Pakistan Journal of Engineering and Technology, PakJET
Multidisciplinary | Peer Reviewed | Open Access
Volume: SI, Number: 01, Pages: 183- 188, Year: 2020

Multi-Class Classification of the YouTube Comments using Machine Learning

Shoaib Nawaz[1], Muhammad Rizwan[2], Samina Yasin[3], Mehtab Ahmed[1], and Umar Farooq[1]

[1]The Islamia University of Bahawalpur, Rahim Yar Khan campus, Pakistan
[2]Khwaja Freed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan
[3]University of Agricultural Faisalabad, Pakistan
Corresponding Author: Shoaib Nawaz (Email: shoaib.03339770257@gmail.com)

Abstract: Due to huge data on Social media the people face difficulty to finding in qualitative content of the video if they find the qualitative content as per their judgment and knowledge, they do not confirm the actual content quality. People put their idea and subscriber watch the videos and put their feedback in the comments. Our purposed study help the subscribers to find the best-experienced idea based on the people personal experience here we classify the collection of comments collected using Google API's and annotate them in different classes which are Experience-positive, Experience-negative, Warning, Suggestions, Questions, Praise, these classes helps us to find the qualitative analysis of the comments but based on the these required classes. Our proposed model shows experiments that the best classifier for us is the SVM have accuracy is 89.18% and F1 score is 0.89 this shows that our model is productive and efficient for classification to help out the user as well as the author to find the best qualitative video content which is experienced and confirmed by the user's experience.

Keywords: *Google APIs, Social Media, Annotation, Machine Learning, SVM.*

I. INTRODUCTION

People use social media for text, image, video sharing for this there is a lot of plate forms are available i.e. YouTube, Facebook, Instagram, and many websites and application are viable for these people to share the ideas and experiences [1]. In the video-sharing Social media websites the big one is YouTube. Two billion users of YouTube daily upload five-hour videos in one second. YouTube is the video-sharing website and also used for entertainment, learning, website people upload videos and create channels and start their sharing idea, tips video, some are medications, some are sharing the home remedies and some are teaching something. YouTube has strong statistics. Today is the world of finding smart and productive and effective solutions [2-3]. People have no time to waste searching for their desired knowledge or solution on the internet.

There are a lot of sites and contents are available for your problem or tip or knowledge or information but some information is not exactly match and sometimes the information is fake and actually not exists. YouTube helps the people to get their information from the YouTube videos because the video is so easy to learn as compare to text to read and learn so important contribution of YouTube in our life. Almost 7 out of 10 find their solution for a problem on the YouTube but by experience [4-6], you will find the searched data which is displayed by the Google ranking algorithm some video contents are productive and some are not and the people are trying the each ranked video and then after a time they will get their required video after a but when they try the method or tip or the cure about some described in the video after an experiment they realize that the said content is not correct and they response the comment with their experience as we said in our work as Experience-Negative(Because they learn from their experience) and some get the Experience-positive and some are given their some suggestions and some are just praising the video content and some ask some Question and some is a comment that we will try this in future in our approach. We named our classes here are Experience-Positive, Experience-Negative, Parise, Suggestions, Questions, Future-Try. We classify the video comments and get the classification of these six classes this will give us the information in the percentage of the people having Experience-Positive and how much the people having the percentage Experience-Negative same as the percentage of Questions, Percentage of Praise, Percentage of the Future-Try and Percentage of the Suggestions. Our focus on these classes will give us the actual picture of how much the qualitative video content on YouTube or other videos on the social media web site.

a) ABOUT YOUTUBE

YouTube [7] was started in 2005 and now [4-8] YouTube have 2 Billion registered users and the Watching of the Study, professional skills, Medication, and learning related a lot of videos on the YouTube you will find and this is increased with 3time per year and 2.5X times increased are see in last two years in the cooking recipes videos the YouTube user age is in between 18-34 year old and 100 countries and more than 80 langue's are used on the YouTube and the one billion watch time on daily basis. These statics describe the YouTube is how much it is important in our lives and how much contribution in our life to solve problems and get the best solution.5 hours video duration are uploaded in the one second on each day YouTube is the product of the Google and its account based on the Gmail account when you sign up in YouTube you can save the playlist, create the channel, Upload videos and also you can download any video even you are not the user of the YouTube. YouTube was started in 2005 the Jawed Kareem is the founder of the YouTube and the first video upload on the YouTube is about the visit in the Zoo was viewed by 2 million peoples this was the start of the YouTube now the YouTube is the best platform for video watching, sharing, uploading and learning and getting the solution for their problems.

It's having a simple interface with one Search Box and the right side is your saved playlists and the right side is next video option and the center is displaying the current video and the bottom of the video is a comments section in which people express their feelings and responses and the comments they also can reply on each comment.

YouTube also had some other aspects of response for YouTube video such aspects are Likes, Dislikes, Subscriber, Viewers, are the basic aspects of the video these aspects also help Google to rank the video [9-12].

Even these aspects cannot describe that how much the truth and quality of the content this will you can get from YouTube comments and read the comments and after reading the comments you will be able to judge the how much the truth and quality of the video content. This is impossible for us to read the billions of comments and so need the machine to read the comments and tell us about some specific information this information will help us to fallow the ideas of video content or ignore them or try them. Here we use the [7-10] Google APIs for extraction the comments from the YouTube and save the files in CSV (comma separated values) and perform the preprocessing on the comments and after this, we annotate the comments as per classes, Experience-Positive in which the people experimented and then feedback on the same video we experienced good and it works and Experience-Positive in contrast some are Experience-Negative when they did some learn from the video content and after the physical experiment they found that the video content ideas are not works and not good so we label this sentence as Experience-Negative. Some said that the said tip about something or talk is so the people face some issue and post the comments in the warning category and some are giving some suggestions and some are asking questions [9-15].

The YouTube statistics describe that the usage of the YouTube it describes that the people having mostly use the YouTube for their problems and the Speaker need to know what the people talk about and what most talks about and the based on our information the speaker make some improved version of the video in the next sessions. This will improve productivity and make the channel productive and progressive [10-18].

II RELATED WORK

There are a lot of tools and techniques are having been used for the YouTube comments analysis people used pythons, R with different text mining packages. The sentiment analysis of the YouTube video comments is not enough for the qualitative analysis of the video content people like to [15-20] study deep inside information extraction topic modeling and subtopic modeling this extraction will describe the more about the user feedback and response about the video content and this study provide a topic modeling technique in which we decide a topic and based on the keyword like “Dance” the match comments are filtered and after the matched comments filter the subtopic and then remove the duplicate comments and then filter the langue this kind of topic are worked called CTFC method [15-19].

In the Data Science, there are two ways to classify the data in defined classes of the YouTube comments before the defined user classes people are only working on the feedback in term of what the people having feelings about the YouTube video content these contents these feedbacks are divided into three categories [17-21] 1.Good or Positive means people like the content 2.Not Good or Negative feedback it’s not the liked video content and some people are giving the non-related feedback this is called the Neutral feedback means people are not like and not dislike video content but they just put their feedback not related with the video content in the data science this kind of feedback is called sentiment analysis.

2015 according to cisco the 90% of the internet traffic is covered by video data means the video-sharing watching are most frequent activity performed on the internet due to this huge traffic how the people find their required video from the Sina Weibo [22] a social networking website most popular website in China the study describes that the video recommendation processes its most difficult processes to recommend the video content to the user and a factor in the social websites the interaction between the users i.e. the one user is the user who shares the maximum time of vides with the community and one user to another user this kind of sharing is mostly neglecting by the recommendation algorithms [16-17] study invent the algorithm which is based on the user discovery model (UDM) and video discovery model (VDM) means which video is best shared and which user mostly shared a video which is mostly shared on the social network this kind of user is called influential user some time it's shared by the trust.

Feedback is not only used in the social network like [13] Facebook, [14] YouTube, [15] Instagram, Blogs, Forums, Discussion websites as well as in News websites this help and support the researchers to find the productive comments to ensure the event (Post, news, video) are how much good, not good, or the people have not interested. This feedbacks also help the research on what platform is what produces what kind of data.

In YouTube, the data science help to find the more deep [15-20] study of the comments in this the comments are happy, angry, disgusting, fearing, sadness based and maybe the feedback is surprising this shell be finds by using supervised and unsupervised learning methods in the text mining with features extraction and the also the Machine learning helps us to classify the comments in the described classes or feedbacks and the Machine learning algorithms in which Naïve Bayes have the F1 Score is 0.682 and Decision tree is 0.782 and the Support Vector machine have the F1 score is 0.815 so the SVM help the researcher in the study having most good results.

The people on YouTube mostly working on the ranking of the YouTube that is based on different parameters and aspects some are like, dislike and number of viewers, subscribers, etc.

These aspects did not represent that the video content is productive and true or fake or just words so the people read the comments in the below section of the video and then after reading the comments they feel what they described in the video content it's not good or it's good or its warning don't use this method such kind of activity will help the user to get the solution form the YouTube that's why the people use the YouTube. Most researches use the technique of sentiment analysis to analyze the comments and the base of the sentiment they describe that the video content is good or not just based on the three feelings shared in the comments like positive comment, negative comment, neutral comment, this approach is so good on some extend but the next people use the sentiment score to use the percentage of the positive comments and percentage of the negative comments and then make produce the [15] recommendation system.

This system describes that the people how to comment positive and the average score of the comments is if greater than 75% than the video content is highly recommended and if the parentage is less than 60% this will be the video content is labeled as recommended and if the percentage of the score is less than 50% the video content is may be recommended and the below 40% is the not recommended this approach help us to get the actual feeling about the video content. The sentiment analysis (Natural Language Processing) NLP helps in a way the score is different and the sentence is the same category for example, this is a good video, see the table. Here we see that all sentences are positive but the score is different on this difference we just get the percentage of the positive sentiment score and get the average sentiment score and then the finally average sentiment score of both approaches 1. Average sentiment score, 2. Average Sentiment score if greater than 50%, finally labeled based on the Average score.

There are extrema video classification is also used in which the YouTube-8M [14] dataset they use deep neural network and solve the 8M videos dataset to classify in 4,716 classes and the model was recorded accuracy 83.90% this was the Kaggle project for competition and the position for this result was 8th.There are different approaches used for the comments classification. Seven Out of 10 are using the YouTube for learning purpose so the people fined the productive videos people see the reviews and try to assess the quality of the video content people find the good cooking [4] recipes collected 20,000 comments of the different cooking recipes and then performing the cleaning the data and first find the actual feedback about the video content and find the productive comments by using Machine learning algorithm the Support vector machine produces an accuracy of F1 score 0.93 and accuracy 83% with using features extraction method TFIDF (Term-Frequency and Invers-Term Frequency).

People are mostly searching to find the informative comments related to the video content they are collected the video form the [8] TED Talk the dataset contains the 1816 videos and 380619 comments and they analyze the people are most likely to search and find them informative videos and the videos which are having the relevant informative comments in this regard the experiments are performed on the 20 videos they performed experiments and finds the comments are mainly based on two parts one is questionable comments and other one are answerable comments and then find the informative comments on the TED Talk videos.

III DATA COLLECTION

We have collected the data from YouTube using Google API's. The details are given in Fig. 1.

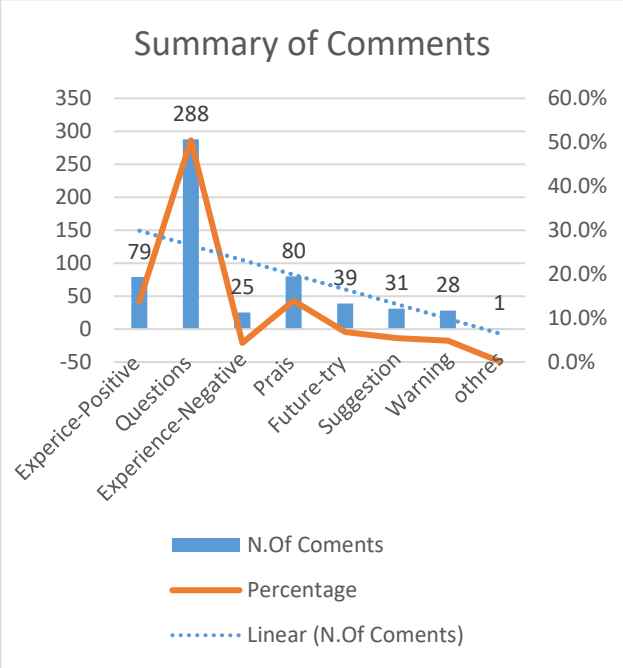

FIGURE 1: Data collection

First, we search videos with different category mostly are related tips, guide and instructive type of videos and collection of video comments we annotate them the dataset values labeled six classes we are trying to get the most required information labels/class like, ***Experience-Positive*** which means the people see the video and do the tip and get an experience that the tip or guide or suggestion is good and true, another type of class if ***Experience-Negative*** which means that the content is not true and the people experience them and relies upon that the tip and guide and suggestion is not true and the video content is just fake. Another class is ***Praise*** this means people not try the tip and guide but they just said is good not the actual good and just praise and next one is ***Suggestions*** some proposed the new related suggestions and gives the people to try them this class represent this psychology, some people are very careful and not to experience the curing tip or suggestion they have more information about the guide, tip, suggestion so they ask more ***questions*** about the content and ask the question if we change something like related ingredients or alternatives and some more good for that tip or idea so our target is also the people how much the percentage the people having the questions about the video content. Some peoples are said we will try this but not now or they actually said the tip is looking nice but we will try in the future we assign this kind of information

in the comments as ***Future-Try*** this means people also agree with the idea but they will do in the future. Some people are sensitive and when they try the tip or the suggestion and idea they experienced that the tip is dangerous when they try they faced issues and post the comments don't apply this in real, or it's dangerous so this kind for information we have actually is ***Warring*** for the people if you have the same issue with you don't do this so categories this in Warring. Our annotated data

details are shown in Fig. 2.

FIGURE 2: Annotated data detail.

Here we see that the [3] word Cloud described that the data set is mostly talking about the Honey and Coffee and they mostly talk about the use and works shows that's the use and works is about the honey or coffee for the face and skin the big one another word is a lemon is also showing here to describe that the lemon is also used and works for the skincare [5].

IV METHODOLOGY

First, we have to search simple search the videos on the YouTube and then see the results from the YouTube ranked videos we randomly select the video and then using the Google API's we extract the comments of the videos and we collect the more than 10 videos comments (2300) data in CSV(Comma Separated Values) format and all the CVS file are combined in the single file which is the main file in which we store all download comments and then make start annotation and then we annotate all the comments and after performing the preprocessing In the preprocessing on the data and then we remove non-English words and then we perform lemmatization remove the video with no comments and if the video has less than 10 comments we also exclude them. After cleaning the dataset. We get the all CSV Files and combined them all and the different peoples annotate the data in the seven classes which are *Experience-positive, Question, Experience-Negative, Praise, Future-try, Suggestions, Warnings* and after the annotation we have to extract features. we use the Machine Learning Algorithms which are SVM 79% and F1 Score is 0.94, Naïve Bayes 89.18% and F1 score is 0.94 with 10-fold cross-validation is used for training dataset we partition the data with 80/20 partition 80% for training and 20% for training this ML Model will help us to get the multi-class classification in which we can focus on the *Experience-positive* class this classified comments product very productive information about the video content specially video uploader will focus such information what are actual impact and how much results about the content which are experienced by the people to more confirm about the video content as this there are other class like *warning* is also most important people will what are the actual issue people are facing by doing as described in the video content. Same as all classes are more important as per their perspective [4].

a) EXPERIMENTS:

We use the python framework Beautiful Soup for comments scrapping also we use the Google Chrome Extension instant data scrapper and scrap the comments of the video and combined them in CSV format and we remove the other than English of all language's comments. We use the Scikit Learn to perform the preprocessing (cleaning data) and we manually annotate the all dataset with classes, *Question, Warning, Experience-positive, Experience-negative, Suggestion, Future-try.* We convert the all classes into [7] numeric format using the label Encoder of *SCIKITLEARN* to encode the classes for the batter and fast processing or learning for machine learning for the comments we use the vectorization methods we use the TIIDF vectorization [16] and use the standard scaler for the comment column and then divide the data into 80% for training and 20% for test the model and we apply Naïve Bayes and Support Vector Machine learning algorithms and when the experiment completed we found that the Support Vector Machine has performed the batter model instead of Naïve Bayes. The Support Vector Machine shows the results of 89.18% Accuracy and F1 0.89. Fig 0.3 shows our working model and this will clearly show that we have corpus or the collection of the comments and after performing the preprocessing the corpus will forward to the machine for learning purpose the ML will build a model and gives us the distribution if the classes based on the support vector machine learning algorithm [17].

Our methodology is represented Fig. 3, the corpus is dataset collection of review / comments and deploy the machine learning algorithm on the corpus and classify the classes and improve the model accuracy. The proposed model target is to classify the classes and find the percentage of each class as shown in the Fig. 3.

Means how much the percentage of the class exists in the video comments and based on that comments we will easily find the quality of the video content from reviews. Details DFD of the methodology shown in the Fig. 4.

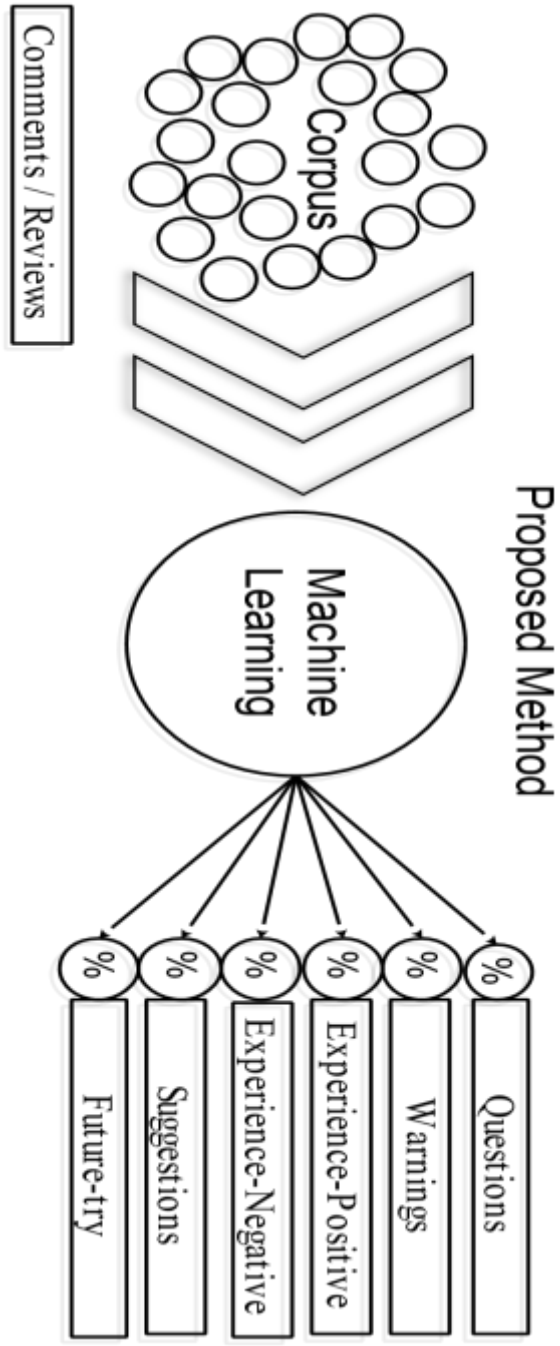

FIGURE 3: Proposed model.

b) PROCESSES MODEL

Here you see the processes mode as shown in Fig. 3. Our methodology is about the classification of the comments but in the effective way our corpus is collected from the YouTube and the deploy the algorithms helps us to classify the comments and find the user experience how the user perform experiments and based on their experience we find the quality in the video content. After deploying the model helps the user and as well as author to get the actual information about the idea discussed in the video and finally, we found the percentage of each class to measure the data in normalized form [10].

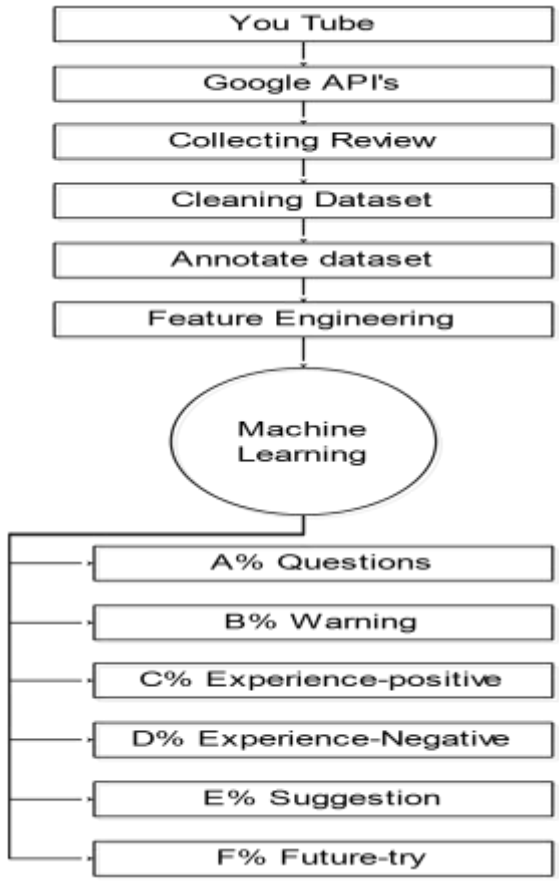

FIGURE 4: Proposed model DFD.

V RESULTS AND DISCUSSIONS:

Algorithm	Accuracy	F1	Precision	Recall
SVM	89.19%	0.89	0.89	0.89
NB	76%	0.86	0.76	1.00

Table I: Algorithms results and accuracy.

The results from our two algorithms given in Tab. I, but the Support Vector Machine perform better than Naïve Bayes. The Experience-positive class noted that having the F1 score 0.94 comparatively with other classes and the praise F1 score is 0.89 and the other classes average results is 0.89 as we know the SVM , NB both are best in binary classification and in the binary classification the results more batter than multi-classification.

VI CONCLUSION

As per our work and experiment the Machin learning helps us to find the actual effective and productive qualitative video content and based on our model accuracy (89.18%) this will beneficial for doctors if they want to check their idea are productive or not same like skin care experts, teachers, and technical experts and many more can get benefits from idea. Our classes Experience-positive and Experience-negative , praise , warning , Suggestion, Future-Try , each class has its own features and each class is not only the one can focus its depends on the dataset or content in which what is your focus class either Experience-positive , Experience-negative or

other classes have importance as per their features related to the video content. The results show the accuracy which means the results will be so productive for video author or content uploader

We majorly focus Experience Positive, Questions, praise these are the important classes and our focus was on these but the other classes Experience-negative, Warning, Future try, Suggestions we also discussed and included in the dataset and experiences shows that the most people watch the video and more than 50% are not exactly believe on the video content they ask more and more question to confirm the idea will work or not others classes show that the Experience-positive, praise is the almost same ratio in the whole dataset this will help us to ensure the authenticity of the video content. This will also help the author to confirm when the idea will be implemented practically (Experience) what will the proximately results are received from video on that bases the author will improve his idea make more research to find the best solution of a problem. This will reduce the fake content and the actually productive and effective content will be promoted.

VII FUTURE WORK

We next work is a more in-depth research on how the people said we have experience-positive and Experience-negative and more important is on what bases the people said warring and find the specific point to find out the reasons behind these classes. We also improve the accuracy of the model.

REFERENCE:

[1] Abbas, Syed Manzar. 2017. "Improved Context-Aware YouTube Recommender System with User Feedback Analysis." *Bahria University Journal of Information & Communication Technologies (BUJICT)* 10 (2).

[2] Agarwal, Neha, Rajat Gupta, Sandeep Kumar Singh, and Vikas Saxena. 2017. "Metadata Based Multi-Labelling of YouTube Videos." In *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*, 586–590. IEEE.

[3] Asghar, Muhammad Zubair, Shakeel Ahmad, Afsana Marwat, and Fazal Masud Kundi. 2015. "Sentiment Analysis on Youtube: A Brief Survey." *ArXiv Preprint ArXiv:1511.09142*.

[4] Benkhelifa, Randa, and Fatima Zohra Laallam. 2018. "Opinion Extraction and Classification of Real-Time Youtube Cooking Recipes Comments." In *International Conference on Advanced Machine Learning Technologies and Applications*, 395–404. Springer.

[5] Chang, Wei-Lun, Li-Ming Chen, and Alexey Verkholantsev. 2019. "Revisiting Online Video Popularity: A Sentimental Analysis." *Cybernetics and Systems*, 1–15.

[6] Chauhan, Ganpat Singh, and Yogesh Kumar Meena. 2019. "YouTube Video Ranking by Aspect-Based Sentiment Analysis on User Feedback." In *Soft Computing and Signal Processing*, edited by Jiacun Wang, G. Ram Mohana Reddy, V. Kamakshi Prasad, and V. Sivakumar Reddy, 900:63–71. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-3600-3_6.

[7] Chen, Yen-Liang, Chia-Ling Chang, and Chin-Sheng Yeh. 2017. "Emotion Classification of YouTube Videos." *Decision Support Systems* 101: 40–50.

[8] Choi, Seungwoo, and Aviv Segev. 2020. "Finding Informative Comments for Video Viewing." *SN Computer Science* 1 (1): 47.

[9] Cui, Laizhong, Lili Sun, Xianghua Fu, Nan Lu, and Guanjing Zhang. 2017. "Exploring a Trust Based Recommendation Approach for Videos in Online Social Network." *Journal of Signal Processing Systems* 86 (2–3): 207–219.

[10] Cui, Zaixu, and Gaolang Gong. 2018. "The Effect of Machine Learning Regression Algorithms and Sample Size on Individualized Behavioral Prediction with Functional Connectivity Features." *Neuroimage* 178: 622–637.

[11] "Google API YouTube." n.d. Accessed April 14, 2020. https://developers.google.com/youtube/v3.

[12] Huang, Sheng-Jun, Wei Gao, and Zhi-Hua Zhou. 2018. "Fast Multi-Instance Multi-Label Learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[13] Iram, Ateya. 2018. "Sentiment Analysis of Student's Facebook Posts." In *International Conference on Intelligent Technologies and Applications*, 86–97. Springer.

[14] Na, Seil, Youngjae Yu, Sangho Lee, Jisung Kim, and Gunhee Kim. 2017. "Encoding Video and Label Priors for Multi-Label Video Classification on Youtube-8M Dataset." *ArXiv Preprint ArXiv:1706.07960*.

[15] Nawaz, S., M. Rizwan, and M. Rafiq. 2019. "RECOMMENDATION OF EFFECTIVENESS OF YOUTUBE VIDEO CONTENTS BY QUALITATIVE SENTIMENT ANALYSIS OF ITS COMMENTS AND REPLIES." *Pakistan Journal of Science* 71 (4 Suppl): 91–97.

[16] Oramas, Sergio, Oriol Nieto, Francesco Barbieri, and Xavier Serra. 2017. "Multi-Label Music Genre Classification from Audio, Text, and Images Using Deep Features." *ArXiv Preprint ArXiv:1707.04916*.

[17] Pereira, Rafael B., Alexandre Plastino, Bianca Zadrozny, and Luiz HC Merschmann. 2018. "Categorizing Feature Selection Methods for Multi-Label Classification." *Artificial Intelligence Review* 49 (1): 57–78.

[18] Samuel, Nardin, Naif M. Alotaibi, and Andres M. Lozano. 2017. "YouTube as a Source of Information on Neurosurgery." *World Neurosurgery* 105: 394–398.

[19] Thelwall, Mike. 2018. "Social Media Analytics for YouTube Comments: Potential and Limitations." *International Journal of Social Research Methodology* 21 (3): 303–316.

[20] "YouTube Data API." n.d. Google Developers. Accessed April 14, 2020. https://developers.google.com/youtube/v3.

[21] Zhan, Ming, Ruibo Tu, and Qin Yu. 2018. "Understanding Readers: Conducting Sentiment Analysis of Instagram Captions." In *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence*, 33–40.

[22] "微博-随时随地发现新鲜事." n.d. Accessed April 14, 2020. https://www.weibo.com/login.php.