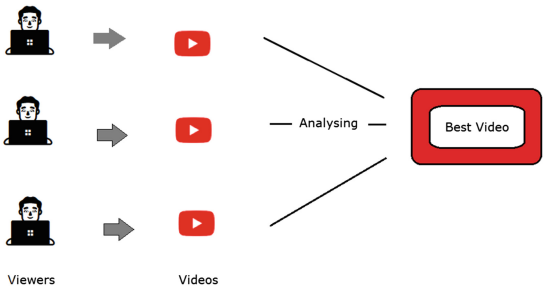

Fig. 4. Recommending the best video

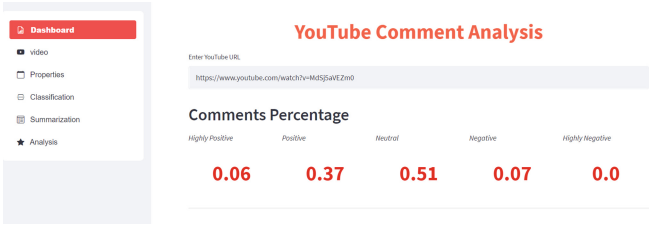

Fig. 5. Home screen of the dashboard

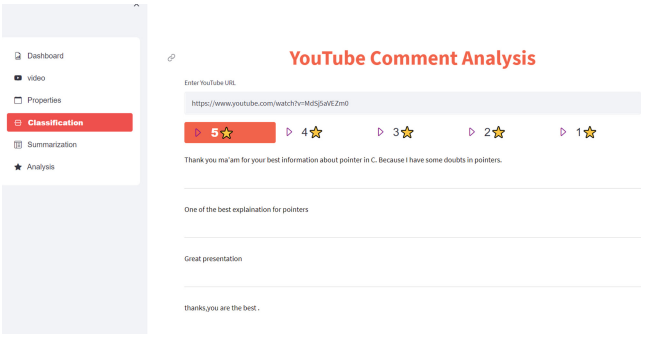

Fig. 6. Sentiment Classification of the video

Fig. 7. Summary of the video

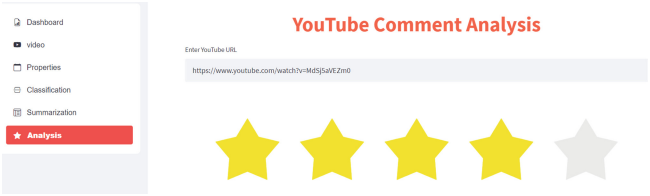

Fig. 8. Rank analysis

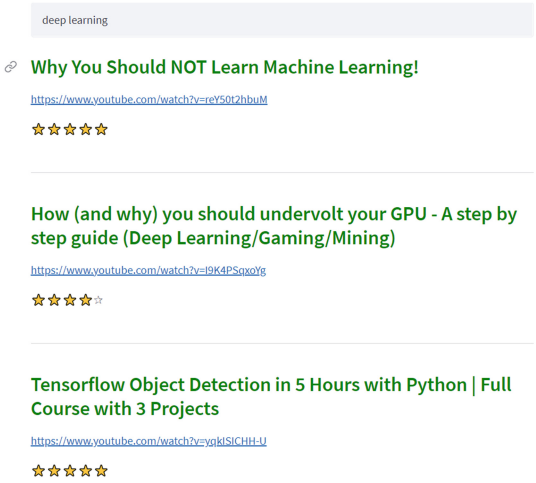

Fig. 9. Prediction of the best videos

5 Results

This section presents the dashboard snapshots.

The homescreen of the dashboard is shown in the figure. It has provision to enter the URL of the respective video. The left panel of the dashboard contains the menu leading to classification, summarization, and rank analysis. The dashboard displays the probability of all five classes for the comments of the entered YouTube video. The classification tab is displayed in the figure. It displays the count of comments belonging to every category. The figure is the summarization window, which displays the summary of the comments in positive, negative, and neutral sections. The assigned rank can be viewed in the Analysis tab as shown in the figure. Finally, the figure shows the application page, which recommends the best videos.

6 Conclusion

With the dashboard, the content creator will know the sentiment of the viewer at all times. They can have a gist of what's working and what's not working.

For instance, the dashboard could show them the positivity and the negativity of a newly released video. A brief summary of positivity and negativity increases their productivity compared to manually filtering their comments. This would save them a lot of time. There have been many classification models for sentiment analysis, and most of them were built on traditional classification techniques. We used lexicon-based techniques, which use a valence dictionary to predict the positivity of a given word and fine tune on threshold values to categorize the comments into some categories.

The comments on YouTube are a massive amount of unstructured data. Most of the existing machine learning models and algorithms can produce meaningful insights when provided with structured data. The categorization of raw comments into the above-mentioned classes adds structure to the comments, Our work can also be used to annotate these comments for building a supervised learning model or the categorized comments can be fed into some existing algorithms. In addition to these, we assigned a rank to the video based on some polarity metrics, and this can be used as a filtered search to extract the video that provides the best content.

References

1. Singh, R., Tiwari, A.: Youtube comments sentiment analysis. Int. J. Sci. Res. Eng. Manage. **5**(5), 1–11 (2021)
2. Mousannif, H.: Using YouTube comments for text-based emotion recognition. Procedia Comput. Sci. **83**, 292–299 (2016)
3. Wadhwani, S., Richhariya, P., Soni, A.: Analysis & Implementation of Sentiment Analysis of User YouTube Comments. No. 7703. EasyChair (2022)
4. Asghar, MZ., et al.: Sentiment analysis on YouTube: a brief survey. arXiv preprint arXiv:1511.09142(2015)
5. Uryupina, O., et al.: SenTube: A corpus for sentiment analysis on YouTube social media. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014) (2014)
6. Potthast, M., Becker, S.: Opinion summarization of web comments. In: Gurrin, C., et al. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 668–669. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12275-0_73
7. Poché, E., et al.: Analyzing user comments on YouTube coding tutorial videos. In: 2017 IEEE/ACM 25th International Conference on Program Comprehension (ICPC). IEEE (2017)
8. Kalra, G.S., Kathuria, R.S., Kumar, A.: YouTube video classification based on title and description text. In: 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). IEEE (2019)