OPEN ACCESS Freely available online

Journal of Remote Sensing & GIS

ISSN: 2469-4134

Research Article

Classification of YouTube data based on Opinion Mining

MUSKAN VERMA,RIYA GARG, RITIKA JAIN*, ANURADHA TALUJA

Department of Computer Science Engineering, Meerut Institute of Engineering and Technology, Meerut, U.P, India

ABSTRACT

Opinion Mining is a matter of great concern and an emerging research field of omnipresentcomputing. Nowadays, people find it very difficult to find the exact content which theyhave searched for due to the enormous amount of content found on google, YouTube ,Unacadamy and Udemy. Using Machine Learning one can make the exact search and getthe best suit or match of the search without spending a lot of time which eventually leads to economizing time. Proposed scheme is more efficient in terms of computation than other traditional schemes.Sentiment analysis systems are being applied in almost every business and socialdomain. It is far beyond just the number , number of likes/comments/shares. It is primarily used to evaluate or analyse the commenter's opinions , sentiments, emotions and attitude from a written piece/comment on a particular post/video. It will fundamentally help in saving a lot of time as the seeker can directly get the overall opinion based on the comments that whether he should invest his time in watching the particular video or go for the other one.

INTRODUCTION

There has been an exponential growth in the use of online resources & there are lot of channels available on youtube for a particular thing or lot of material available on different sites for a particular thing which ultimately makes it challenging for people to choose which one is better for them, for example, students find difficulty to find better education platform for them or the best youtube channel for a particular subject or for a particular topic.

A person to search a lot to find a simple thing on the internet nowadays due to the presence of large amounts of data over the internet and it becomes very tedious to find the best platform to enhance skills and causes a lot of wastage of time. The only way a person is left with is just surfing and searching on the internet & then to subscribe to the channel he/ she finds the most appropriate. It requires a lot of time & as a result users get deprived from a lot of good channels on youtube or websites available over the internet. That's why Opinion Mining comes into picture to provide the solution of existing problems.

Opinion mining refers to the use of natural language processing, text analysis andcomputational linguistics to identify and extract subjective information from the written source material. This project will easily give users the overall review of the content whether it is positive, negative or neutral. This can be used for various websites like Youtube, Unacademy etc.This approach of opinion Mining saves a lot of time. It helps others to find better choices of channels on Youtube Or Unacademy in just a few minutes of search for their betterment & enhancement.

LITERATURE REVIEW

A lot of work in this field done already but because of increasing usage of online resources few more effort needs to put in. Now a days, there are lot of youtubers on social site youtube but because of number of different channels everyone is so confused to follow which particular channel or video. If a user is able to see number of positive or negative reviews for a particular video with the help of proposed system then it would be great for every individual.

Every individual will able to analyze everyone's prospective with the help of given comments by them on a video.

Mostly work is done in this field using Naïve Bayes classifier but in this proposed system, Decision Tree classifier is used to test or train the data.

*Correspondence to: **Correspondence to:** Jain R, Department of Computer Science Engineering, Meerut Institute of Engineering and Technology, Meerut, U.P, India; Tel No: 8630752405; E-Mail: ritika.jain.cse.2017@miet.ac.in

Received date: April 29, 2021; **Accepted date:** September 03, 2021; **Published date:** September 13, 2021

Citation: Jain R (2021) Classification of YouTube data based on Opinion Mining.J Remote Sens GIS, 10:110

PROPOSED APPROACH

This paper suggests a framework that shows how the comments are extracted from the platform, how the classification of the data is done and how the opinion analysis is done based on the classification of data.

Framework consists of different steps:-

- YouTube API.
- Text/Comments Extraction.
- Clustering of Data.
- Language Detection.
- Text Processing.
- Feature Extraction.
- Classification of Comments.
- Opinion Analysis.

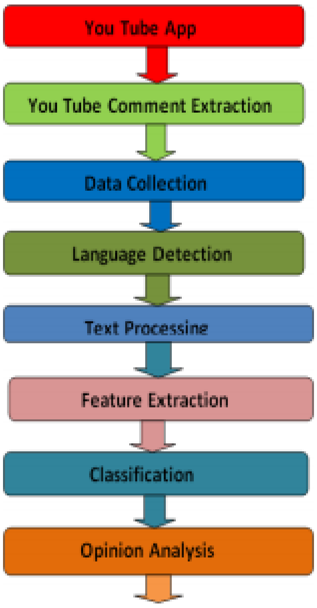


Figure1. Proposed Framework.

YouTube API

The very first step is to interact with YouTube API. For the sake of interaction with YouTube or to abstract the data from YouTube. This application is developed on YouTube developer site.

Text/Comment Extraction

As soon as the application is created the proposed system will extract the comments from YouTube. The comments are in the form of the text.

Clustering of Data

Now, the comments which include the textual data will be extracted. The textual data are the comments or keywords entered by the commenter in the comments such as "good", "bad" and so on. The table1 given below shows the description of data which contains some comments used in the research paper.

Language Detection

Language Detection is a mechanism which helps in recognizing the language of the text. In this research paper, Analysis has been performed on English language . The language of the comments has been detected on the basis of the character set of English language. Every language has its own unique set of letters which helps to recognize that language.

Text Processing

The data that has been extracted from YouTube is first in an asymmetrical structure containing both beneficial and non-beneficial text. So, before analysing that data there is a need to remove those additional texts and extract the beneficial data. Text processing involves the removal of extra spaces, special characters etc from the comments.

Feature Extraction

It is the most important step in which the data is shifted out. In other words, it is the reprocessing of data which gives essential features from the gathered data. Filtration of the gathered data is done by removing the stop words or the words which are not necessary.

Classification

It is the step where extracted features are classified into positive, negative or neutral depending upon its polarity.

Figure 2: Extracted features are classified into positive, negative or neutral depending upon its polarity.

Opinion Analysis

If polarity is greater than 0, it will be considered as a positive comment. If polarity is less than 0, it will be considered as a negative comment. If polarity equals to zero, it will be considered as neutral.

Steps/Procedure involved in the implementation:-

- We have to extract dynamic data (comments) from YouTube of a particular video to analyze its overall feedback by its users. To extract the data from any API, it requires permission. This is possible with the help of a developer key.
- For dynamic extraction of data from any API using python then this could be possible
- with the help of selenium webdrivers.
- With the help of python script, it will open youtube where the chrome is controlled by automated test software.

Python library called "**TextBlob**" is used for analyzing the sentiments of comments into positive, negative or neutral.

Textblob is an extremely powerful NLP(Natural Language Processing) library. It is used for processing the textual data. It returns two properties named as polarity & subjectivity. Polarity lies between [-1,1] where -1 represents negative comment, 0 represents neutral comment & 1 represents positive comment. Subjectivity lies between [0,1]. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information.

- Then a .csv file is created which consists of comments with their polarity (Positive, Negative or Neutral)& subjectivity. This will help the user in analyzing that particular video on youtube having how much positive reviews or negative reviews. By analyzing the number of positive & negative reviews on a particular video, the user can identify which video will best suit them. After knowing the count of positive comments or negative comments, users can either prefer the particular video or not.

EXPERIMENTAL RESULTS

This section firstly contains a table named Data Description which consists of a set of words which have been used in our project to evaluate the polarity of the comment.

Figure 3: Classification of comments depending upon its polarity.

Data Description	Positive	Negative	Neutral
Text Data	Amazing, Excellent, Lovely, Awesome, Success	Poor, Disrespect, Hate, Bad, Angry	Calm, Sensitive

TABLE 1: Data Description

The shown below is the required dataset of comments. These comments are of a particular movie named as "Sadak 2". It is considered because this movie comprises of negative as well as positive feedbacks from people. This dataset is extracted dynamically from Youtube which consists of comments, polarity, sentiment type & subjectivity.

	Comments	Polarity	Sentiment type	Subjectivity
0	Who is after here 13M disliked	Negative	-0.200000	0.600000
1	Who's here just to see how many dislikes lol	Neutral	0.650000	0.600000
2	Who searched this trailer just for dislike?	Positive	0.000000	0.000000
3	Now the son of the king will not be the king, The king will become the brave dog	Neutral	0.000000	0.000000
4	I created 10 google accounts for disliking 10 times	Neutral	0.000000	0.000000
5	Sushant Singh fan from Tamil Nadu	Neutral	0.000000	0.000000
6	Alia do u know who Sushant is now ? Welcome to reality	Positive	0.800000	0.900000
7	It's not a big deal disliking this trailer. But when the movie releases, we should not watch it to prove that we have the power. My humble request to people, please do not watch this movie in hotstar forever.	Negative	-0.100000	0.250000
8	I don't know hindi but just for sushant A proud dislike from tamilnadu	Positive	0.800000	1.000000

9	Alia asked who is Sushant in koffee with karan!!! Now she will never forget his name !!!	Neutral	0.000000	0.000000
10	Fun fact: we all searched this trailer just to dislike and report	Positive	0.300000	0.200000
11	O God how many dislikes LITERALLY CAN'T stop laughing	Positive	0.500000	0.500000
12	Most dislike trailer in the world this is a new record for dislike	Positive	0.318182	0.477273
13	Target 15M dislikes. for SSR soul rest in peace.Who is with me comment.	Neutral	0.000000	0.000000
14	The trailer actually looks crap. Deserve a dislike by itself!	Negative	-0.500000	-0.450000
15	Sushant Singh Rajput's Fan Like	Positive	0.600000	0.900000

TABLE 2: Dataset of Top 15 comments of Movie Sadak

The shown below is the visualization of comments which shows that the number of positive & neutral comments is more than the negative comments. For analyzing purposes, 80 comments of a movie are extracted from youtube.

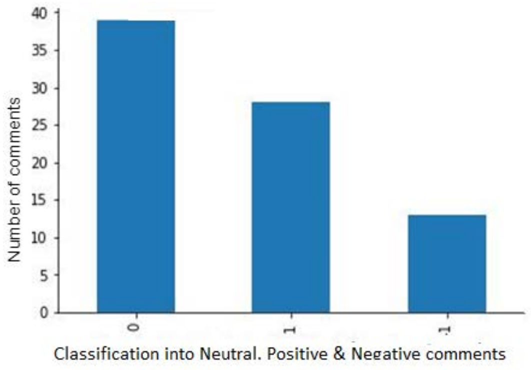
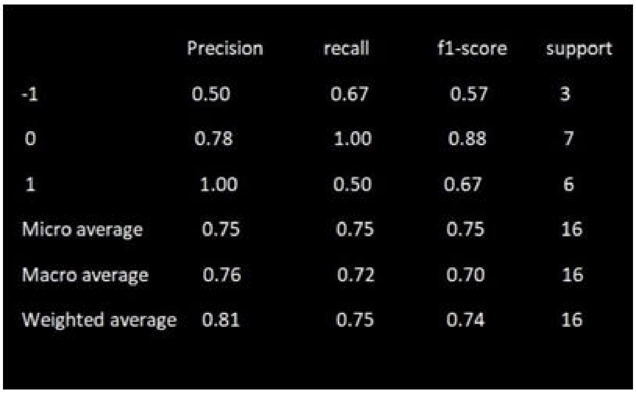


Figure 4: Plot shows no. of positive, negative & neutral comments

Decision Tree Algorithm is supervised learning Algorithm. After applying Decision Tree Algorithm on dataset of 80 comments, below shown classification report is generated:
Accuracy_Score = 0.75

Classification Report:



	Precision	recall	f1-score	support
-1	0.50	0.67	0.57	3
0	0.78	1.00	0.88	7
1	1.00	0.50	0.67	6
Micro average	0.75	0.75	0.75	16
Macro average	0.76	0.72	0.70	16
Weighted average	0.81	0.75	0.74	16

Figure 5: After applying Decision Tree Algorithm on dataset of 80 comments.

CONCLUSION

Sentiment Analysis is an analysis of opinion, sentiments or emotions. Sentiments can be classified into positive, negative or neutral. If a video has more positive comments than the other video on youtube, then it will be preferred to the user which will be the best option for a user to find out which is more suitable for him/her from an enormous amount of videos over the internet today.

REFERENCES

1. Joachims,T.A probabilistic Analysis of the Rocchio Algorithms with TFIDF for Text Categorization. In proceeding of the Fourteenth international conference on Machine Learning D.H. Fisher, Ed. Morgan Kaufmann Publishers, San Francisco, CA, , 1997,143-151.
2. Gamgarn Somprasertsri, Pattarachai Lalitrojwong , Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization, Journal of Universal Computer Science, vol. 16, no. 6 (2010), 938-955.
3. Hu, and Liu, "Opinion extraction and summarization on the web", AAAI., (2006), pp. 1621-1624.

4. Jia, C. Yu, andW. Meng , “The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness”, In Proceedings of CIKM ,2009

5. Weitong Huang, Yu Zhao, Shiqiang Yang, Yuchang Lu, “Analysis of the user behavior and opinion classification based on the BBS” , Applied Mathematics and Computation 205 (2008) 668–676 .

6. Dave, D., Lawrence, A., and Pennock, D. Mining the Peanut Gallery: Opinion Extraction

7. and Semantic Classification of Product Reviews. Proceedings of International World Wide Web Conference (WWW’03), 2003.

8. Ding, X., Liu, B. and Yu, P. A Holistic Lexicon-Based Approach to Opinion Mining. Proceedings of the first ACM International Conference on Web search and Data Mining

9. (WSDM’08), 2008.

10. LLee, and Pang Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, 2008, pages 1-135.

11. Bing Liu.2011.Sentiment Analysis Tutorial - Given at AAAI-2011, San Francisco, USA.

12. R. Prabowo and M.Thelwall, Sentiment Analysis: A combined Approach, Journal of Informetrics, 2009, pages 143-157.

13. Bing Liu. 2010. Sentiment Analysis: A Multi-Faceted Problem.

14. B.Ohana, B. Tierney.2011.Opinion Mining with SentiWordNet.