

Youtube Comment Analyzer

Mohammed Arsalan Khan^{1*}, Sumit Baraskar², Anshul Garg³, Shineyu Khanna⁴,
Asha M. Pawar⁵

^{1,2,3,4,5}Dept. of Computer Science, MIT-School of Engineering, Maharashtra Institute of Technology, Pune, India

*Corresponding author: sonukhanroxxx@gmail.com Tel: +919867697881

Available online at: www.isroset.org

Received: 15/Jun/2021, Accepted: 30/Jul/2021, Online: 31/Aug/2021

Abstract— The goal of this study article is to assist a content creator or anyone who wants to know about the audience's thoughts, emotions, on a particular video. We have studied the literature papers first and then identified the basic functionality of it and then we get to know about its dimensions from the paper.

Keywords— Sentiment Analysis; YouTube comment

I. INTRODUCTION

Here, we get to know about the two terms i.e., Sentiment Analysis and TensorFlow JS. Sentiment Analysis is the process of getting the sentiments, thoughts, emotion of a particular text or video that user post on social media platforms and TensorFlow JS is a library by Google Brain Team in 2015 which helps in developing ML models and it is used by JavaScript developers which indeed help the to deploy ml model in web development.

Social media is drastically increasing day by day and according to a survey done by confirming.co there has been Ninety-Eight percentage of the people use at least four social media platforms daily. YouTube (92.77%) is the top-ranking social media that is used once a day.

YouTube is most popular with **millions** of videos uploading and **billion** mobile **views** and comments for all of its videos. Due to several limitations in current sentiment dictionaries, detecting sentiment polarity in social media, particularly on YOUTUBE, is a difficult process.

It is clear from the studies that the web traffic is 20% and Internet traffic is 10% of the total YOUTUBE traffic. There are many techniques of YOUTUBE for the judgment of opinions and views of users on a video. Voting, rating, favorites, and sharing are some of the techniques available. Sentiment Analysis of user comments is a source of user data that may be used for a variety of applications. Comment filtering, personal suggestion, and user profiling are examples of future applications. content became very difficult sometimes and to solve this problem we are trying to help them give comment statistics about a particular video.

II. RELATED WORK

Research has been done on a range of aspects of YouTube video features. Comments also reflect the user's behavior and will be accustomed locate potential trolls. Moreover, by analyzing the sentiment of comments, it's possible to seek out the user's positivity or negativity about the video. Altingovde et al. developed some way to boost video retrieval supported basic features and social features. Lehner et al. used YouTube comments to point out how a viewer's perception of a video (like/dislike) changes with comments. These methods allow us to seek out the foremost popular videos by employing a number of features in order that they'll be used for the best and handiest searches. Even though the 2 proposed methods demonstrated impressive results for video retrieval, they utilized the like/dislike and views feature to urge an inaccurate result. On the contrary, we only analyze the quantity of comments (rather than the kind/view) feature, to seek out videos which may be of interest to YouTube users.

III. METHODOLOGY

There are four phases to the suggested procedure as shown within the following section, supported NLP-based evaluation of user sentiment comments. The text first goes through some language pre-processing before it goes through NLP-based methods to get data sets. the info sets are then analyzed with a sentiment classifier to calculate the scores per comment. Finally, apply the quality Deviation to get the rating result. [A] Comment Collection and Pre-processing The goal of this section is to gather comments on a selected YouTube video. A focused crawler is implemented to fulfil this objective. It displays comments (up to 100 comments of that video) supported the video id. However, the comments are diverse in terms of the languages and ideas utilized by the users As a result,

we performed some pre-processing on these unstructured remarks so as to make the info sets. Following the extraction of the comments, the subsequent adjustments are made: Remove all expressions that are not associated with the suggested technique, like date ("Dec 2- 2010" or "2-12-2010"), link (www.imdb.com, www.tmdb.com), numerals (12, 20, etc.) and special characters, and different language such as Chinese, Arabic, Bangla, Hindi etc.). The comment collecting and pre-processing module first pulls data (comments) from a specific YouTube video so processes it.

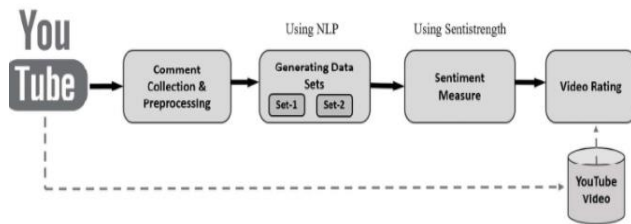


Fig. 1 Working process

Overall work process of evaluation of user comment. IEEE International Conference on Signal and Image Processing Applications (IEEE ICSIPA 2017), Malaysia, September 12-14, 2017475 • Remove any punctuation marks, like the amount ("."), space (" "), commas (","), semicolon (";"), and hash ("#") among others.

[B] Generating Data Sets

For each assessing video, there are two datasets created by the proposed technique. Both are produced using the handled comments text. to create datasets first, within the handled content MySQL stop word is applied to eliminate all the stop words and afterward con-vert each one of the words into their particular structure and hence make dataset 1. Then, for Dataset 2 each one of the descriptors (significant expressions of the remark text) of the remarks are gathered. Observationally and from the self-examination on YouTube video remarks it appears to be that descriptors are the most pointers for a client's inclination and selection about the video's quality and pertinence. Subsequently, to create the rundown of descriptors of the remarks Stanford POS Tagger is applied and distinguish each one of the modifier words and along these lines make dataset.

In this segment, we will in general utilize the SentiStrength3 equivalent locator in each dataset to work out the overall notion of client remarks. SentiStrength could be a sentiment dictionary investigation classifier that estimates the strength of positive and negative assumption of the remark words.

It reports 2 conclusion qualities: - one (not negative) to - 5 (very negative) and one (not positive) to five (incredibly sure). In the unlikely event that a word inside a sentence got a <1 rating, the classifier picks it's anything but a negative word, and whenever got a >=1 rating, then, at that point the classifier chooses it's anything but a positive word.

For the partner degree model, Table I shows the assessment live for a chose remark ("He altered the planet as we as a whole know it and stays very modest. plentiful regard.") per the arranged procedure. From Table I, it looks that, what that remark is which means concerning the video, is it certain or negative. By and large, when hard the slant worth of all remarks the quality deviation worth is determined (by applying change procedure, depicted in next part) of those qualities for the subsequent stage of the strategy.

Table I: Sentiment measure of a comment

Word	pos	neg
changed	1	-1
world	1	-1
remains	1	-1
super	1	-1
humble	1	-1
respect	3	-1

IV. RESULTS

In this segment, we will in general square measure expecting to specify the test we tend to did with the ways. to pass judgment on the arranged methodology the trial is led on 1,000 recordings of YouTube that were first class aimlessly.

Notwithstanding, precisely ten classes (schooling, science, and innovation, redirection, animation, and so forth) were tip top for those recordings. for each class 100 recordings were contemplated and for each video, 1,000 remarks were pondered. Prior to the analysis, a manual survey was performed on at irregular tip top 100 recordings of YouTube.

Ten volunteers were enlisted to legitimize the aftereffects of the video through the arranged strategy. Volunteers checked the association and nature of the video. Result educated that singular video dataset having > zero.5 normal American state is ideal and significant per the inquiry. it had been moreover tracked down that the recordings square proportion of top quality, awesome, important, and standard once the normal American state becomes > zero.5 considering each dataset along. consequently, for the investigation of the arranged technique ology, the verge cost was set to zero.5.

Test result's displayed in Table II. The outcome shows that to search out the significant and standard video in YouTube exploitation remarks square measure viable. From Table II, it appears, for dataset one, the outcome's much better than for dataset a couple of. the best precision 75. 435% of the arranged methodology happened for science and innovation associated recordings any place the most reduced exactness fifty one.21% has been determined

for animation classes of video. furthermore, for dataset one, the run of the mill exactness of the arranged methodology was 77.462% once entire remarks were considered. Regardless of the fact that precision was low for the dataset a couple of anyway exploitation descriptors gives rather righter outcome regularly.

Subsequently, considering each dataset may give the \$64000 picture of a chose video than exploitation explicit one dataset. For the Associate in Nursing model in Table III, when a video was looked through like "Huawei Ascend p7 Review", inside the hunt board it had been shown that the underlying few recordings were pertinent and great. anyway, to work out which one was the awesome more famous and which one ought to show up first? For this, dissecting each dataset was fundamental. From the outcome in Table III, it appears to be that the principal video of dataset 1 gives preferred precision over the second video of dataset 1. Notwithstanding, the exactness of dataset 2 of the subsequent video is obviously superior to dataset 2 of the principal video.

Table III: An example of measuring video relevancy and quality through the proposed method

Video	Dataset	Accuracy when considering each dataset	Accuracy when considering both dataset together
Video1	Dataset1	1.134411513	0.866642
	Dataset2	0.598874	
Video2	Dataset1	1.120566227	0.939081
	Dataset2	0.757596	

Accordingly, in such a case considering each dataset gives us the genuine image of those recordings' quality and flawlessness. Like, for this situation, the subsequent video is awesome and much pertinent than the first. Along these lines, in the unlikely event that both datasets are considered to discover the video on YouTube, the outcome may be greatly improved. From the outcome, it very well may be presumed that, if the investigation of YouTube video per-shaped dependent on the remarks text, then, at that point our proposed approach may submit the great outcome.

In any case, it depends however much semantically the content is dissected, the more we break down the content the more opportunity to improve exactness.

V. CONCLUSION

This paper delineates A programmed technique for finding supportive recordings by estimation examination of client's remarks Natural Language measure (NLP).

Our methodology assessed the norm, relevance, and acknowledgment of YouTube recordings considering the connection of client's feelings communicated in remarks. we will in general examine an example of pretty much 1,000,000 YouTube remarks.

Huge scope investigations of YouTube video Meta data (remark) abuse the data science and SentiStrength un-

covered the significance of client sentiments. The exploratory outcome shows the strength of the extended methodology by revealing most 75.435% exactness to recovering powerful video and conjointly gives the bearing to seek after the work to attempt to additional examination on the remark.

ACKNOWLEDGMENT

We wish to communicate a genuine feeling of appreciation towards our guide, Prof. Asha Pawar who at exceptionally discrete stages in the investigation of this venture, contributed their significant direction and assisted with taking care of each difficult that emerged. We likewise thank our H.O.D Dr. Rajneesh Kaur Sachdeo Bedi, for giving the essential offices to the finishing of this task work in our school. We might want to thank our companions and partners who inspired us to do the exploration.

REFERENCES

- [1] Shi Yuan, Junjie Wu, Lihong Wang and Qing Wang, "A Hybrid Method for Multi-class Sentiment Analysis of Microblogs", ISBN- 978-1-5090-2842-9, **2016**.
- [2] Neethu M S and Rajasree R, "Sentiment Analysis in Youtube using Machine Learning Techniques"
- [3] Aliza Sarlan, Chayanit Nadam, and Shuib Basri, "Youtube Sentiment Analysis", 2014 International Conference on Information Technology and Multimedia (ICIMU), Putrajaya, Malaysia **November 18 – 20, 2014**.
- [4] B. Gupta, M. Negi, K. Vishwakarma, G. Rawat, and P. Bandhani, "Study of Youtube Sentiment Analysis using Machine Learning Algorithms on Python," Int. J. Comput. Appl., **vol. 165, no. 9, pp. 29-34, May 2017**.
- [5] "Computationally Efficient Learning of Quality Controlled Word Embeddings for Natural Language Processing," 2019 IEEE Comput. Soc. Annu. Symp. On, p. 134, 2019. Opinion Mining", Kluwer Academic Publishers. Printed in the Netherlands, **2006**.
- [6] Hearst, M., "Direction-based text interpretation as an information access refinement", In Paul Jacobs, editor, Text-Based Intelligent Systems. Lawrence Erlbaum Associates, **1992**.
- [7] Das, S., and Chen, M., "Yahoo! for Amazon: Extracting market sentiment from stock message boards", In Proc. of the 8th Asia Pacific Finance Association Annual Conference (APFA 2001), 2001. unsupervised classification of reviews". In Proc. of the ACL, **2002**.
- [8] Argamon-Engelson, S., Koppel, M., and Avneri, G., "Style-based text categorization: What newspaper am I reading? ", In Proc. of the AAAI Workshop on Text Categorization, **pages 1-4, 1998**.
- [9] Pang, B. & Lee, L., "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", Association of Computational Linguistics (ACL), **2004**.
- [10] Jin, W., & HO H. H., "A novel lexicalized HMM-based learning framework for web opinion mining", Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Quebec, Canada, ACM: 465-472, **2009**.
- [11] Brody, S., & Elhadad, N., "An unsupervised aspect-sentiment model for online reviews", Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California, Association for Computational Linguistics: 804-812, **2010**.
- [12] Wiebe, J., Wilson, T., and Cardie, C., "Annotating expressions of opinions and emotions in language". Language Resources and Evaluation, **2005**.