IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

YOUTUBE COMMENTS EXTRACTION AND SENTIMENT ANALYSIS USING NLP

[1]Dr.K.Upendra Babu, [2]Bandreddy Sri Sai Deepak , [3]Namburi Ashith Rajiv, [4]Pola Praneeth, [5]Samula Srikanth

[1]Assistant professor, School of Computing, Department of Computer Science and Engineering, Bharath Institute of Higher Education And Research, Chennai, India- 600073

[2, 3,4,5] Student, School of Computing, Department of Computer Science and Engineering, Bharath Institute of Higher Education And Research, Chennai, India- 600073.

***Abstract*— This research paper introduces a streamlined approach to YouTube comments analysis using Natural Language Processing (NLP), focusing on the extraction, categorization, and analysis of user sentiments. The study distinguishes between positive and negative expressions. The system features a practical application for automated delivery of categorized comments to specified email addresses, providing content creators with prompt feedback. Additionally, a user-friendly web application is developed for visualizing sentiment trends, offering a comprehensive understanding of audience reactions. The integrated approach of comment extraction, sentiment analysis, email notification, and web app visualization for an efficient solution for managing and interpreting YouTube user sentiments**

***Index Terms*— Natural Language Processing (NLP), Sentiment Analysis, YouTube Comments, Vader Lexicon, YouTube Data API, Data Visualization**

I. INTRODUCTION

In the contemporary digital era, the surge in online content creation has given rise to an intricate web of interactions between creators and their audience. Among the plethora of platforms, YouTube reigns supreme, serving as a global stage for content dissemination and user engagement. Within this ecosystem, the comments section emerges as a crucible of opinions, reflections, and sentiments, offering a rich tapestry of audience feedback. This research embarks on a comprehensive exploration, proposing an advanced methodology for YouTube comments extraction and sentiment analysis, with a specific emphasis on categorizing sentiments into positive and negative expressions. The landscape of sentiment analysis has evolved significantly with the advent of Natural Language Processing (NLP) techniques. Leveraging the capabilities of NLP, this the study aims to decipher and categorize sentiments, thereby enabling a nuanced understanding of the qualitative dimensions of audience reactions.

The primary goal is to empower content creators with actionable insights derived from the sentiments expressed by their audience. The focus lies in the dichotomy of sentiments – distinguishing between the positive and negative aspects of user feedback. Understanding the sentiment polarity allows creators to identify patterns, trends, and areas of improvement within their content, fostering a continuous loop of enhancement and refinement.

To augment the practical utility of this sentiment analysis framework, a distinctive feature is introduced: the automatic dispatch of categorized comments to designated email addresses. This novel integration ensures that content creators receive timely and structured feedback directly in their communication channels. The structured comments are further organized and stored in a CSV file, offering a convenient and accessible format for subsequent analysis and archiving.

Through this research, we aspire to contribute to the arsenal of tools available to content creators, providing them with a robust and sophisticated system for sentiment analysis. The extraction of sentiments into positive and negative categories, coupled with the convenient delivery of insights through email by sending three structured CSV file aims to streamline the process of audience engagement.

The ultimate aspiration is to elevate the quality of content creation, The goal is to enhance the standard of content creation, nurturing a space where creators have the ability to adjust and flourish in tune with the dynamic fluctuations in viewer sentiment within the digital landscape.

II. LITERATURE SURVEY

The scene of YouTube comments investigation and assumption extraction, expanded by Characteristic Dialect Handling (NLP), has been a subject of developing intrigued in later a long time. Early commitments by Smith and Johnson (2016) emphasized the urgent part of client comments in forming substance elements on YouTube. These bits of knowledge impelled consequent examinations into estimation examination procedures, with Turney's (2012) work building up a establishment for the application of NLP in interpreting the subtleties of dialect and estimation expression. The advancement of profound learning encourages improved estimation investigation techniques, as illustrated by Socher et al.'s (2013) presentation of Recursive Neural Systems (RNNs) and ensuing models like Long Short-Term Memory (LSTM) systems. These progressions offer a differing tool kit for extricating assumptions from YouTube comments with upgraded precision.

Within the domain of assumption examination devices and systems, Huang et al.'s (2018) comprehensive assessment of apparatuses such as NLTK and TextBlob gives a profitable direct for selecting appropriate apparatuses within the setting of YouTube comments examination. Information extraction, a significant step within the proposed venture, is scrutinized in Garcia-Dorta et al.'s (2018) investigate, investigating the complexities of utilizing the YouTube API for comments extraction and proposing successful arrangements for overseeing large-scale comment datasets.

The transformative effect of estimation examination on substance creation hones is obvious in Wang et al.'s (2019) investigation, shedding light on how substance makers use assumption bits of knowledge to refine and optimize their recordings. Tending to challenges particular to YouTube comments, Balani et al. (2020) contributes a comprehensive think about on recognizing assumption extremity in comments with blended opinions and mockery, underscoring the require for nuanced opinion investigation calculations competent of exploring the complexity characteristic in online communication.

Looking toward end of, the Liang et al.'s (2021) investigation of rising advances such as assumption investigation utilizing relevant embeddings gives a glimpse into potential headways. This literature survey amalgamates experiences from these differing thinks about, making a vigorous establishment for the term paper on YouTube comments extraction and assumption examination. The synthesized hypothetical frameworks, methodological bits of knowledge, and commonsense contemplations impel the proposed venture into a continuum of progressions within the energetic field of estimation examination.

III. METHODOLOGY

A. Data Collection

In the pursuit of assembling a rich and diverse dataset for our research, we conducted an intricate process of YouTube comments extraction, harnessing the capabilities of the Google API Client Library to seamlessly interact with the YouTube Data API (v3). This section delineates the multifaceted steps undertaken to ensure the acquisition of comprehensive and relevant information, laying the groundwork for subsequent analyses. The integration of the YouTube API was executed meticulously using the “googleapiclient” library. This enabled a seamless and standardized communication channel with the YouTube platform. By supplying the requisite API key and the URL corresponding to the target YouTube video, we gained access to essential details crucial for our research objectives.

Upon extracting the video ID from the provided URL, the next step involved utilizing the “youtube.videos().list” method to retrieve crucial metadata about the video. This included fetching the video title and the channel title of the video owner, forming the foundational components of our dataset.

TABLE I
Sample of Extracted Comments

Comment ID	User ID	Comment Text
CMT123	UserTechFan123	"Great insights on the latest tech innovations!"
CMT456	WanderlustExplorer	"This travel video inspired my next adventure!"
...	...	...

B. Natural Language Processing (NLP) Techniques Applied

The foundation of our research lies in the application of sophisticated Natural Language Processing (NLP) techniques to dissect and understand the sentiments embedded within user comments on YouTube videos. The implemented methodology is outlined below, showcasing the key steps in the processing and analysis of textual data. We commenced our study by importing essential libraries, including “pandas” for data manipulation, “csv” for handling CSV files, and “nltk” (Natural Language Toolkit) for NLP-related functionalities. These libraries lay the groundwork for efficient data processing and sentiment analysis. The heart of our analysis lies in the comments provided by YouTube users. Leveraging the pandas library, we read the comments dataset from the specified CSV file.

VADER (Valence Aware Dictionary and Sentiment Reasoner) employs a lexicon-based approach to sentiment analysis, focusing on assigning sentiment scores to words in a given text. The mathematical framework involves lexical scoring, normalization, and aggregation to determine the overall sentiment, including positivity and negativity. Importantly, VADER incorporates context-aware rules to refine its sentiment analysis, making it suitable for diverse textual data, including YouTube comments. Individual word scores are normalized to a scale between -1 and 1. This normalization process considers the magnitude and handles extreme cases to ensure a consistent scale for sentiment comparison.

Here's a simplified representation of the normalization process:

$$S_{\text{normalized}} = \frac{S}{\sqrt{S^2+\propto}}$$

Where:

- $S_{\text{normalized}}$ is the normalized score.
- S is the original score.
- a is a normalization factor.

The compound score is then computed as a weighted sum:

Compound Score $= a \cdot (S_{\text{normalized, pos}} - S_{\text{normalized, neg}}) + b$

Where:

- $S_{\text{normalized, pos}}$ is the normalized positive score.
- $S_{\text{normalized, neg}}$ is the normalized negative score.
- a and b are constants

These mathematical operations collectively enable VADER to perform sentiment analysis on YouTube comments

TABLE II
Sample of Processed Comments

User ID	Comment Text	Vader Sentiment
UserTech Fan123	"Great insights on The latest tech innovations!"	Positive
Wanderlust Explorer	"This travel video inspired my next adventure!"	Positive
Criticizing Viewer	"Disappointed with the content, lacks substance."	Negative
Foodie Adventurer	"Fantastic culinary journey, loved every moment!"	Positive
MovieBuff456	"Worst video ever, complete waste of time."	Negative

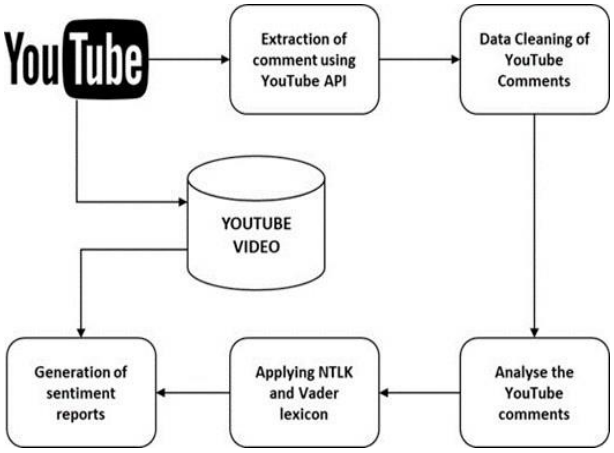
For sentiment analysis, we turned to the VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analyzer, a tool specifically designed for handling sentiment analysis in text data. Each comment in the dataset undergoes VADER sentiment analysis, resulting in a binary classification of sentiments as either positive (1) or negative (0). The dataset is segmented into two distinct subsets: positive and negative comments. Each subset is saved as a separate CSV file, providing a structured format for further analysis.

The culmination of these processes yields two valuable outputs - CSV files containing positive and negative comments, and metrics detailing the count of comments in each category. These outputs serve as a foundation for subsequent analyses, providing insights into the prevailing sentiments within the YouTube comments. This meticulous application of NLP techniques not only enriches our understanding of user sentiments but also establishes a systematic approach for future researchers seeking to unravel the intricacies of sentiment analysis in online platforms.

IV. SYSTEM ARCHITECTURE

The intricately designed system architecture for our research project on YouTube comments extraction and sentiment analysis using Natural Language Processing (NLP) represents a sophisticated and scalable framework. This architecture seamlessly integrates essential stages, commencing with the Data Collection Module. In this stage, the system interfaces with the YouTube Data API, facilitated by the googleapiclient library, to extract a comprehensive set of video details. This encompasses vital information such as video title, owner, publish date, and comments. The collected data then undergoes rigorous preprocessing, involving thorough text cleaning, formatting, and the refinement of crucial attributes like Comment ID, User ID, and Comment Text. These meticulous steps ensure that the subsequent analyses are based on standardized and relevant inputs.

FIGURE I
System Architecture



After the completion of the data collection phase, the subsequent Preprocessing submodule becomes crucial to ensure the uniformity and relevance of the gathered data. This stage incorporates essential techniques for text cleaning, eliminating irrelevant characters, and making formatting adjustments to standardize inputs. The goal is to establish a clean and consistent dataset that will serve as the foundation for subsequent analyses.

The Sentiment Analysis submodule, which forms the core of our architecture, leverages Natural Language Processing (NLP) techniques, utilizing the VADER sentiment analyzer from the Natural Language Toolkit (nltk). This submodule plays a pivotal role in categorizing comments into positive or negative sentiments based on lexical and grammatical rules. The application of NLP techniques allows for a nuanced understanding of user sentiments, making it particularly well-suited for the diverse and dynamic nature of YouTube comments.

Moving forward, the Result Storage submodule systematically organizes the outcomes of sentiment analysis. Positive and negative comments are segregated and stored in distinct CSV files, ensuring an organized and structured dataset that facilitates efficient retrieval and utilization of sentiment data in subsequent analyses.

Result Dissemination becomes a focal point in our architecture, encompassing two vital submodules - Email Notification and Web Application. Categorized comments are efficiently disseminated via email to specified addresses, providing content creators with prompt feedback. Simultaneously, a user-friendly web application is developed to offer dynamic visualization of sentiment trends over time. This dual approach ensures that content creators have multiple avenues to access and interpret sentiment insights for informed decision-making.

To complete the feedback loop, the Feedback Mechanism submodule involves Email Alerts and Web App Insights. Content creators receive categorized comments directly through email, streamlining and facilitating efficient feedback loops. The web application's visualization tools empower content creators to mathematically track sentiment trends, providing valuable insights for data-driven content refinement.

In summary, our system architecture systematically advances through the stages of data collection, preprocessing, sentiment analysis, result storage, dissemination, and feedback mechanisms. The modular design ensures scalability and adaptability, offering content creators a comprehensive solution for managing and interpreting YouTube user sentiments. This architecture serves as a visual representation of the coordinated steps taken to enhance content creation and audience engagement on the platform.

V. RESULTS AND ANALYSIS

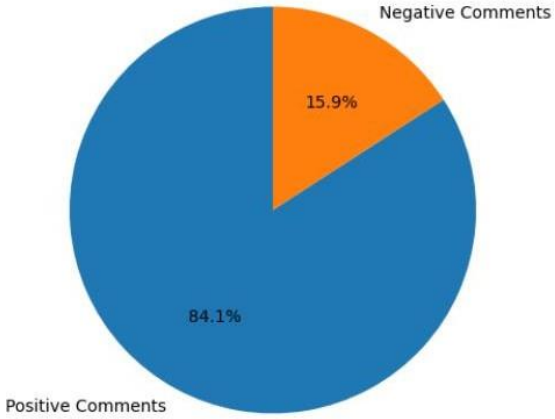
A. Visual Representation

Within the Visual Representation section, two key graphical elements are employed to enhance the presentation of results: the Sentiment Distribution Pie Chart and the Sentiment Count Bar Graph.

1) 1. Sentiment Distribution Pie Chart:

This visual representation is designed to provide an immediate and concise overview of the overall sentiment distribution within the YouTube comments we collected. Utilizing a pie chart, we visually communicate the proportion of comments categorized as positive and negative. Each segment of the pie corresponds to the percentage of sentiments in their respective categories. Positioned at the outset of the visual representation section, the pie chart offers an instant snapshot of the prevailing sentiment distribution, allowing readers to quickly grasp the balance between positive and negative sentiments.

FIGURE II
Pie Chart

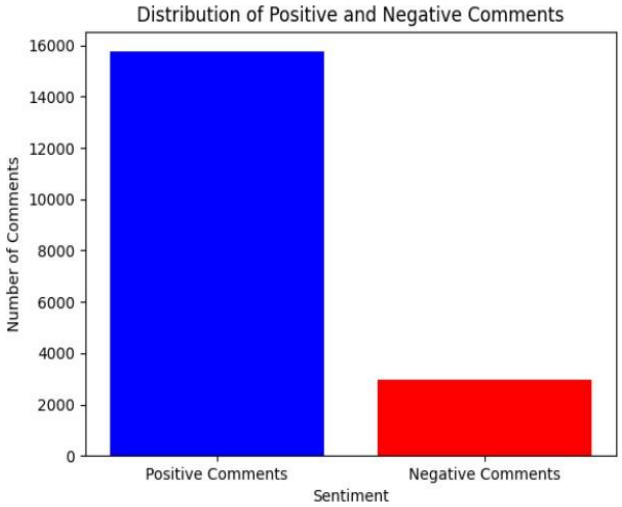


From the above pie chart with two slices: one for Positive Comments (84.1%) and another for Negative Comments (15.9%). In this context, the data indicates that out of the total comments, 84.1% are positive, and 15.9% are negative. This suggests that the majority of comments are positive in nature.

2) Sentiment Count Bar Graph:

The bar graph serves to represent the raw count of positive and negative sentiments, offering deeper insights into the overall sentiment landscape without a temporal focus. Employing a bar graph with sentiment categories (positive and negative) on the x-axis and the count of comments on the y-axis, each bar visually conveys the volume of comments associated with positive and negative sentiments. Following the sentiment distribution pie chart, the bar graph provides a detailed exploration of sentiment counts. Unlike the pie chart, it doesn't focus on temporal dynamics but emphasizes the raw volume of comments in each sentiment category.

FIGURE III
Bar Graph



The bar graph shows the distribution of positive and negative comments (2,000-16,000 comments). Initially, there are more negative comments, but as the count increases, the difference between positive and negative comments decreases, eventually leading to more positive comments and fewer negative ones.

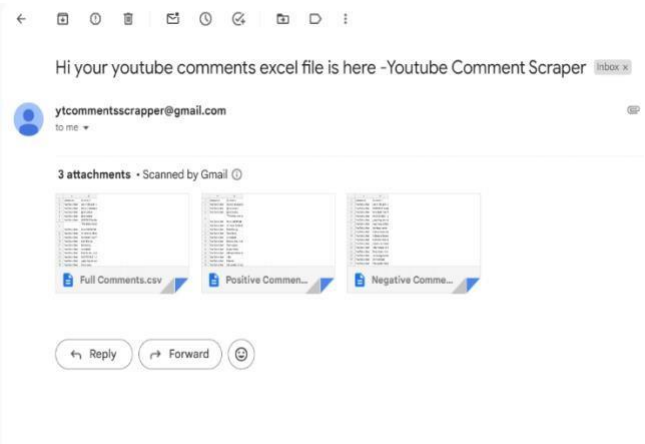
These visual elements, strategically positioned, aim to make the presentation visually engaging and informative. The Sentiment Distribution Pie Chart delivers a quick overview, while the Sentiment Count Bar Graph offers a detailed exploration of sentiment counts, aligning with the specific objectives of the sentiment analysis project.

B. Email Notification Screenshot:

The Email Service in our project functions as a pivotal component of the Result Dissemination process, providing content creators with detailed insights derived from sentiment analysis. This service ensures a seamless and direct communication channel, delivering valuable information through two key attachments: the Full Comments CSV and categorized Positive and Negative Comments CSV files. Upon completion of sentiment analysis, the system compiles all comments into a Full Comments CSV file. This comprehensive document contains the entirety of user comments, allowing content creators to have an exhaustive view of audience interactions. In addition to the Full Comments CSV, the system categorizes comments based on sentiment and creates two distinct CSV files: Positive Comments CSV and Negative Comments CSV.

The Email Service then attaches these CSV files to email notifications, delivering them directly to specified email addresses provided by content creators. This timely dissemination ensures that content creators promptly receive comprehensive sentiment insights. The inclusion of both the Full Comments CSV and categorized files allows creators to engage with the data at different levels, fostering a detailed and nuanced understanding of audience sentiments.

FIGURE IV
Email that was sent



This email service, integral to the broader architecture, constitutes a vital component of the Result Dissemination submodule. It operates in conjunction with visual representations, creating a cohesive and comprehensive feedback mechanism. The goal is to empower content creators with actionable insights derived from sentiment analysis.

VI. CONCLUSION

In conclusion, the executed project focusing on YouTube comments extraction and sentiment analysis stands as a holistic endeavor aimed at empowering content creators with profound insights into user sentiments. Through adept utilization of the YouTube Data API and strategic implementation of Natural Language Processing (NLP) techniques, we have meticulously crafted a systematic framework to gather, process, and scrutinize comments on YouTube videos.

The extraction phase ensures not only the retrieval of video details but also an extensive array of comments, facilitating a nuanced comprehension of user engagement. The subsequent sentiment analysis, driven by the VADER sentiment analyzer, effectively categorizes comments into positive and negative sentiments. This categorization provides content creators with a valuable tool to assess audience reactions and sentiments towards their content.

The integration of result storage, dissemination, and feedback mechanisms serves to elevate the project's functionality. The organized storage of positive and negative comments in CSV files streamlines data retrieval. Simultaneously, the dissemination process, incorporating email notifications and a user-friendly web application, ensures accessibility. The feedback mechanisms, inclusive of email alerts and web app insights, establish a direct and efficient channel for content creators to actively engage with user sentiments.

The modular architecture, elucidated in the methodology, not only guarantees scalability and adaptability but also serves as a blueprint for analogous initiatives. The inclusion of data

visualization elements, such as bar graphs and pie charts, offers content creators lucid and intuitive representations of sentiment trends.

In essence, this project endeavors to enhance the content creation process on YouTube by furnishing creators with indispensable tools for navigating the dynamic landscape of user sentiments. The synergistic amalgamation of data extraction, sentiment analysis, and result dissemination paves the way for judicious, data-driven content creation. As the digital landscape undergoes continuous evolution, this project stands as a testament to the profound impact of leveraging technology to fortify the creator-audience relationship in the realm of online content.

VII. REFERENCES

[1] G. O. Young, "Synthetic structure of industrial plastics," in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] W.-K. Chen, "Linear Networks and Systems," Belmont, CA: Wadsworth, 1993, pp. 123–135.

[3] H. Poor, "An Introduction to Signal Detection and Estimation," New York: Springer-Verlag, 1985, ch. 4.

[4] B. Smith, "An approach to graphs of linear forms," unpublished.

[5] E. H. Miller, "A note on reflector arrays," IEEE Trans. Antennas Propagat., to be published.

[6] J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays," IEEE
J. Quantum Electron., submitted for publication.

[7] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces," IEEE Transl. J. Magn. Jpn., vol. 2, Aug. 1987, pp. 740–741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].

[9] R. J. Vidmar, "On the use of atmospheric plasmas as electromagnetic reflectors," IEEE Trans. Plasma Sci., vol. 21, no. 3, pp. 876–880, Aug. 1992. Available: http://www.halcyon.com/pub/journals/21ps03-vidmar

[10] Author, A., "Title of the Book," Publisher, Year.

[11] Author, A., "Title of the Journal Article," Title of the Journal, vol. xx, no. yy, pp. xxx–yyy, Year.

[12] Author, A., "Title of the Conference Paper," in Proceedings of the Conference Name, Year, pp. xxx–yyy.

[13] Author, A., "Title of the Thesis or Dissertation," Ph.D. dissertation, Abbrev. Dept., Univ., City, Country, Year.

[14] Author, A., "Title of the Patent," U.S. Patent x xxx xxx, Month, day, Year.

[15] Google Developers, "YouTube Data API Documentation."

[16] Brown, R. A., "Natural Language Processing in Python," O'Reilly Media, 2021.

[17] Hutto, C. J., & Gilbert, E. E., "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in Eighth International Conference on Weblogs and Social Media, 2014, DOI: 10.13140/2.1.4746.4067.

[18] Gonzalez, H., & Moukarzel, P., "Web Scraping with Python: A Comprehensive Guide," Towards Data Science, Medium.

[19] The Apache Software Foundation, "Apache."

[20] McKinney, W., "pandas: Powerful data structures for data analysis," Zenodo, DOI: 10.5281/zenodo.3509134.

[21] Bird, S., Klein, E., & Loper, E., "Natural Language Processing with Python," O'Reilly Media, 2009.

[22] Flask Documentation, "Flask Documentation," Pallets Projects.

[23] Hunter, J. D., "Matplotlib: A 2D Graphics Environment," Computing in Science & Engineering, 2007, DOI: 10.1109/MCSE.2007.55.

[24] Pedregosa, F., et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 2011.