Predicting ratings of YouTube videos based on the user comments

Conference Paper · September 2021

CITATIONS
0

READS
22

2 authors, including:

Pirunthavi Sivakumar
University of Vavuniya
5 PUBLICATIONS **0** CITATIONS

SEE PROFILE

FARS 2021
The 2nd Faculty Annual Research Session
Faculty of Applied Science
University of Vavuniya, Sri Lanka
September 15, 2021

Predicting ratings of YouTube videos based on the user comments

Pirunthavi, S.
Department of Information and Communication Technology
Rajarata University of Sri Lanka
psivakum@tec.rjt.ac.lk

Jayalath, E.
Department of Computer Science and Informatics
Uva Wellassa University of Sri Lanka
jayalath@uwu.ac.lk

ABSTRACT

This project attempted to develop a model to predict the rating of a YouTube video based on the user comments. We extracted the user comments from many YouTube videos to make the sentimental analysis. The keywords were extracted from the user reviews using the Natural Language Processing technique, and those reviews were categorized into positive or negative predicated on the sentimental analysis. The Naïve Bayes model was trained to utilize the user reviews extracted from YouTube to presage the rating of a video. The model was tested on original datasets, and the precision of that was evaluated respectively. Conclusively, one conclusion has been met that the rating of a video cannot be presaged through the user comments. The performance of the model is decent enough compared to the subsisting models in the literature. YouTube sanctions extract an inhibited number of user comments, and hence, this factor could negatively affect the rating presage's precision.

Keywords: Naïve Bayes, Natural language processing, Sentimental analysis, Video rating.

INTRODUCTION

Visually examining videos on YouTube has become one of the most popular regalement factors in the 21st century. YouTube is an online video platform where users watch around 6 billion hours of videos every month, and hundreds of hours of video content are uploaded to YouTube servers every minute (GCFGlobal 2005). The rating of a video is primarily subject to individual discernment. In the present computerized world, an individual's assessment is as remarks found on the web for the most part. The research topic proposes a model for predicting the rating of a video on YouTube based on the user comments. Data mining techniques allow predicting the rating of a video found online (Meenakshi et al., 2018). The fundamental trouble is to extract valuable data from YouTube in the arrangement of the source information. In this exploration, modified word references were utilized from the web, where different words that users generally use in surveys will be assembled and omitted at a particular rate depending on the administrator's decision. By utilizing the Naïve Bayes Algorithm, we have tried to predict the rating of a video found on YouTube (Meenakshi et al., 2018).

METHODOLOGY

Data collection

Python language was used to extract user comments from YouTube of each video using the YouTube APIs. A thousand remarks for each video were gathered since YouTube permits the greatest number of 1000 remarks for every video to be gotten to through the APIs (Krishna, 2014). For extracting the YouTube comments, the queries will be asked by the system from the user, such as video ID of the video which comments to be extracted and the number of comments are to be extracted; it should be less than 1000 (Krishna, 2014)

Data pre-processing

Data scrapping: Other online tools need more processing power and storage (Pirunthavi et al., 2020); thousand of comments were scrapped from YouTube using YouTube API.

Data cleaning: The information was cleaned to choose just the applicable traits, which would help in the examination and expectation. The accentuations, emojis and numerous pitfalls were removed out from the user remarks in this stage (Pirunthavi et al., 2020).

Tokenization: In this step, longer text strings were divided into smaller pieces or tokens like words, keywords, phrases, and other elements called tokens.

Stemming: Various types of words were decreased, and suffixes/prefixes in sentences were eliminated by utilizing data stemming.

Analysis

Data labeling: The words were labelled as positive (+) and negative words (-). The positive and negative dataset was acquired from the web, and as per those datasets, the prepared information was marked into positive and negative words.

Overall measure: User comments may contain positive or negative features. So, the percentage of positive and negative features can be calculated. By this rate, the rating of a video can be acquired out of a hundred. By utilizing this, the rating of a film can be determined out of ten. Utilizing the Naive Bayes algorithm, the mean qualities were determined for the dataset gathered from YouTube.

Table 1: Rating prediction sample output

Video name (movie)	**Expected rating of the video**	**Actual rating of the video**
Avengers	0.8	0.35
Birth of the Dragon	5.5	2.50

Implementation

Naive Bayes algorithm is typically used in text classification, and it is anything but an indistinguishable strategy to anticipate the likelihood of different classes supporting different qualities. This Algorithm was utilized to anticipate the rating of a video by utilizing client remarks. After pre-preparing (data cleaning, tokenization, and stemming), the calculation changed over the dataset into a frequency table.

YouTube API was utilized to separate the remarks from YouTube recordings. By giving the video IDs, the remarks of every single video were extricated. Here, the video ID and several remarks were given as data sources and got positive and negative slant investigation rate as yield. YouTube API permits designers to get to recordings insights and YouTube channels information through a REST API call. YouTube permits the greatest number of 1000 remarks for each video to be gotten through the APIs (Krishna, 2014). To acquire better exactness, 1000 remarks were removed from every video; in any case, the precision could be decreased when the quantity of remarks is less. Then, at that point, it is anything but a likelihood table that discovering the conceivable outcomes. For example, the likelihood of positive words is 0.38 (38%), and the likelihood of negative words is 0.61 (61%).

Testing and evaluation

Nowadays, people will in the everyday quest for information and sentiments to assemble their own judgment concerning. Verbal is seen as a critical wellspring of information for the people (Krishna, 2014). These electronic action words infer positive or negative decrees made on YouTube remarks. Whether or not it is anything but a productive or opposite remark, people will assemble their own assessment concerning the everyday quest for the remarks. The dominant parts of the people give likes or dislike upon their own assumptions. YouTube remarks with more responses will go to the most elevated mark of the line. Nowadays, a single negative comment can spread rapidly (Krishna, 2014). Thus, that kind of remark will get a more substantial number of responses than the other positive remarks. So in most of the accounts, negative remarks exist in the most elevated place of the line. Exactly when we are rejecting client remarks from YouTube, we can isolate them from the top. We can't get all of the remarks, considering the way that YouTube grants a restriction of 1000 remarks for each video to be got past the APIs (Krishna, 2014). Remarks with more responses will be on the top and removed through the APIs. Thus most bad remarks are removed from the positive remarks. Usually, the decided mean assessment of negative conclusion is more noteworthy than the decided mean assessment of positive slant. Table 1 indicates the yield of the expected and actual ratings of some movies and expected ratings are obtained from the IMDb website (IMDb 2021).

A dummy YouTube channel was made to check if the created model is performing precisely. In this way, a few recordings were transferred to the made YouTube channel, and a portion of the clients was approached to put remarks for the recordings. At first, they were approached to put just adverse remarks, in view of those remarks, the model determined negative assessment as 100% and positive assumption as 0%. Then, at that point, half of the users were approached to put just adverse remarks and half of them to put just certain remarks; in light of those remarks, the model determined negative slant as half (50%) and positive notion as half (50%). Then, at that point, the users were approached to put just certain remarks; in light of those remarks, the model determined positive assumption as 100% and negative opinion as 0%.

CONCLUSION

The Naïve Bayes Algorithm has been used as a primary line in tasks related to texts as it is excellent in classifying texts. YouTube API separates more negative remarks than positive remarks as remarks with more reactions (likes/dislikes) will go to the first spot on the list since humans react to negative remarks rather than positive ones based on the Word-of-Mouth and their opinion. For the most part, the yield of the model gives a more negative rate than a positive rate. Along these lines, more often than not, the model can't provide precise outcomes. At last, a choice has been met, that we can't anticipate the rating of a YouTube video through the person's remarks. The created model is working effectively; however, we unable to acquire the regular rating depending on the user remarks.

REFERENCES

GCFGlobal. (2005) What is YouTube?. Website of Goodwill Community Foundation. [online] <https://edu.gcfglobal.org/en/youtube/what-is-youtube/1/> [Accessed 15 June 2021]

IMDb. 2021. IMDb: Ratings, Reviews, and Where to Watch the Best Movies & TV Shows. [online] Available at: <https://www.imdb.com/> [Accessed 15 July 2021].

Krishna, A. (2014) *Polarity trend analysis of public sentiment on YouTube*. MSc Thesis, Iowa State University, Ames.

Meenakshi, K. Maragatham, G., Agarwal, N., and Ghosh, I. (2018) A Data mining Technique for Analyzing and Predicting the success of Movie. *Journal of Physics*. DOI: 10.1088/1742-6596/1000/1/012100.

Sivakumar, P, Rajeswaren, V.P, Kamalanathan, A, Ekanayake, E.M.U.W.J.B, and Mehendran, Y (2020) Movie Success and Rating Prediction Using Data Mining Algorithms. *Journal of Information Systems and Information Technology*, 5(2): 72-80.