International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

YouTube Video Analyzer Using Sentiment Analysis

Goli Sushma[1], Vadicherla Raju[2], Rajashekar Kandakatla[3], Neethipudi. Sashi Prabha[4], Dr. Rajesh Saturi[5], Dr. L. Mohan[6]

Submitted: 06/12/2023 **Revised**: 12/01/2024 **Accepted**: 30/01/2024

Abstract: Sentiment analysis is a method used to ascertain the opinions and viewpoints of people regarding any service or product. With millions of views, YouTube is one of the most widely used sites for sharing videos. With the ever-increasing popularity of online videos and the exponential growth of user-generated content, understanding the quality and relevancy of content has become crucial for viewers by looking over the comments, number of views, and number of likes manually. The goal of this paper is to create a thorough methodology and a useful tool for assessing user sentiment on YouTube videos. The suggested method extracts text comments and transcripts from YouTube videos for examination using cutting-edge natural language processing algorithms and categorizes them into opinions that are neutral, positive, negative, relevant, and irrelevant along with a transcript summary. Ultimately, this research endeavors to revolutionize the way YouTube videos are analyzed, facilitating informed decision-making and enhancing user experience on the platform.

Keywords: *Sentiment analysis, Natural Language Processing, Opinion extraction, Decision–making.*

1. Introduction

YouTube is a useful tool for data collection because of its wide variety of user-generated material. It offers a vast dataset for a variety of analytical and research applications because of its billions of videos and millions of users. Since social scientists use YouTube data to investigate cultural and social phenomena, and advertisers rely on it for targeted marketing, YouTube data is an invaluable resource for a wide range of academic fields. The platform's comments and debates provide a wealth of information that may be used to study public sentiment and opinion on a range of topics. Furthermore, YouTube collects data for educational research, which aids academics and institutions in evaluating the efficacy of instructional tactics and online learning. YouTube's enormous content library, which includes everything from news and personal vlogs to entertainment and tutorials makes it a priceless tool for academics, companies, educators, and marketers wishing to leverage the wealth of data produced by users and their interactions with the platform. As content creators and content are increasing rapidly there evolves an issue of data relevancy. Some people seekto watch long tutorials on various subjects and if the content is just slightly relevant or mostly irrelevant then the user's time is wasted. To overcome such problems, we perform NLP-based sentiment analysis on real-time data collected from YouTube i.e., YouTube comments and transcript of the videos, then analyze and classify them using various libraries such as Textblob and Spacy, and finally generate the output in the form of graphs and summary. There are two graphs, one represents the classification of the comments into positive, negative, and neutral based on polarity values, and the other one into relevant and irrelevant. Transcript summarization is done using spacy based on the normalized frequency of sentences under the condition of required parameters. Based on the output, the user can decide whether to watch the video, just read the transcript, or proceed to the next one.

[1]*Assistant Professor, Department of CSE (Data Science), Gurunanak institutions Technical Campus, Hyderabad sushmasri227@gmail.com,*
[2]*Assistant Professor , Department of CSE(Data Science), Vignana Bharathi Institute of Technology, Hyderabad, India, vadicherla.raju@vbithyd.ac.in*
[3]*Assistant Professor, Department of CSE (Artificial Intelligence and Machine Learning), Vaagdevi College of Engineering(Autonomous), Bollikunta, Warangal, Telangana, rajashekar6902@gmail.com*
[4]*Assistant Professor , Department of Computer Science and Engineering, Gokaraju Rangaraju Institute of Technology, Hyderabad, prabha1735@grietcollege.com*
[5]*Associate Professor Department of Computer Science and Engineering, Vignana Bharathi Institute of Technology, Hyderabad, India, rajesh.saturi@vbithyd.ac.in*
[6]*Associate Professor, Dept. CSE(Internet of Things), Balaji Institute of Technology and Science(Autonomous), Narsampeta, Warangal, Telangana. lavmohan@gmail.com corresponding*
Author: Dr.Rajesh saturi, rajesh.saturi@vbithyd.ac.in

2. Related Work

Sentiment analysis on YouTube was relatively nascent, and research mainly revolved around text-based comments and textual metadata analysis. Researchers explored basic sentiment lexicons and rule-based approaches to gauge audience reactions.

Many research papers are published in the field of sentiment analysis and text summarization. We

thoroughly examined the following papers to acquire a comprehensive understanding of this field. The review papers and their descriptions are presented below with utmost attention to detail.

- Mohammed Arsalan Khan, SumitBaraskar, Anshul Garg, Shineyu Khanna, and Asha M. Pawaranalyzed the comments of the YouTube video and determined the video rating through estimation examination of the client's remarks Natural Language measure, and uncovered the significance of client sentiments using sentistrength.
- K.M. Kavitha, Asha Shetty, Bryan Abreo, and Akshara Kondanapresented a comparative study of classifiers utilizingthe Bag of Words and Association list techniques. They further concluded that non-contiguous phrases should also be investigated as features.
- Aditya Baravkar, RishabhJaiswal, and Jayesh Chhoriya analyzed the nature of comments by adding a parameter of sentiment to training data. Further,they concluded that the model can be trained using true comments fetched from API for more accurate decision-making.
- Deepali K, Gaikwad, and C. Namrata Mahender proposed a system that summarizes a text by identifying and highlighting significant sentences in the original text and concatenating them to create a succinct summary and found that abstractive summarization is relatively more complicated than the extractive approach, requiring more reasoning and expertise.
- DuyDucAn Bui PhD, Guilherne Del Fiol MD, PhD, John F. Hurdle MD, PhD, and Siddhartha Jonnalagadda PhD presented an extractive text summarization system to prioritize sentences employing sentence ranking and key phrase extraction.They also used an ensemble approach, combining three different extraction methods to achieve the best extraction performance.

3. Proposed System

The proposed system considers the sentiment of comments and transcript of the given video to provide relevancy classification and accurate results as per the user's request. The main aim of this paper is to develop an innovative YouTube Video Analyzer that leverages the power of sentiment analysis to gain valuable insights into viewers' emotions, opinions, and attitudes towards the video content through comments and provide a video summary using the transcript. The system utilizes natural language processing techniques and machine learning algorithms i.e., Gaussian NB and Logistic Regression to assess the sentiment of comments on YouTube videos automatically. It classifies comments as positive, negative, neutral, relevant, and irrelevant based on polarity values. The system uses real-time data from comments and video transcripts to train its sentiment analysis model. The primary goal is focused on the user's decision-making strategy within less range of time by providing the statistics of comments and ensuring accurate sentiment analysis irrespective of complex language expressions.

A simple web interface is created using Flask where the user can enter a valid YouTube video URL that needs to be analyzed. As soon as the user clicks on the analyze button, the application searches for comments and transcripts related to the video in the backend. Then, the comments related to the video are sent to the sentiment analysis model for classification and the result in graphical terms. Simultaneously, the transcript of the video is sent to the summarization model which creates a pipeline to generate a summary and return it. As a result, graphical representations of sentiment distribution of comments and transcript summary will be displayed to the user.

4. Methodology

The project mainly consists of three sub-modules. These three modules are as follows:

a) YouTube Video Comment Classification
b) YouTube Video Transcript Summary Generation
c) Integration and User Interface

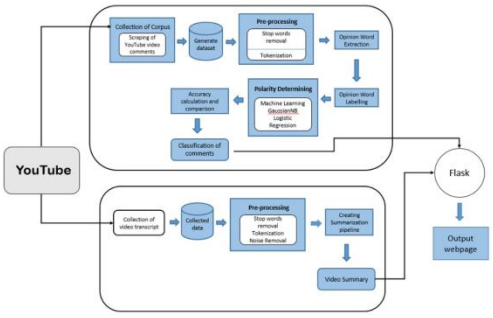

Fig – 1: Architecture of YouTube Video Analyzer

4.1 YouTube Video Comment Classification

User comments related to the YouTube video provided by the user are collected by making use of the YouTube Data API v3 which allows developers to access video statistics by making REST and XML-RPC calls using its URL. These scraped comments are stored in a response object and are processed to extract only comments text ignoring other metadata, and stored in a list to create a data frame of comments.

The collected comments in the data frame undergo preprocessing, including lowercasing, tokenization using WordNetLemmatizer, and removal of special characters,

new line characters, punctuations, multiple spaces, references, hashtags, and stopwords using the substitute method of regular expressions package.

Later, Sentiment analysis is implemented on these preprocessed comments. Polarity is calculated for each comment in the data frame using the TextBlob library and transformed the polarity values into a set of -1, 0, and 1. Comments with polarity values greater than 0 are labeled as positive, those with values less than 0 as negative, and those with values equal to 0 as neutral. A classification machine learning model i.e., a Logistic Regression classifier is trained with the labeled sentiment data and predicts the results for test data.

In parallel with sentiment analysis, comments are classified as relevant or irrelevant to the video's content based on the same polarity values. The comments whose polarity value is greater than 0.5 or less than -0.5 are labeled as relevant, and those are in between the range of -0.5 to +0.5 labeled as irrelevant. GaussianNB classifier is trained with the labeled sentiment data and predicts the results for test data.

Finally, the module outputs sentiment analysis results, relevance classification for each comment, and classification reports. These results can be visualized using bar graphs plotted using the Matplotlib library, allowing users to understand the overall sentiment and relevance distribution.

4.2 YouTube Video Transcript Summary Generation

Initially, the transcript of the YouTube video is obtained through a third-party youtube – transcript – api by passing the corresponding video ID as a parameter to it. The extracted transcript undergoes preprocessing, including lowercasing, tokenization, and removal of special characters, new line characters, punctuations, multiple spaces, references, hashtags, and stopwords imported from the spacy library. This step ensures the transcript is free from noisy text.

After preprocessing the video's transcript, a word frequency table is created in which the frequency of each word i.e., the number of times each word has appeared in the transcript is recorded as a key–value pair. Thereafter, the frequency of each word in the word frequency table is normalized by dividing it by the maximum frequency obtained from the word frequency table. The frequency of each transcript sentence is measured by adding the normalized frequency of each word in that particular sentence. This step ensures the transcript is ready for summarization.

As a result, the summary of the video is generated using the nlargest method of the heapq library by providing it with the preprocessed transcript, calculated sentence scores for each sentence of the transcript, and maximum length or maximum percentage of summary to be generated. Depending upon these parameters, the function will ignore the sentences with the least frequency and include those with the highest frequency to generate the summary. This summary is designed to provide users with a quick overview of the video's content.

4.3 Integration and User Interface

The outputs of both modules above are integrated into a single interface using Flask, a web framework of Python. The user interface comprises input fields for the video URL and buttons to initiate analysis. Users can input a YouTube video URL, and the system processes its comments and transcript to provide sentiment distribution graphs, relevance classification, and the generated video summary in an intuitive manner.

5. Results

The required comments and transcript of a video are fetched successfully and trained classification models with that data resulting in an accuracy of 90%. Sentiment and relevancy distribution graphs along with the summarized transcript of the provided video URL are displayed. This output showcases the relevancy of the YouTube video to a respected user and they can decide whether to watch the video or proceed with the next one by saving their time.

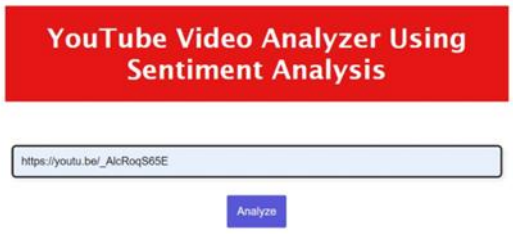

Fig – 2: YouTube Video Analyzer Webpage

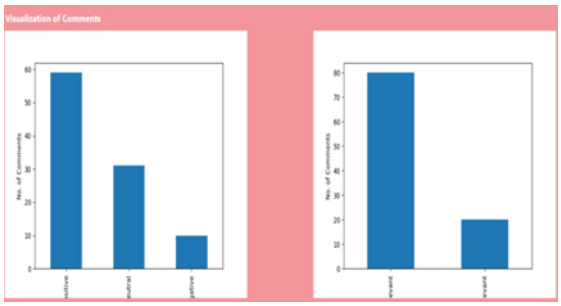

Fig – 3: Positive, negative, neutral, relevant, irrelevant classifications of the given YouTube video

Fig – 4: Transcript Summary of the given YouTube video

Table - 1: Classification report of positive, negative, and neutral comments

	Precision	**Recall**	**f1-score**	**Support**
negative	0.00	0.00	0.00	1
neutral	0.83	1.00	0.91	10
positive	1.00	0.89	0.94	9
accuracy			0.90	20
macro avg	0.61	0.63	0.62	20
weighted avg	0.87	0.90	0.88	20

Table - 2: Classification report of relevant and irrelevant comments

	Precision	**Recall**	**f1-score**	**Support**
Relevant	0.92	0.71	0.80	17
Irrelevant	0.29	0.67	0.40	3
accuracy			0.70	20
macro avg	0.60	0.69	0.60	20
weighted avg	0.83	0.70	0.74	20

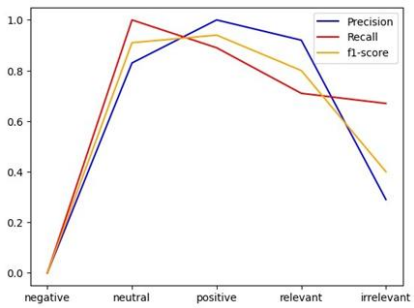

Fig – 5: Precision, recall, and f1-score graph for sentiment classification of the given YouTube video

6. Conclusion

As a result, this paper offers a very reliable YouTube video analyzer that makes use of NLP-based sentiment analysis to provide a useful tool for helping viewer sentiments and improving content understanding.This researchmakes a significant contribution to the field of sentiment analysis and its applications in the evaluation of video material since it has the potential to revolutionize content analysis on YouTube. The model is trained using true comments from an API for more accurate decision-making. The paper reviews sentiment analysis methods for YouTube videos, categorizing them into positive, negative, neutral, relevant, and irrelevant levels. This classifier-based technique automatically classifies YouTube comments according to how related they are to the video content. It can be combined with a recommendation system for increased reliability and productivity. It can be extended to categorize YouTube comments that contain multilingual phrases or words and assess the video content based on its audio for those videos whose transcript and comments are disabled as future work. Abstractive summarization requires more learning and reasoning but provides more meaningful and appropriate summaries. However, there is limited work on abstractive methods in Indian languages, suggesting potential for further exploration.

References

[1] Mohammed Arsalan Khan, SumitBaraskar, Anshul Garg, Shineyu Khanna, Asha M. Pawar, “YouTube Comment Analyzer”, International Journal of Scientific Research in Computer Science and Engineering, August 2021

[2] Aditya Baravkar, RishabhJaiswal, JayeshChhoriya, “Sentimental Analysis of YouTube Videos”, International Research Journal of Engineering and Technology (IRJET), Volume: 07, Issue: 12, Dec 2020

[3] K.M. Kavitha, Asha Shetty, Bryan Abreo, Adline D’Souza, AkarshaKondana, “Analysis and Classification of User Comments on YouTube Videos”, ScienceDirect, Nov 2020

[4] Deepali K., Gaikwad, C. NamrataMahender, “A Review Paper on Text Summarization”, International Journal of Advanced Research in Computer and Communication Engineering

[5] DuyDucAn Bui PhD, Guilherne DelFiol MD, PhD, John F. Hurdle MD, PhD, Siddhartha Jonnalagadda PhD, “Extractive text summarization system to aid data extraction from full text in systematic review development”, Journal of Biomedical Informatics, Oct 2016.

[6] 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.

[7] Neethu M S and Rajasree R, “Sentiment Analysis in Youtube using Machine Learning Techniques”

[8] 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.

[9] 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.

[10] “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.

[11] Hearst, M., “Direction-based text interpretation as an information access refinement”, In Paul Jacobs, editor, Text Based Intelligent Systems. Lawrence Erlbaum Associates, 1992.

[12] 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.

[13] Argamon-Engelson, S., Koppel, M., and Avneri, G., “Stylebased text categorization: What newspaper am I reading? ”, In Proc. of the AAAI Workshop on Text Categorization, pages 1–4, 1998.

[14] 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.

[15] 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.

[16] Wiebe, J., Wilson, T., and Cardie, C., “Annotating expressions of opinions and emotions in language”. Language Resources and Evaluation, 2005.

[17] https://ieeexplore.ieee.org/document/9396049

[18] https://www.sciencedirect.com/science/article/pii/S1877050920323553

[19] https://www.isroset.org/pdf_paper_view.php?paper_id=2461&6-ISROSET-IJSRCSE- 06279.pdf