Journal of Applied Information Science
11 (1) 2023, 16-20
http://www.publishingindia.com/jais

TubeRate: YouTube Rating System

Prahlad Gurjar[1*], Naitik Yadav[2], Ahimsa Jain[3] and Nisha Rathi[4]

[1]Student, B.Tech (Computer Science and Information Technology), Acropolis Institute of Technology and Research, Indore, Madhya Pradesh, India. Email: prahladgurjarci19@acropolis.in
[2]Student, B.Tech (Computer Science and Information Technology), Acropolis Institute of Technology and Research, Indore, Madhya Pradesh, India. Email: naitikyadavci19@acropolis.in
[3]Student, B.Tech (Computer Science and Information Technology), Acropolis Institute of Technology and Research, Indore, Madhya Pradesh, India. Email: ahimsajainci19@acropolis.in
[4]Professor, B.Tech (Computer Science and Information Technology), Acropolis Institute of Technology and Research, Indore, Madhya Pradesh, India.
*Corresponding Author

Abstract: YouTube is one of the comprehensive video information sources on the web where video is uploaded continuously in real time. It is one of the most popular sites, where users interact with sharing, commenting and rating (like/views) videos. Generally the quality, relevancy and popularity of the video is maintained based on this rating. Sometimes irrelevant and low quality videos ranked higher in the search result due to the number of views or likes, which seems untenable. To minimize this issue, we present a Natural Language Processing (NLP) based sentiment analysis approach on user comments. A sentiment classifier model is built which identifies tweets positive, negative or neutral. In this technique, the collected corpus was divided into 3 sets namely positive emotions- happiness, amusement or joy; Negative emotions- sadness, anger or disappointment and Neutral-text doesn't contain emotions. YouTube data is also automatically classified into positive, negative and neutral according to query terms used in user review comments. In the paper the author uses Parts Of Speech (POS) polarity technique and tree kernel technique. Research work uses two types of resources such as a hand dictionary of emotions and a dictionary collected from the web. Different types of classification and feature extraction algorithms are used [1]. This analysis helps to find out the most relevant and popular video of YouTube according to the sentimental analysis of the comments posted by users on YouTube videos. Using sentiment analysis, these users' opinions and emotions can be extracted and quantified. TubeRate examines the current papers on sentiment analysis on YouTube comments as well as present the work done on proposed idea of user comment based YouTube video rating that is analyzed through polarity and segregation as positive, negative or neutral. This can be useful in predicting the like proportion of a YouTube video. These ratings will be on a scale from 1 to 5, where 1 means extremely dissatisfied and 5 means extremely pleased with the content of the video.

Keywords: Machine learning, Sentimental analysis, Support vector machine, YouTube.

I. Introduction

In recent years YouTube has made space for millions of users to share their information and opinion with each other. It is a popular online video site that has uploaded a huge amount of videos on different areas that may be description about a product, their services, or tutorials about software, likewise YouTube received a huge amount of comments from the viewers. YouTube is recognized as the second most popular website in the world by Alexa Internet [2].

The current YouTube algorithm only utilizes the statistics of user engagement while suggesting content. Through the use of the TubeRate app, the user will be able to view ratings of the videos which will be based on sentiment analysis performed on the comments of those videos [3]. Perform sentiment analysis on 2053 movie reviews collected from the web Movie Database (IMDb). They examined the hypothesis that sentiment analysis is often treated as a special case of topic-based text classification. Their work depicted that standard machine learning techniques such as Naive Bayes or Support Vector Machines (SVMs) outperform manual classification techniques that involve human intervention. However, the accuracy of sentiment classification falls in need of the accuracy of ordinary topic-based text categorization that uses such machine learning techniques. They reported the simultaneous presence of positive and negative expressions. Sentiment analysis is useful for quickly gaining the whole idea by using large amounts of text data and it will be helpful to understand the user's opinion. The sentiment is decided by the frequency of the positive-oriented and negative-oriented words. So, in these lexicon-based methods [4], domain dependency makes them less suitable for the domains without specialized lexicons.

However, these techniques often have a low precision rate due to the lack of powerful linguistic resources. Sentimental analysis is additionally referred to as opinion mining that means to find out or identify the positive, negative, neutral opinions, views, attitudes, impressions, emotions and feelings indicated in the text. Comments serve the purpose of helping the community to filter relevant opinions more efficiently [5, 6, 7]. For the proposed work, we collected the data from the YouTube comments of the public and measured the attitude of the users towards the aspects of a video which they describe in text to rate the video in our search engine.

II. Related Work

Many research papers are published in the field of sentimental analysis. We have reviewed the 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 [8] 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 [9] proposed a system to perform sentiment analysis of English tweets using a 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 and Priza Pandunata [10] have used Naive Bayes and Support Vector Machine to classify comments of YouTube as positive and negative.

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

Weilong Yang and Zhensong Qian [11] have shown some deep understanding of characteristics of videos from YouTube of different categories. Study included 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.

III. Literature Review

Title of Project	Name of Author	Year of Publication	Description	Limitations	Improvement
YouTube Comments Sentiment Analysis	Ritika Singh [12]	2021	Presented a Sentiment analysis system for YouTube comments, by using different machine learning classifiers to process the data and optimize the classification results.	Considered the emojis too, for the classification of the dataset.	In our proposed model we used different features to classify the comments which removes the emojis, the abusing words and the repetition of words.
Sentimental Analysis of YouTube Videos	Aditya Baravkar [13]	2020	The proposed system considers the sentiment of top comments of every video 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. Only positive comments are taken into consideration.	The project is only based on educational content on YouTube. Only positive comments are taken into dataset, all the negative and neutral comments are ignored.	In our proposed model we considered all the videos present on YouTube. The polarity is generated by considering positive, negative and neutral comments.
Analysis and Classification of User Comments on YouTube Videos	Hanif Bhuiyan [14]	2020	This paper illustrates an automatic process for finding useful video by sentiment analysis of user's comments based on Natural Language Processing. Their approach was to evaluate the quality, relevancy and popularity of YouTube videos considering the relationship of user's sentiments expressed in comments.	They included the repetition of words, which leads to inaccuracy of results sometimes. They only considered 10 topics from YouTube randomly for the data testing.	In our proposed model, we have taken random YouTube videos for testing and worked on the important features that would lead to inaccuracy of results sometimes, such as reputation of words, ignoring abusive language.

Title of Project	Name of Author	Year of Publication	Description	Limitations	Improvement
Retrieving YouTube Video by Sentiment Analysis on User Comment	K. M. Kavitha [15]	2018	They extracted the comments using the video URL and manually categorized them into four classes. They considered a few random comments and the description of the video.	YouTube video should have an associated description.	It is not necessary for a YouTube video to have a description, and there the project fails. In our proposed model, we have taken the top comments for the optimization of results.

IV. Motivation

Thousands of hours worth of content is being created and shared on YouTube throughout the day by content creators of YouTube. It has become really difficult to spend time watching such videos which may have a longer duration than expected and sometimes our efforts may become futile if we couldn't find relevant information out of it. There are times when we are searching for something on YouTube and there are a lot of videos about it and we can't just rely on the number of likes/dislikes a particular video has. More often than not, people like/dislike certain videos randomly or without even watching them at all. The current YouTube algorithm only utilizes the statistics of user engagement while suggesting content. Through the use of the TubeRate app, the user will be able to view ratings of the videos which will be based on sentiment analysis performed on the comments of those videos.

V. Methodology

Our project is mainly divided into six modules:

a) Extracting comments and preprocessing
b) Data gathering
c) Feature extraction
d) Creating ML model and training
e) Sorting based on ratings
f) Integrating in a web application

A. Extracting Comments and Preprocessing

The comments for each of the videos are extracted using YouTube Data API in JSON format, which is then converted into a Pandas DataFrame. This comments dataset is then cleaned and preprocessed to get rid of symbols, emojis and stopwords.

B. Data Gathering

Labeled data (positive/negative/neutral) was needed to train our classifiers using supervised learning. Two labeled datasets of YouTube comments were found and used. The first dataset consists of about 2900 comments from the video "The Boring Company | Tunnels" by the channel The Boring Company [16]. The comments were manually labeled positive, neutral, or negative [17].

The second YouTube dataset was found on Github [18] and features about 1100 YouTube comments. The comments appear to belong to one or several videos from the channel Dude Perfect and some videos featuring news footage. Automatic labeling, the lowest accuracy scenario, was assumed as no information was given about manual or automatic labeling.

Testing data was taken from the Kaggle dataset Trending YouTube Video Statistics and Comments [19], consisting of YouTube videos that were trending in the United States of America and the United Kingdom during a period in 2020. Data includes among other information video likes, video dislikes and comment texts. The data is separated per country. All videos from the UK dataset and only the videos in the US dataset that were not also featured in the UK dataset were used. For videos that were trending several times and thus featuring several times in the dataset, all comments were added along with the last appearing like count and dislike count.

C. Feature Extraction

Feature extraction is an important stage in text mining or SA, and the methods used for extracting the features significantly impact the results. Deep learning models have recently been adopted in the field of SA for learning word embeddings. Word embeddings aim to capture similarities between words and their linguistic connections [20].

D. Creating ML Model and Training

Support Vector Machine was used to train the model with five classes- each representing a rating from 1 to 5. The classifier uses five SVMs since it is a One-to-Rest approach. Each SVM would predict membership in one of these classes. The data used consisted of comments taken from several YouTube videos across different categories, along with their ratings.

E. Sorting Based on Ratings

The most relevant videos against a search term are sorted based on the comments of those videos. Unlike the default order

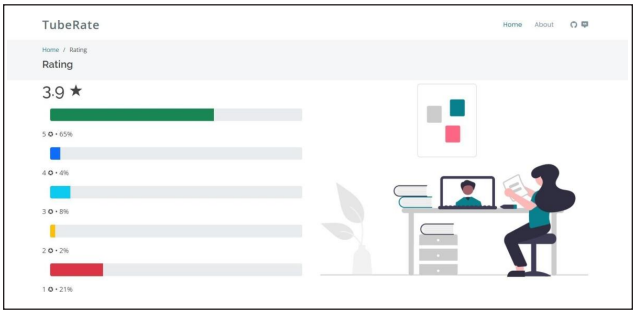
which has videos sorted based on likes, views, hashtags, titles, etc. this sorting is solely based on the ratings predicted by the ML model.

F. Integrating in a Web Application

The web application is built using Flask. It provides the user with the functionality to view ratings of individual videos by entering their URLs as well as have videos sorted based on their ratings by entering relevant search keywords.

VI. Result

The essential data for ordering videos is successfully fetched using API. Top video comments are passed through the ML model to detect sentiments and generate ratings. The final ratings are calculated based on the percentages of these ratings for each video. The sorting employs the final ratings to display relevant videos.



VII. Conclusion

This paper illustrates an automatic process for finding useful video by sentiment analysis of user's comments based on Natural Language Processing (NLP). Our approach evaluated the quality, relevancy and popularity of YouTube videos considering the relationship of user's sentiments expressed in comments. We analyzed a sample of almost 1,000 YouTube comments. Large-scale studies of YouTube video metadata (comment) using the NLP and SentiStrength revealed the importance of user sentiments.

References

[1] V. Sahayak, V.Shete, and A. Pathan, "Sentiment analysis on twitter data," *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, vol. 2, no. 1, Jan. 2018.

[2] E. Momeni, C. Cardie, and M. Ott, "Properties, prediction, and prevalence of useful user-generated comments for descriptive annotation of social media objects," In *Proceedings of ICWSM*, 2013.

[3] B. Pang, L. Lee, and S. K. Vaithyanathan, "Sentiment classification using machine learning techniques," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002.

[4] M. E. Basiri, and A. Kabiri, "HOMPer: A new hybrid system for opinion mining in the Persian language," *J Sci*, vol. 46, no. 1, pp. 101-117, 2020.

[5] S. Chelaru, C. Orellana-Rodriguez, and I. S. Altingovde, "How useful is social feedback for learning to rank YouTube videos?," *World Wide Web*, vol. 17, no. 5, pp. 1-29, 2013.

[6] S. Siersdorfer, S. Chelaru, J. S. Pedro, I. S. Altingovde, and W. Nejdl, "Analyzing and mining comments and comment ratings on the social web," *ACM Transactions on the Web (TWEB)*, vol. 8, no. 3, pp. 1-39, 2014.

[7] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. S. Pedro, "How are your comments? Analyzing and predicting YouTube comments and comment ratings," In *Proceedings of the 19th International Conference on World Wide Web (ACM)*, 2010, pp. 891-900.

[8] S. al Osaimi, and Md. B. Khan, "Sentiment analysis of Arabic tweets using RapidMiner," Dept of Information System, Imam Muhammad ibn Saud Islamic University, KSA.

[9] P. Tripathi, S. K. Vishwakarma, and A. Lala, "Sentiment analysis of English tweet using RapidMiner," In *International Conference on Computational Intelligence and Communication Networks*, 2015, pp. 668-672.

[10] Md. A. Nizar, S. Bukhori, and P. Pandunata, "Sentiment analysis of positive and negative of YouTube comments using Naïve Bayes – Support Vector Machine (NBSVM) classifier," In *2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*, 2019.

[11] W. Yang, and Z. Qian, "Understanding the characteristics of category-specific YouTube videos," Dec. 2011, citeseerx.

[12] R. Singh, and A. Tiwari, "YouTube comments sentiment analysis," In *International Journal of Scientific Research Engineering and Management*, vol. 5, no. 5, 2021.

[13] A. Baravkar, R. Jaiswal, and J. Chhoriya, "Sentimental analysis of YouTube videos," In *International Research Journal of Engineering and Technology*, vol. 7, no. 12, 2020.

[14] H. Bhuiyan, J. Ara, R. Bardhan, and Md. R. Islam, "Retrieving YouTube video by sentiment analysis on user comment," In *International Conference on Signal and Image Processing Applications*, 2017.

[15] K. M. Kavitha, A. Shetty, B. Abreo, A. D'Souza, and A. Kondana, "Analysis and classification of user comments on YouTube videos," In *International Workshop on Artificial Intelligence for Natural Language Processing*, 2020.

[16] The Boring Company, "The Boring Company | Tunnels," Apr. 2020. Accessed: Apr. 7, 2021. [Online]. Available: https://www.youtube.com/watch?v=u5V_VzRrSBI

[17] K. Eklekta, "Relevance and sentiment of YouTube comments of a vision video," Version version 1, Zenodo. Feb. 2021, doi:10.5281/zenodo.4533302.

[18] K. Verma, "Labeled comments," 2020. Accessed: Apr. 1, 2021. [Online]. Available: https://github.com/pranaykmr/YouTubeSentimentAnalysisandPredictingTrends/tree/master/data

[19] M. Jolly, "Trending YouTube video statistics and comments," Oct. 2020. Accessed: Apr. 3, 2021. [Online]. Available: https://www.kaggle.com/datasnaek/youtube

[20] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics*, vol. 9, no. 3, 2020.