Sentimental Analysis of YouTube Videos

Aditya Baravkar[1], Rishabh Jaiswal[1], Jayesh Chhoriya[1], Prof. Bhanu Tekwani[2]

[1]Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, Maharashtra, India

[2]Professor Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, Maharashtra, India

ABSTRACT

YouTube is the second most popular social media platform with two billion users. Every minute around thousand hours of videos are uploaded. On an average people watch one billion hours of YouTube videos per day. Every youtuber want their video to get popular and make best possible efforts. However, video can crawl at the top of search with help of clickbait, etc. which compromises the content of video. There are videos whose relevancy and quality are top-notch but cannot make to top five or ten. People watching such videos interact by commenting, liking and subscribing. Especially in education category, viewers interested in watching long marathons or tutorial series have to make choice wisely to avoid wastage of time. To get favorable videos on top list, the sentiments of comments, no. of likes, views, comments is considered. The objective is to provide well analyzed and relevant educational videos to the budding students by reducing valuable search time.

Keywords: YouTube, Sentiments, Comments

I. INTRODUCTION

Ever increasing population has created a competitive environment among youth generation. YouTube enables inexpensive distribution of educational content, including course materials from educational institutions. Young minds prefer available free content on YouTube than spending on coaching institutes.

Educational tutorial series or marathons available on YouTube may be preferred by one student and no other. It entirely depends on knowledge possessed previously. The viewers of that particular series or marathon can help students to get idea about the quality and relevancy of content since the videos had been watched by beginners, intermediate and professional people too. The nature of comments posted by viewers, number of likes, views help in judging the videos.

The project considers sentiments of comments, number of comments, number of views and number of likes to perform customized sorting on the top videos provided by YouTube according to its ranking. We have used YouTube API to fetch data related to specific videos and extracted parameters like comments, number of comments, likes and views. A

machine learning model is created which is trained using mobile product reviews obtained from Amazon as it had over 3 lakh entries. The model works on logistic regression.

The idea in this project is mainly motivated by the existing algorithm of YouTube to show search results related to educational content. YouTube has great features of adding comments to videos. This can help to creator to engage audience as well. This feature can be used to great extent and they can modify their working of algorithm. Their main aim is to engage audience, so they try to show those content which are based on their search history or type of videos people are watching. It tracks the behaviour while you are searching.

II. LITERATURE REVIEW

Osaimi and Badruddin have done immense work on the sentiment analysis of the tweets in the Arabic language. In this, they built different classifiers and trained them. They used a proper dataset and analysed the accuracy and result of these classifiers to predict the correct sentiments [1].

Pragya Tripathi, Santosh Kr Vishwakarma, Ajay Lala have submitted their work on the sentiment analysis of English tweets with the help of rapid minor. They collected the dataset from the twitter in natural language and applied the techniques of text mining and used to build the sentiment classifier [2].

K.Bhuvaneswari and R. Parimala have proposed in their work, a method for sentiment classification using correlation-based feature selection. They applied different data pre-processing techniques, and then used a correlation attribute method for feature selection, and then finally two classifiers namely Naïve Bayes and Support Vector Machine are implemented and results were evaluated [8].

Farhan Laeeq, Md. Tabrez Nafis and Mirza Rahil Beg have proposed a work on sentiment classification of social media. In their work, they used three classifiers namely K-NN, Naïve Bayes and Decision Tree Classifier for sentiment classification and obtained a result that shows the accuracy of K-NN, Naïve Bayes, and Decision Tree classifier is 76.50%, 81%, and 79% respectively [9].

Mangal Singh, Md.Tabrez Nafis and Neel Mani have worked on Similarity Evaluation and Sentiment Analysis on the Reviews for Heterogeneous Domain product. They had demonstrated scaling and sentiment classification with similarity evaluation among the reviews on the product. And the Reviews data is pre-processed and cleaned for the data pre-processing [10].

Mnahel Ahmed Ibrahim and Naomie Salim had on the sentiment analysis of Arabic text tweets extracted from Twitter, and with the help of various classifiers like SVM, K-Nearest Neighbour, Naïve Bayes were used to find best results [11].

Eibe Frank and Albert Bifet proposed challenges that Twitter data define, focused on classification problems and consider for sentiment analysis [12].

Isabella et al., have proposed movie reviews for sentiment analysis and evaluated feature selectors to improve the performance of the classifiers [13].

Pang B. in 2002 measured the problem of characterizing reports by general feeling, for example, deciding if an audit s positive or negative. Utilizing film surveys for information, that basic machine learning strategies absolutely outflank user created baselines. Nonetheless, the techniques that were utilized are Naïve Bayes, Support Vector Machine and Maximum Entropy. These techniques don't do well on classification of sentiments as on

customary subject related classification of sentiments. The author finishes up by looking at factors that make the sentiment classification issue much difficult [3].

Bin Liu.2012 expressed that Opinions are vital to every single human activity and are key influencers of our practices. This isn't valid for people yet additionally valid for organizations. Opinions and its connected ideas, for example, feelings are the concern of investigation of sentiments. The initiation and fast development of the field match the online networking e.g., surveys, sites, micro blogs, Twitter, and social networks, on the grounds that without precedent for mankind's history, an immense volume of obstinate information is being recorded in computerized frames. Since mid2000, sentiment examination has become a standout amongst the most dynamic research regions [4].

Gautam G. and Yadav P in 2014 define another method for communicating the feelings and emotions of people. In general, it is a way which tremendously measures the information where clients can see the emotions of different clients which are categorized into various classes of sentiments and are progressively developing as a main factor in basic leadership. The work defined is useful to view the data as the quantity of tweets where sentiments or emotions are either good or bad, or neutral [5]

Lo Y.W, in2013expressed that the Web has significantly changed the best approach to express sentiments on specific items that has been obtained and utilized, or for administrations that we have gotten in the different businesses. Sentiments or surveys can be effortlessly posted, for example, in dealer destinations, audit entryways, web journals, Internet gatherings, and considerably more. This information is usually alluded to as client created service or client created media. Both the item

makers, and in addition potential clients are extremely intrigued by this online verbal, as it gives item makers data on their clients different preferences, and additionally the positive and negative remarks on their items at whatever point accessible, giving them better learning of their items restrictions and favourable circumstances over contenders; and furthermore giving potential clients helpful and direct data on the items or potentially administrations to help in their buy basic leadership procedure[6]

John C and Jonathon R. specified an imperative sub-errand of emotions examination is classification of polarity, in which content is delegated being good or bad. Machine learning methods can play out this classification adequately. Be that as it may, it requires a substantial corpus of preparing information, and various examinations have shown that the great execution of supervised models is reliant on a decent match between the preparation and testing information as for the area, subject, and time period. Pitifully supervised procedures utilize an extensive accumulation of unlabeled content to decide sentiment, thus their execution might be less subject to the area, subject [19].

III. PROPOSED SYSTEM

A. Overview

When a user of YouTube search for educational content, the results given to the user is based on how good is engagement of audience with video. This includes number of views, likes and comments. These parameters are considered to be judging parameters for video and depending upon that the videos are ranked and displayed. The parameters seem to be incomplete without understanding the sentiments of user. The sentiment of user can be obtained from comment section. Suppose if a user is interested in watching long videos or tutorial series then all results

appear to be similar and user has no idea which series to go with. So here by observing positive sentiment of top combined comments, user can get somewhat idea.

The proposed system considers the sentiment of top comments of every video along with other parameters and display the results. The system is web application which takes input as search keyword and displays top nine videos related to educational content only. Also, the application displays statistics of each video with graphical representation.

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.
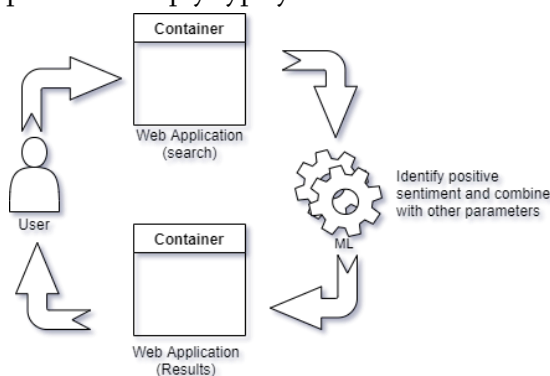
Figure 1. Overview

The system has user interface where a layman user can interact with the system. The interface hides the unnecessary information from user like the working behind and exposes them only to the functionalities available. The user has flexibility to search educational content on the system. The user search keywords are processed and analysed deeply in the backend. All the related content is first fetched from YouTube into the system. The data fetched is semi-structured data that is in the JSON format. The data is segregated and required parameters are selected. It checks the category code matches or not.

There are multiple search results for single search keyword. Some of the results are processed especially the comment section. Top N comments of top M results are selected for processing otherwise the data would be massive. To determine nature of comments, a machine learning model is created. The model is

prepared by training with large data set that classify the comment as positive or neutral. All N comments are processed and analysed. One particular score is given to each of the M results based on N comments. Apart from scores, the likes, dislikes, views data are also captured of all M results and stored in a list. The list is then sorted considering all the parameters and the ranking of videos are decided.

The user types the keywords of educational content and click on search. In the backend, the application fetches information required for analysis of top YouTube results with the help of YouTube API. We fetch number of views, likes and comments as well as the top user comments. A machine learning model is created which identify the sentiment of comments and give result as number which determines positivity. So now for each video there will be additional parameter sentiment. Combining all parameters together the videos are again sorted based on some algorithm and then displayed to user. The video with good positivity score, views and likes will top the results.

B. Flowchart
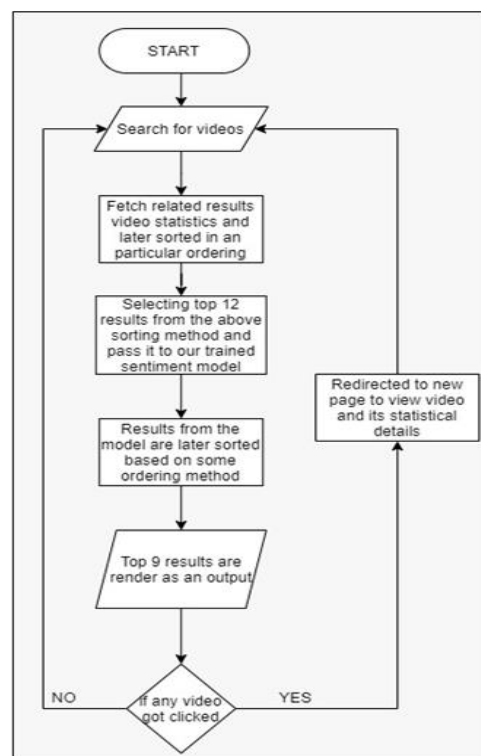
Figure 2. Flowchart

When user opens application there is simple interface with search input and search button. To search any video the user needs to type search keyword related to education category and click search button. The front end is design with bootstrap and in the backend, we have python managing routes, working with API, analysing comments, performing sorting operation and displaying results.

Using the users query input, an actual search is made on YouTube without showing it to the user. YouTube API is created which helps to collect the metadata from YouTube. Parameters that were extracted were likes, dislikes, number of views and comments. For a particular user query fifty video metadata is collected. For each video from the result top hundred comments are considered for processing. All these comments are passed through a machine learning model to give output.

The machine learning model is supervised model. It is trained using amazons' mobile reviews dataset to determine sentiment of unknown comment. Total there are fifty videos and each has hundred comments, so total of five thousand comments are passed through the model. Each video will be assigned one particular score based on its hundred comments analysed. So, all the fifty videos will get some score based on their processing. The output will be displayed as video links which are in order according to sentiments. The user can click and play the video where he can see graphical analytics of sentiments also.

IV. IMPLEMENTATION

The project is divided into four modules
1. Fetching required data using API
2. Creating a ML model and training
3. Sorting the videos
4. Integrating with web application

A. Fetching data using API

The data such as number of likes, views, comments and top hundred comments are fetched for each top video from YouTube. Fetching data of YouTube can be done using two methods first is web scraping in which we have to create a crawler which take an input of video-id and based on the id it will start scraping that web page and look for all the details need for further analysis. Main problem with web scraping is that it required too much time to run and given 10 video-id and scraping video comments can take up to 5-10 mins for a single run. Second method is to use API provided by google itself which is used by lakhs of developer on their personal or live projects. YouTube API Key can be found on google developer website, create a google account, sign in and enable YouTube Api and you are ready to start using it.

Using API, we'll send a request to YouTube with a given video id and it will return the response with json format instantly. Hence enabling us to make our project much faster for end user. In python there is api client module which helps to build a YouTube api given that you provide it with YouTube api key. Using this now we have access to fetch data using various different methods provided by YouTube, reference can be found in google developer website. Here is simple example of how to search for a video of query python which is restricted to education category only and max search results should be 50.

Figure 3. YouTube API

Figure 4. API Code

Above output are same as if you had searched python on YouTube, it would have return same results only. This method it also used to found comments of each results video id. Comment data are limited to max result of 100 which can be solved using next page token provided in the response itself and loop it through till it found all the comments.

B. Machine Learning Model

A machine learning model is created to predict sentiment of comments. The model is trained using Amazon's Mobile Product Reviews as it had sufficient three lakhs' entries. It uses logistic regression to classify the sentiment. In our model we are using Amazon's Mobile Product Reviews which is quite similar to how to people on YouTube comments on a video. Here is a sample of our data set.

From above we can see that they are pretty much same as some kind of YouTube comments. The column which is required are Rating and Reviews column only. Rating column gives us some idea about how the review will be, positive or negative. So, we will be ignoring neutral comments because we don't want our model to predict some bad reviews or suggestion or questioning as a positive sentiment. So, while training we had ignored all the data set which contains rating of 3. Positive sentiment is of rating 4

and 5. We are not focusing on negative or neutral sentiment.

Figure 5. Dataset

Above pic is the real snippet of our code to create model for our project. First, we split our data into two set one for training and other for testing. Next, we used count vectorizer to fit our training data with parameter of min_df=5 and ngram_range of 2 to 3, by setting ngram our vocabulary set will change accordingly. Then we passed it to our logistic regression model derived from sklearn.linear_model. After transforming test data set, we can predict the output of our result and thus the accuracy score is defined. Our model accuracy score is about 95%.

```
X_train, X_test, y_train, y_test = train_test_split(df['Reviews'],
                                                    df['Positively Rated'],
                                                    random_state=42)

vect = CountVectorizer(min_df=5, ngram_range=(2,3)).fit(X_train)

X_train_vectorized = vect.transform(X_train)
print(X_train_vectorized.shape)
model = LogisticRegression(C=500)
model.fit(X_train_vectorized, y_train)

predictions = model.predict(vect.transform(X_test))
print('AUC: ', roc_auc_score(y_test, predictions))
```

Figure 6. Training and Prediction

C. Sorting

Along with count of views, likes and comments, additional parameter called sentiment is added. Considering all these parameters, the ranking of video is decided. The sorting is done using lambda

function. Python as already has sorted methods which take iterable as input and return a new sorted list from the items in iterable. Sorted has additional parameter of key which take in what order or how you want your iterable to be sorted in. So, we had passed tuple with 4 different items in some specific order which then sort it in that manner. We had passed tuple of which consist of Like count, Comment Count, View Count, like to dislike ratio respectively. Hence using this approach, we get our result which is not dependent on only one parameter.

D. Web Application

The web application is created using flask which is a framework of python. The web application serves as user interface where user will search and obtain the results. To build web application first we have to decide on which framework to start with. Web Application Framework or simply Web Framework represents a collection of libraries and modules that enables a web application developer to write applications without having to bother about low-level details such as protocols, thread management etc.

V. RESULTS AND DISCUSSION

A. Web Application

The required data for ranking videos is successfully fetched using API. Top YouTube video comments related to keywords are passed through machine learning model to detect sentiment and positive sentiment of each video is identified. The sorting makes use of number of likes, views, comments and sentiment of comments.
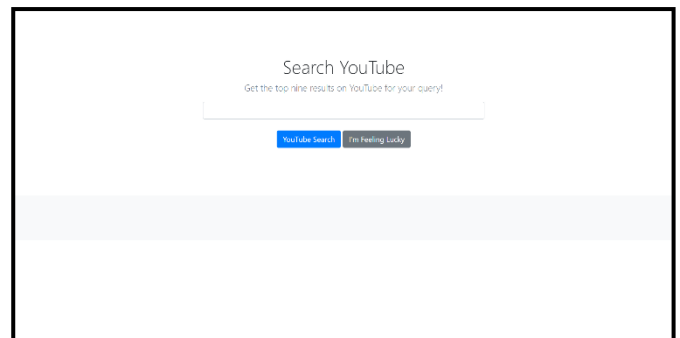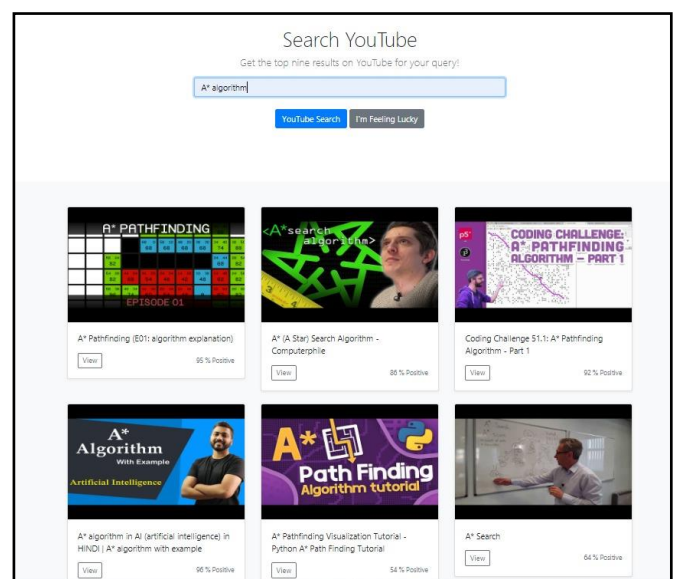

Figure 7. GUI

Figure 8. Result against search

Figure 9. Embedded Video Player

Figure 10. Sentiment Chart

VI.CONCLUSION

Sentiment analysis is not a new term nowadays, but the utilization of this concept plays very important role. Proper interface is created for better user interaction. YouTube API is used effectively to fetch all the required parameters. Obtained data is cleaned before processing. Proper scores are allotted to positive and neutral sentiments. A machine learning model has been trained for the sentiment analysis of the YouTube comments followed by the pre-processing of the dataset.

For classification of comments, logistic regression algorithm is used and achieved an accuracy score of approximate 90% on amazon's mobile reviews dataset which had over 3 lakh entries. Objective of proposed system is to provide reliable and quality results of search against the keywords provided by user is achieved by displaying top nine results as per order.

VII. FUTURE SCOPE

One use case is analysing the trending video as many times, there are videos with more views and likes on trending but if you will use sentiment analysis you will easily be able to find the most useful video of a particular channel, celebrity, category, etc. To know about customer reviews about a particular service or product, the proposed system will help to know about their views. The project can be further be extended to create technology of writing positive, optimistic and professional emails or content. This system can be hosted as a standalone application online for public use. It can be combined with recommendation system to increase reliability and productivity. Since we are focusing primarily on educational content, it can be thought of extending it to other categories too.

VIII. ACKNOWLEDGEMENT

IX. REFERENCES

[1]. "Sentiment Analysis of Arabic tweets Using RapidMiner" Salha al Osaimi and Khan Muhammad Badruddin, Dept of Information System, Imam Muhammad ibn Saud Islamic University, KSA.

[2]. "Sentiment Analysis of English Tweet Using Rapidminer", Pragya Tripathi, Santosh Kr Vishwakarma, and Ajay Lala, International Conference on Computational Intelligence and Communication Networks, 2015, pp. 668-672.

[3]. "Opinion mining and sentiment analysis in Found Trends Inform Retriev", Pang B, Lee L, 2 (2008), pp. 1135.

[4].   "Sentiment analysis and opinion mining", Liu B in Synth Lect Human Lang Technol (2012)

[5].   "Sentiment Analysis of Twitter Data using Machine Learning Approaches and Semantic Analysis", Gautam G, Yadav P.

[6].   A review of opinion mining and sentiment classification framework in social network, Lo, Y.W, Potdar, V.

[7].   Weakly Supervised Techniques for Domain-Independent Sentiment Classification, Jonathon Read, John Carroll.

[8].   "Correlation Base Feature Selection for Movie Review Sentiment Classification", K. Bhuvaneshwari and R. Parimala, IJARCCE, vol. 5, no. 7, July 2016.

[9].   "Sentimental Classification of Social Media using Data Mining", Farhan Laeeq, Md. Tabrez Nafis and Mirza Rahil Beg, in IJARCS.

[10].   "Sentiment Analysis and Similarity Evaluation for Heterogeneous-Domain Product Reviews", Mangal Singh, Md. Tabrez Nafis, and Neel Mani, in IJCA, vol. 144, no. 2, June 2016.

[11].   "Sentiment Analysis of Arabic Tweets: With Special Reference Restaurant Tweets", Mnahel Ahmed Ibrahim and Naomie Salim, in IJCST, vol. 4, no. 3, May – June 2016, pp. 173–179.

[12].   "Sentiment Knowledge Discovery in Twitter Streaming Data", Albert Bifet and Eibe Frank, from University of Waikato, Hamilton, New Zealand.

[13].   "Analysis and Evaluation of Feature Selectors in Opinion Mining", J. Isabella & Dr. R.M.Suresh, Indian Journal of Computer Science and Engineering, (ISSN: 0976-5166), Dec 2012-Jan 2013, Vol. 3 No.

[14].   "Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier," A. N. Muhammad, S. Bukhori and P. Pandunata, 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), Jember, Indonesia, 2019, pp. 199-205, doi: 10.1109/ICOMITEE.2019.8920923.

Cite this article as :