International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 12 | Dec 2020 www.irjet.net p-ISSN: 2395-0072

Sentimental Analysis of YouTube Videos

Aditya Baravkar[1], Rishabh Jaiswal[2], Jayesh Chhoriya[3]

[1-3]*Student, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, 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 these 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, and comments is considered. A machine learning model is prepared to identify sentiments of top comments of each video and ranking the videos by sorting with help of multiple parameters. A user-friendly web application is created where user can search for videos and see the statistics. The objective is to provide well analyzed and relevant educational videos to the budding students by reducing valuable search time.*

***Key Words*: Sentiment Analysis, Logistic Regression, Machine learning, Flask Web Application, Statistics.**

1. 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 as it gave the best accuracy score which was 96.2% as compared to other algorithms.

2. SUPERVISED ALGORITHM

Supervised learning refers to an approach that teaches the system to detect or match patterns in data based on examples it encounters during training with sample data. Supervised learning can be used with data that is known to predict outcomes and results. In supervised learning, the job of the algorithm is to create a mapping between input and output. The primary applications for supervised learning are in systems that solve classification or regression problems.

2.1. Logistic Regression: Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables

3. RELATED WORK

Many research papers are published in field of sentimental analysis. We have reviewed following papers to get a better understanding of this field. The review papers and their description is given below.

Salha al Osaimi and Khan Muhammad Badruddin [1] proposed an automatic approach to predict sentiments for informal Arabic language. They made use of Natural Language processing along with artificial intelligence. They further came to a conclusion that emotion icons play a vital role in development of an accurate classifier.

Pragya Tripathi, Santosh Kr Vishwakarma, and Ajay Lala [2] proposed a system to perform sentiment analysis of English tweets using rapid miner platform. They built two classifiers and also tested the dataset using Rapid Miner. Further they compared both the classifiers in order to find the better results.

Abbi Nizar Muhammad, Saiful Bukhori, Priza Pandunata [3] has used Naive Bayes and Support Vector Machine to classify comments of YouTube as positive and negative.

The data set is divided into 7:3 ratio .i.e. 70% training and 30% testing data set. The two algorithms are combined and acquired precision of 91%, recall 83% and flscore of 87%.

Song Qin, Ronaldo Menezes, Marius Silaghi [4] have created YouTube recommender network that works on variety of features and such as high rating, most viewed count, etc. and then captures important characteristics among them. The videos are ranked based on the information obtained from social network of users. Watching a particular video on youtube recommends videos of similar genre or domain which is not always correct as user may tend to love other domain. The data is collected from youtube API. An undirected and weighted graph is created where nodes represents the videos and edges is link between two nodes if there is user who commented on both nodes. The distribution of tags inside communities is diverse which is demonstrated. The weight of edge predicts the strength of relation between two nodes. Utility value is assigned to the nodes. Videos are first recommended to users from highest utility value to lowest and are able to recommend categorized videos through community characteristics.

Weilong Yang and Zhensong Qian [5] have shown some deep understanding of characteristics of videos from youtube of different categories. Study included of video duration, user engagement, view source, view counts and growth trends. Analysis of growth trend, view counts were done. Those patterns were very different but intuitive.

Tim O'Keefe and Irena Koprinska [6] introduce a number of feature selection and feature weighting methods for sentiment analysis. They used more three feature weighting methods (SWN-SG, SWN-PG and SWN-PS) and where compared by their performance with the standard and popular feature such as frequency, feature presence and TF-IDF methods. All the experiments were conducted using two main classifiers, SVM and NB, over the movie review data set. The results were promising as it is comparable with the previous state of the art of 87.3% and 91% with smaller uses of features.

4. PROPOSED SYSTEM

The proposed system considers the sentiment of top comments of every video along with other parameters and displays 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. 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 custom algorithm and then displayed to user. The video with good positivity score, views and likes will top the results.

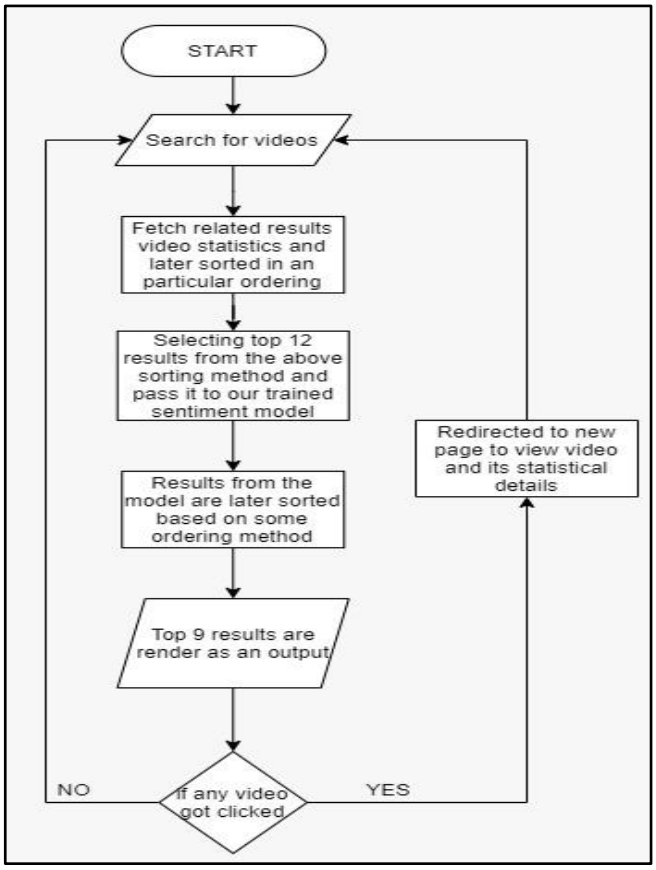

Fig -1: Workflow Diagram

A simple web application is created using Flask where user can search for reliable videos. Top results according to normal YouTube search are collected. Positive sentiment score is identified for each video. Combining number of likes, views, comments and obtained sentiment is used to sort and rank the videos. Top nine results are displayed with statistical details of each video along with graphical representation.

5. METHODOLOGY

Our project mainly divided into four sub modules. These four modules are as follows:

a) Fetching required data using API
b) Creating ML model and training
c) Sorting based on parameters

d) Integrating in a Web application

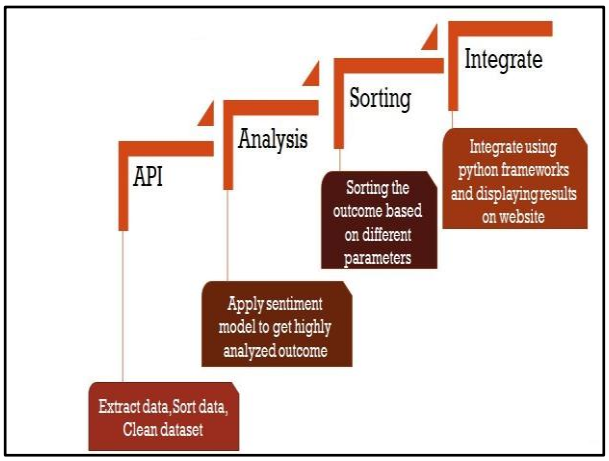

Fig -2: Methodology

5.1. Fetching require data using API

The YouTube Application Programming Interface (YouTube APIv3) allows developers to access video statistics by making REST and XML-RPC calls using URL or some request modules. This is used as the mediator between our query request and YouTube. Comments of each video are fetched using this API, provided expected url request.

5.2 Creating ML model and training

Logistic Regression model was used to train model because we only concerned with positive reviews so the logistic regression model helps to classify if a certain class or event exist or not.

The data which was used to train this model was based on Amazon mobile reviews which tells provide us quiet a large bag of words on reviews which resembles to the comments made on YouTube videos.

```
z=vect.transform(['Not that bad ',
                  'cat is not very good animal',
                  'Very satisfied with the service.',
                  'was not in good condition but does work good',
                 'the reasons for the 3 star rating was it was in my opinion better than my iphone 4s but it tends to randomly c
                  "Just... not good. The phone has great screen resolution, storage is low, you need an SD card to do anything,
v=model.predict(z)

print(v)

[0 0 1 0 1 0]
```

Fig -3: Sentiment classification

5.3 Sorting based on parameters

Rankings of the list of videos where sorted based on like count, comment count, sentiment stats from ML model, view count respectively in order. This helps the low ranking YouTube videos who may not have that many views compare to first ranking videos on YouTube to get attention.

5.4 Integrating in a 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.

6. RESULTS

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.

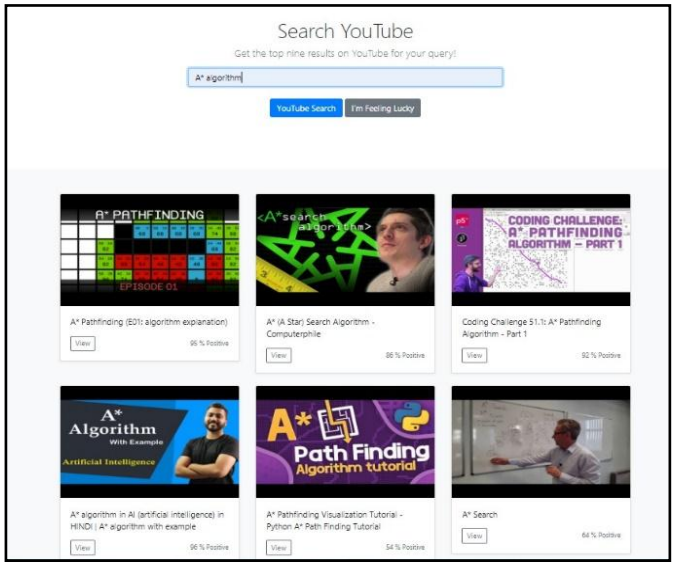

Fig -4: Web Application

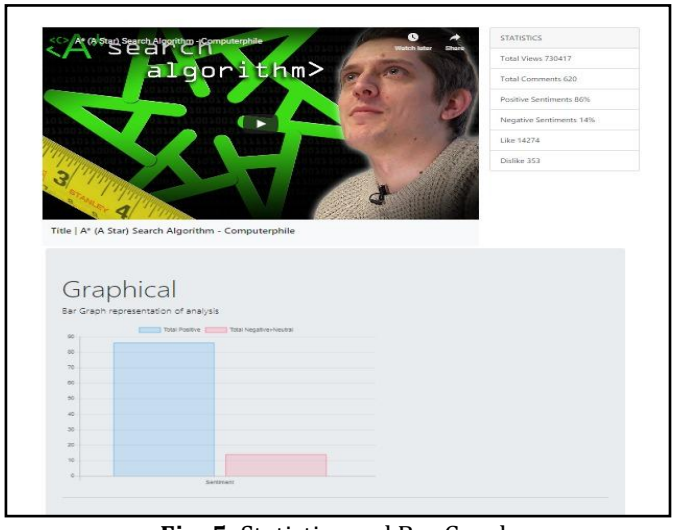

Fig -5: Statistics and Bar Graph

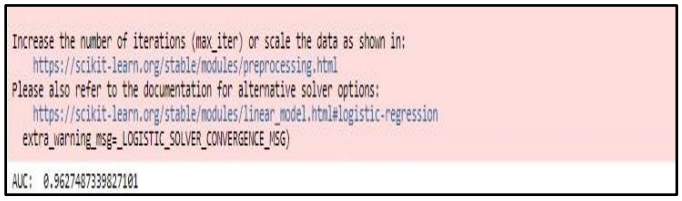

Fig -6: AUC Score

7. CONCLUSIONS

Less relevant tutorial series or marathons can rank up their videos using targeted keywords, audience retention and video engagement. Video engagement includes sharing of video, number of subscribers, likes and views. Everything depends on number. The nature of comments is not taken into consideration. The result of search changes by adding sentiment as parameter along with other. This helped hidden gem videos to get noticed. One more parameter is added called sentiment which analyzes the nature of comments. Top video comments are passed through machine learning model and considering multiple parameters the videos are sorted and displayed. The system 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. Model can be trained using true comments fetched from API for more accurate decision making. Project can be used for more specific topic related structure.

ACKNOWLEDGEMENT

The success and final outcome of project requires lot of guidance and assistance we are extremely privileged to have the same for completing our project synopsis. All that we have done is due to supervision and assistance and would not forget to thank them. We express gratitude towards our project guide Prof. Swati Sharma for her valuable and timely advice during various phases of project despite of busy managing schedule. We would thank her for having faith in our capabilities, providing required support and encouragement, flexibility and patience. Finally, we would like to thank everyone who is involved directly or indirectly in our project synopsis.

REFERENCES

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

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

[3] Abbi Nizar Muhammad, Saiful Bukhori, Priza Pandunata, “Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier”, 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE).

[4] Song Qin, Ronaldo Menezes, Marius Silaghi, “A Recommender System for Youtube Based on its Network of Reviewers”, 2010 IEEE Second International Conference on Social Computing.

[5] Weilong Yang, Zhensong Qian, “Understanding the Characteristics of Category-Specific YouTube Videos”, December 2011-citeseerx.

[6] O’Keefe. T and Koprinska I, “Feature Selection and Weighting in Sentiment Analysis,” in Proceeding of 14th Australasian Document Computing Symposium, Dec 2009, Sydney, Australia.