iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VIII **Month of publication:** Aug 2023

DOI: https://doi.org/10.22214/ijraset.2023.55579

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VIII Aug 2023- Available at www.ijraset.com

From YouTube Comments to Insights: A Sentiment Analysis of Opinions on Productivity Tools

Devang Shah[1], Masumee Parekh[2]
[1, 2]*Department of Computer Engineering, NMIMS University, Mumbai, Maharashtra, India*

Abstract: ***YouTube, a diverse content hub, acts as a prominent platform for both viewing videos and engaging through comments that express opinions. By carefully analyzing these comments, sentiments associated with products or topics become evident. This research delves into sentiment analysis of YouTube comments, specifically focusing on sentiments towards productivity tools and technologies. Through this examination of conveyed sentiments, it deepens understanding of public perceptions and attitudes toward these tools. Utilizing YouTube's reach and engagement, the study systematically collects comment data using the YouTube API and employs the established VADER sentiment analysis framework. Employing a Bag-of-Words approach, the research incorporates machine learning algorithms like Naive Bayes and Random Forest, achieving accuracy rates of 84.01% and 91.34%, respectively. In-depth temporal analysis of user engagement patterns uncovers trends, with heightened engagement followed by decline, correlating with external events, notably the Covid-19 pandemic. These insights enhance sentiment analysis and illuminate dynamics between societal occurrences, user sentiments, and digital dialogues, offering perspectives on evolving opinions about productivity tools and technologies.***

Keywords: Sentiment Analysis, Random Forest Algorithm, Decision Trees Algorithm, YouTube Comments, VADER.

I. INTRODUCTION

This In an era driven by rapidly advancing digital technologies, the profound influence of social media platforms on society cannot be understated. Among these platforms, YouTube has emerged as a powerhouse for disseminating information, entertainment, and fostering global connectivity. With its diverse array of content creators and extensive user engagement, YouTube stands as an ideal medium to gauge public sentiment and perceptions to extract a plethora of comments made by individuals concerning productivity tools. As a platform that caters to a wide range of demographics, YouTube comments provide a valuable repository of opinions, emotions, and perspectives that offer insight into how digital technologies are shaping individuals' lives.

The pervasive integration of digital tools and technologies into various aspects enhanced productivity, streamlined communication, and facilitated efficient time management. Notion, EverNote, Jira, Slack, and ToDo List are just a few examples of such tools that have gained popularity in various professional and personal domains. These technologies empower individuals to organize tasks, collaborate seamlessly, and optimize their work processes, thereby potentially alleviating stressors and contributing to a sense of accomplishment. However, amidst this transformative landscape, the question of whether these advancements genuinely contribute to self-improvement remains a topic of substantial discussion.

Moreover, we gather data for analysing sentiments from YouTube by using its API, tapping into the large pool of comments made by users. The sentiment analysis itself is conducted utilizing the VADER (Valence Aware Dictionary and Sentiment Reasoner) lexicon, a well-established tool for analysing sentiments from text data. To extract insights from the comments, we used a Bag-of-Words approach. This allows for the creation of a structured representation of text data, capturing the frequency and distribution of words in the comments. Subsequently, machine learning algorithms including Naive Bayes and Random Forest are employed to classify sentiments and find patterns within the data.

II. LITERATURE SURVEY

This research paper explores the sentiments and opinions expressed by users regarding productivity tools and technologies. Alhujaili and Yafooz [1] provide insights into how sentiment analysis works on YouTube, acknowledging the complexities of accurately categorizing diverse user viewpoints. Aufar et al. [2] present a case study focusing on sentiment analysis of comments related to Nokia products, employing Decision Tree and Random Forest algorithms to classify sentiments as positive, negative, or neutral, and subsequently assessing the performance of these algorithms. Furthermore, Singh and Tiwari [3] emphasize the vital role of sentiment analysis in unearthing trends and insights from YouTube comments, highlighting the utilization of various machine learning algorithms to discern sentiments' alignment with real-world events.

©IJRASET: All Rights are Reserved

Bhuiyan et al. [4] propose a Natural Language Processing-based sentiment analysis technique applied to user comments, aimed at enhancing the retrieval of relevant and high-quality YouTube videos. Additionally, the research by Solanki et al. [6] contributes a broader perspective by investigating the development of advanced software tools, demonstrating the iterative enhancement of user experience and productivity through innovative interfaces. Collectively, these references offer a comprehensive understanding of sentiment analysis techniques, their application in deciphering public sentiment, and their relevance to comprehending users' opinions on productivity tools and technologies within the dynamic YouTube ecosystem.

III. METHODOLOGY

Figure 1 shows the layout of the system architecture. Data is collected using YouTube APIs, followed by a preprocessing phase involving tasks such as punctuation removal, lowercase conversion, tokenization, stemming, and stop word elimination. Subsequently, sentiment labelling is achieved utilizing the VADER tool. The core of the architecture involves the construction of a bag-of-words model through counter vectorization, enabling the representation of comments in a structured format. To extract sentiment insights, the pre-processed and labelled data is then subjected to machine learning algorithms, including Naive Bayes and Random Forest. Through this systematic approach, the process effectively discerns and categorizes the sentiment of comments as either positive, neutral, or negative, contributing valuable insights into the perceptions surrounding productivity tools and technologies.

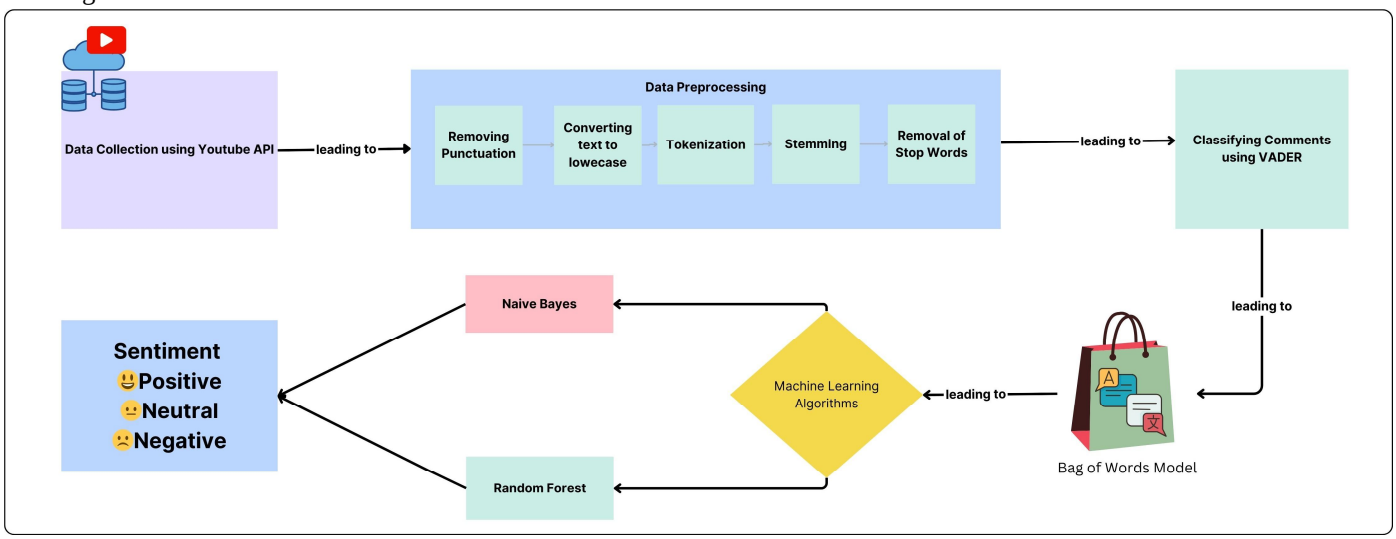


Fig. 1 System Architecture

A. Data Collection

The data collection process involves utilizing the YouTube Application Programming Interface (API) to retrieve comments posted by users in response to videos discussing productivity tools and technologies. The collected data includes the timestamp of each comment as well as the comment text. The timestamp provides the basis for analysing the interactivity of the video over a specified period, enabling insights into viewer engagement dynamics, whereas the comments serve as the primary textual data for subsequent sentiment analysis.

B. Data Preprocessing

Before subjecting the data to sentiment analysis, a series of text data pre-processing steps were undertaken. These steps are essential to enhance the quality and consistency of the text data, thereby improving the accuracy of sentiment analysis results. The purpose of these pre-processing steps is to standardize the text data and reduce noise.

1) *Removal of punctuation marks*: A Punctuation marks such as periods, commas, exclamation points, and question marks are removed from the raw text data. This step is crucial as it streamlines the text by removing non-essential elements. As a result, the content of the comments becomes more focused and conducive to accurate sentiment analysis, allowing for a deeper understanding of the sentiments expressed by the users.

2) *Converting text to lowercase:* A All text in the comments is converted to lower case. This normalization step ensures that the text is treated consistently, regardless of the original casing. It prevents duplicate entries of words with different capitalizations and reduces the complexity of the text data.
3) *Tokenization:* Tokenization involves splitting the continuous text into individual words or tokens. This process breaks down the comments into smaller units, which serves as the foundation for various text processing tasks. Tokenization enables accurate analysis of the sentiment of individual words, allowing VADER and machine learning algorithms to precisely comprehend the context and sentiment embedded within YouTube comments.
4) *Stemming:* Stemming is applied using the Porter stemmer algorithm. This process reduces inflected or derived words to their base or root form, thereby consolidating similar variations of words. Stemming helps to standardize the text data and reduces the dimensionality of the feature space, making it easier for machine learning algorithms to handle.
5) *Removal of stop words: Stemming* Stop words, such as and, the, is, in, of, with, this, on, href, etc. are removed from the comments. These common words do not typically carry significant sentiment information and can be safely excluded without affecting the overall sentiment analysis results. Removing stop words reduces noise in the data and focuses the analysis on more meaningful content.

The pre-processed data is then employed for sentiment analysis using VADER (Valence Aware Dictionary and Sentiment Reasoner).

C. Implementation of VADER

VADER, a lexicon-based sentiment analysis tool, was utilized to assess the sentiment polarity of each pre-processed comment. The VADER sentiment scores provide a quantitative measure of sentiment for each comment. The positive, negative, and neutral scores indicate the proportions of sentiment expressed in the comment, while the compound score represents the overall sentiment intensity ranging from -1 (most negative) to +1 (most positive). Comments with compound scores close to zero are considered neutral.

D. Implementing Bag-of-Words Model

The Bag of Words (BoW) model is a widely used technique in Natural Language Processing that represents text data as a sparse matrix of word occurrences. To implement the Bag of Words model, the Counter Vectorization technique was employed. This involved the creation of a vocabulary containing all unique words across the entire dataset. Each comment was then transformed into a feature vector by counting the occurrences of each word in the comment and mapping it to the corresponding dimension in the vocabulary. The resulting feature vectors represent the comments in a numerical format suitable for machine learning algorithms.
This transformation facilitates the application of various machine learning algorithms, enabling the development of predictive models to classify sentiments effectively.

E. Implementing Bag-of-Words Model

The Bag of Words (BoW) model is a widely used technique in Natural Language Processing that represents text data as a sparse matrix of word occurrences. To implement the Bag of Words model, the Counter Vectorization technique was employed. This involved the creation of a vocabulary containing all unique words across the entire dataset. Each comment was then transformed into a feature vector by counting the occurrences of each word in the comment and mapping it to the corresponding dimension in the vocabulary.
The resulting feature vectors represent the comments in a numerical format suitable for machine learning algorithms. This transformation facilitates the application of various machine learning algorithms, enabling the development of predictive models to classify sentiments effectively.

IV. RESULTS

In this study, we present the results of our sentiment analysis on YouTube comments, which focuses on evaluating the opinions of users regarding productivity tools and technologies. The analysis involved utilizing the YouTube API for data collection, preprocessing the comments, performing sentiment analysis using the VADER lexicon, and subsequently applying machine learning algorithms, namely Naïve Bayes, and Random Forest, based on a bag of words model. The evaluation of these algorithms was based on precision, recall, F1 score, and accuracy metrics.
The following subsections discuss the outcomes of both algorithms and interpret the implications of these metrics in the context of users' opinions about productivity tools and technologies.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

TABLE I
NAÏVE BAYES ALGORITHM

Parameters	Naïve Bayes		
	Negative Class	Neutral Class	Positive Class
Precision	0%	92.77%	65.22%
Recall	0%	88.41%	67.57%
F1 Score	Undefined (0 precision)	90.54%	66.37%
Accuracy	84.01%		

All The Naïve Bayes algorithm demonstrated a moderate performance in sentiment classification (Table I). In terms of precision, which measures the proportion of comments classified as a particular sentiment (negative, neutral, or positive) that are truly representative of that sentiment. It achieved precision for positive class sentiment was 65.22%, 92.77% for neutral class, and 0% for negative class precision.

Recall, representing the algorithm's ability to identify all instances of a given class, was substantial for neutral (88.41%) and positive (67.57%) sentiments, indicating that the algorithm successfully captured a significant portion of these sentiments. However, the negative class recall was 0%, highlighting a limitation in correctly identifying negative sentiments.

The F1 score, which considers both precision and recall, revealed that the algorithm's overall performance was the best for neutral sentiment (90.54%), followed by positive sentiment (66.37%), and undefined for negative sentiment due to zero precision. The accuracy of the Naïve Bayes algorithm was 84.01%, showcasing its proficiency in classifying sentiments on a general level.

TABLE II
RANDOM FOREST ALGORITHM

Parameters	Naïve Bayes		
	Negative Class	Neutral Class	Positive Class
Precision	99.34%	100%	74.42%
Recall	100%	78.05%	100%
F1 Score	99.67%	87.62%	85.38%
Accuracy	91.34%		

The Random Forest algorithm outperformed Naïve Bayes in sentiment classification, especially in precision and recall (Table II). The precision values for negative (99.34%) and neutral (100%) sentiments were impressively high, indicating that the algorithm excelled in correctly classifying comments into these categories. The positive sentiment precision was 74.42%, suggesting that the algorithm showed a good ability to identify positive sentiments, although it was relatively lower compared to the other classes.

Recall was outstanding for negative (100%) and positive (100%) sentiments, indicating the algorithm's robustness in capturing instances of these sentiments. The neutral class recall was 78.05%, showing that the algorithm performed slightly less effectively in identifying neutral sentiments.

The F1 score for negative sentiment was 99.67%, reflecting a balance between precision and recall. The F1 score for neutral sentiment was 87.62% whereas the F1 score for positive sentiment was 85.38%, demonstrating the algorithm's ability to find a balance between correctly classifying positive sentiments. The accuracy of the Random Forest algorithm was 91.34%, showcasing its superiority in sentiment classification compared to Naïve Bayes.

F. Comparative Analysis

The comparison of metrics between the Naïve Bayes and Random Forest algorithms reveals significant differences in their performance.

For positive sentiments, the Random Forest algorithm's higher F1 score suggests its effectiveness in identifying and encapsulating users' expressions of satisfaction, appreciation, or excitement towards productivity tools and technologies. In contrast, Naïve Bayes might struggle to accurately discern such nuanced positive sentiments.

Regarding negative sentiments, the Random Forest's elevated F1 score implies a better ability to navigate the intricacies of sarcastic language, subtle criticism, or indirect phrasing often used by users, potentially enabling it to capture more genuine negative opinions. Naïve Bayes, with a lower F1 score, may falter in accurately detecting these complexities.

In terms of neutral sentiments, Random Forest's superior F1 score signifies its proficiency in capturing informative and balanced opinions that provide valuable insights into users' experiences without strong emotional bias. Naïve Bayes, while displaying a comparatively lower F1 score, might struggle to interpret such neutrality accurately.

The accuracy metric serves as a vital indicator of a sentiment analysis algorithm's overall performance. The higher accuracy demonstrated by the Random Forest algorithm across all sentiment classes signifies its robustness in capturing the intricacies of users' sentiments towards productivity tools and technologies. This enhanced accuracy, especially in classifying negative sentiments, ensures that content creators and developers can accurately gauge users' reactions, whether they are enthusiastic, dissatisfied, or providing informative insights. By leveraging the Random Forest algorithm's superior accuracy, stakeholders can make more informed decisions, refine their strategies, and optimize the user experience based on a more reliable sentiment analysis, which is essential for the success and improvement of productivity tools and technologies.

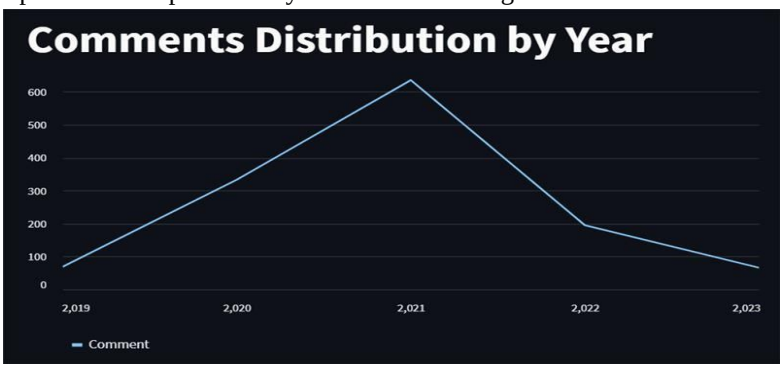


Fig. 2 Distribution of Comments

The observed trend in the line chart (Figure 2), with a notable increase in interactivity of comments from 2019 to 2021 followed by a subsequent decline, strongly suggests the influence of significant external events, particularly the Covid-19 pandemic, on user engagement with productivity tools and technologies content on YouTube. This phenomenon provides valuable insights into how such events can impact user interactivity and shape their perceptions.

The sudden surge in interactivity from 2019 (75 comments) to 2020 (350 comments) can be attributed to the onset of the Covid-19 pandemic. As people around the world were forced to adapt to remote work and new lifestyles, there was a heightened demand for information on productivity tools and time management strategies. Content creators responded by producing relevant videos, which led to increased user engagement and discussions in the comments section. By this time, individuals had settled into new routines and were seeking continuous support and guidance on maintaining productivity in the changed circumstances. The higher engagement in 2021 suggests that users were actively seeking, sharing, and discussing insights related to productivity tools, technologies, and time management strategies as they adjusted to the "new normal."

This analysis emphasizes the critical role of external events in shaping online interactions and offers insights into users' adaptive behaviours and changing priorities in response to significant global events.

V. CONCLUSION

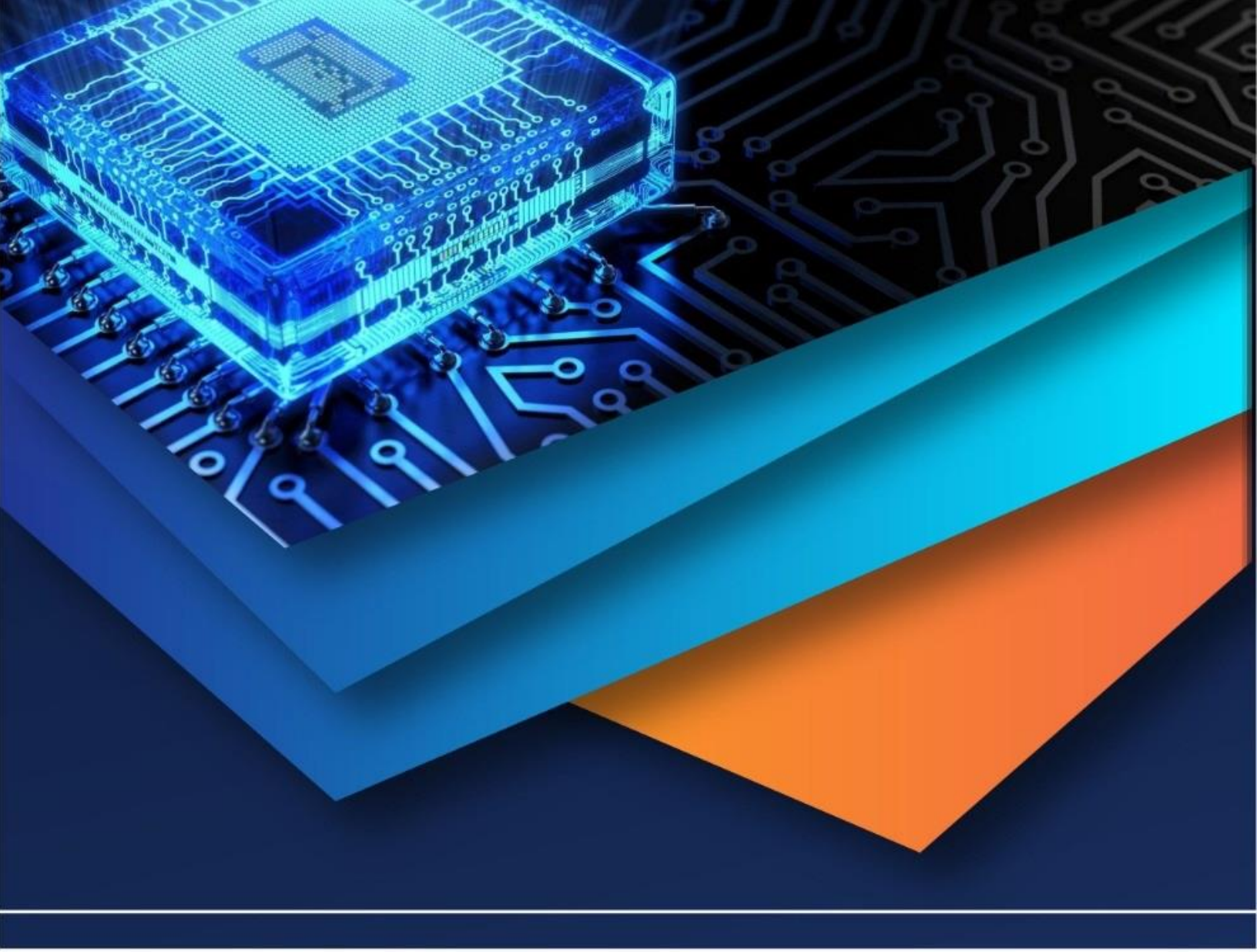
This research delved into the field of sentiment analysis of YouTube comments pertaining to opinions on productivity tools and technologies, utilizing a robust methodology involving data collection through YouTube APIs and sentiment labelling via the VADER framework. The application of machine learning algorithms, specifically Naive Bayes and Random Forest, achieved accuracies of 84.01% and 91.34% respectively, in categorizing sentiments, with the Random Forest algorithm exhibiting higher level of effectiveness in distinguishing negative and neutral comments, indicative of its ability to capture aspects like sarcasm and subtle critique. Notably, the temporal analysis of user interactivity, as depicted in the line chart, showed a noticeable trend of increased engagement from 2019 to 2021, succeeded by a subsequent decline.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

This highlights the significant impact of external factors, especially Covid-19 pandemic, on user perceptions and interactions with content related to productivity tools and technologies on YouTube videos. These findings demonstrate the intricate relationship between significant societal events and online user behavior, helping one understand the transformative impact of such occurrences on digital discourse and user sentiment. This research not only contributes to the field of sentiment analysis but also provides valuable insights into the dynamic interplay between technological trends, user sentiments, and global events, offering a comprehensive perspective on the evolving landscape of users' opinions on productivity tools and technologies.

REFERENCES

[1] R. F. Alhujaili and W. M. S. Yafooz, "Sentiment Analysis for Youtube Videos with User Comments: Review," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 814-820, doi: 10.1109/ICAIS50930.2021.9396049.
[2] M. Aufar, R. Andreswari and D. Pramesti, "Sentiment Analysis on Youtube Social Media Using Decision Tree and Random Forest Algorithm: A Case Study," 2020 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 2020, pp. 1-7, doi: 10.1109/ICoDSA50139.2020.9213078.
[3] Singh, R. and Tiwari, A. (2021) 'YOUTUBE COMMENTS SENTIMENT ANALYSIS ', International Journal of Scientific Research in Engineering and Management (IJSREM), 5(5).
[4] Hanif Bhuiyan et al. (2017) 'Retrieving YouTube Video by Sentiment Analysis on User Comment', IEEE International Conference on Signal and Image Processing Applications (IEEE ICSIPA)
[5] Pai, Aiswarya R, Maria Prince and C. V. Prasannakumar. "Real-Time Twitter Sentiment Analytics and Visualization Using Vader." 2022 2nd International Conference on Intelligent Technologies (CONIT) (2022): 1-4.
[6] C. Solanki, G. Bana, R. Singh, M. K. Goyal, B. Sharan and R. Gupta, "Enhancing Productivity and User Experience with Advanced Notepad: A Comprehensive Study," 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 1120-1123.
[7] H. A. Patrick, P. G. J, M. H. Sharief and U. Mukherjee, "Sentiment Analysis Perspective using Supervised Machine Learning Method," 2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT), Erode, India, 2023, pp. 1-4, doi: 10.1109/ICECCT56650.2023.10179807.
[8] Muhammad Zubair Asghar et al. (2020) 'Sentiment Analysis on YouTube: A Brief Survey', Institute of Computing and Information Technology (ICIT) [Preprint].
[9] Olga Uryupina et al. (2020) SenTube: A Corpus for Sentiment Analysis on YouTube Social Media [Preprint].
[10] G. Veena, A. Vinayak and A. J. Nair, "Sentiment Analysis using Improved Vader and Dependency Parsing," 2021 2nd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2021, pp. 1-6, doi: 10.1109/GCAT52182.2021.9587829.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)