Decoding Social Sentiments : A Study on Social Media Comments Using Machine Learning Techniques

Aastha Gade
School of Engineering and Applied Sciences
University at Buffalo
New York, USA
aastha6.gade@gmail.com

Aayushi Pandey
School of Engineering and Applied Sciences
University at Buffalo
New York, USA
aayuship.221@gmail.com

Aishvarya Salvi
School of Engineering and Applied Sciences
University at Buffalo
New York, USA
aishwarya.salvi28@gmail.com

Avipsa Bhujabal
School of Engineering and Applied Sciences
University at Buffalo
New York, USA
avipsa.bhujabal@gmail.com

***Abstract*— Social media is a highly used world-wide platform which gives access to any kind of information. People often express how they feel about anything very avidly giving us great insights on all the topics. This study explores sentiment analysis of two of the social medias initially – Twitter and YouTube. The project employs many models including Logistic Regression, SVM, Random Forest etc. Our experimentation shows that most of the users use similar words/ images to express their emotions. Experimental results stated that the overall data can be divided into categorical data and numerical data, and then be used for model training. The proposed design approach for social media sentiment analysis helps in understanding user sentiments on a plethora of topics, products.**

Keywords*— *Social Media, Sentiment Analysis, Twitter, YouTube, machine learning, artificial intelligence

I. INTRODUCTION

The evolution of social media platforms has fundamentally changed how audiences consume and interact with content. Platforms like YouTube, TikTok, and Twitter not only allow users to share and view content but also to engage with it through comments and discussions. Understanding the sentiments behind these comments is essential for content creators, marketers, and researchers who wish to connect with their audiences effectively [1]. Through years and years of research on sentiment analysis we have experimented This proposal outlines a comprehensive project aimed at implementing sentiment analysis on comments extracted from these platforms.

Using social media sentiment analysis, we aim to computationally categorize opinions and then classify them on a range of emotions. We have chosen YouTube comments and Tweets because these are the most widely used public broadcasting channels.

With the new advancements in social media and technology, it is apparent that such platforms, which serve a vast population will generate tons and tons of data. This data eventually, when transformed and processed, provides insights which can alter data interpretation entirely. YouTube and Twitter are platforms which offer high amounts of public feedback. YouTube helps us identify positive and negative sentiments in terms of the comments. These comments include texts, GIFs, emojis and many other such visual or literal sentiments. Twitter on the other hand is a platform where information sharing is done in various forms for example videos, audios, texts etc.

II. LITERATURE REVIEW

Many researchers have made a significant contribution in social media sentiment analysis and have published their works. In [1], Shevtsov et al. (2023) found that Twitter communities correlate with YouTube communities, one of the reasons why we chose these two platforms only. [2] Monitoring the Social Media activities is a good way to measure customers loyalty, keeping a track on their sentiment towards brands or products.[3] Panguila and Chandra (2019) evaluated several classifiers and found that "Neural Networks methods such as Multi-layer Perceptron (MLP) and Convolutional Neural Network (CNN) performed better than others classifier in general" for Twitter and consumer review data. Considering this we implemented several other methods to retrieve data but ended up using the ones mentioned in the paper later.

YouTube has emerged as a dominant video-sharing and social media platform, offering diverse content for users of all ages. Its growing influence is challenging traditional media, particularly among younger audiences, as content creators and their videos attract significant viewership and engagement [4]. Many popular videos accumulate a large number of views and comments, which often extend beyond the video's content to various related topics [5].

Sentiment analysis of these video comments has become a focal point for researchers aiming to understand the correlation between user sentiment and video popularity. This analysis employs various methods to explore the relationship between comment sentiment polarity scores and the overall popularity of videos [6]. However, the task of sentiment analysis on YouTube comments presents several challenges for machine learning and neural network models. These challenges include linguistic irregularities, typing errors, niche vocabulary, sarcastic and ambiguous content, and the influence of unrelated events on comment themes [4][5].

Despite these obstacles, sentiment analysis of YouTube comments remains a valuable tool for content creators and researchers alike. It provides insights into audience reception, helps in understanding user behavior, and can potentially predict video performance. As YouTube continues to evolve and compete with traditional media, the ability to accurately analyze and interpret user sentiment will become increasingly important for optimizing content strategies and understanding audience preferences [6].

The authors in [7], have analyzed the customer sentiments in marketplace on twitter platform by performing NLP, CBR , ANN, SVM and API, and have future scoped to develop a web linux application. The study in [8], gives insights on TSA and Opinion Mining [OM] and how the tweet text content can be converted into a feature vector , using N-gram, Part of Speech Tagging (POS). The authors have discussed the various approaches and classifiers for TSA, including Probabilistic, Rule based and Lexicon based classifier, to indicate the strength of the emotions expressed. In [9], the authors have surveyed and provided a comparative study of ML techniques for Opinion Mining, They have investigated that the use of bigram model provides maximum word feature entropy, as it combines two words as a feature.

III. METHODOLOGY

Key features in any form of analysis of YouTube comments would include the following: how YouTube Data API v3 works, authentication through a library named `google-api-python-client`; it basically allows the API to access securely, either by API key or by OAuth 2.0; then there is the data collection, which in this case has been done through the `commentsThreads().list()` method, fetching comments of a certain video-in cases of high volume, pagination. The preprocessed text collected includes tokenization of comments into words, followed by making them uniform in case by changing all the words to lower case, and removal of punctuation and special characters. After that, stop words are removed using NLTK.

Next, the comments will undergo word frequency distributions in order to show what terms are the most frequent-a quantitative look at the content. These are then visualized using the WordCloud library, which builds a word cloud where the size of each word is determined by frequency within the dataset. The cloud is then displayed using `matplotlib`. The methodology has employed the use of the following Python packages: `google-api-python-client` for API interactions, `nltk` for text processing, and `wordcloud` and `matplotlib` for visualization. This guarantees that the analysis of the comments on YouTube is orderly and productive.

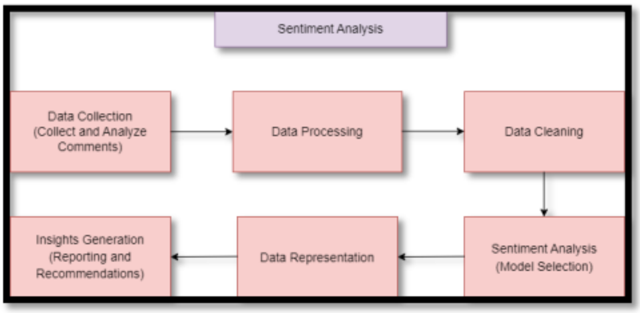
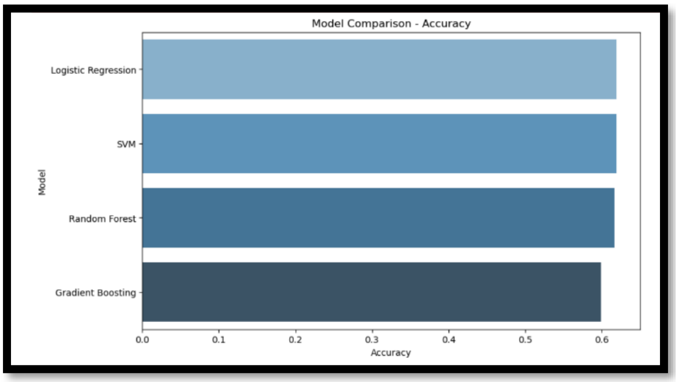


Fig. 1. System block diagram

A. *Youtube*

The code will interface directly with the YouTube Data API v3 to collect and analyze comments on YouTube videos. This is a very new area in which social media analytics and NLP have become very important, especially because many civic discussions are nowadays carried out with the help of digital platforms. Other works approached different types of analyses, like sentiment analysis, which classifies comments by machine learning models into positive, negative, and neutral sentiments; topic modeling, such as Latent Dirichlet Allocation, in finding major themes; network analysis of structure in comment threads and the interaction between users; hate speech and toxicity detection as a means of counteracting online harassment and misinformation; and multimodal analysis, putting comments together with video content for deep insights into viewer engagement. For this type of research effort, the YouTube Data API, which allows for the extraction of large-scale comments for analysis, is indispensable.

This code does everything in order: from API authentication, it collects the data with commentsThreads().list(), followed by text preprocessing like tokenizing, converting to lowercase, removing punctuation, and removal of stop words via NLTK. The analysis would work towards developing word frequency distributions to view the most frequent terms present in the comments. Words can be visualized effectively with the help of a WordCloud library as this represents words in word-cloud format. Each word's size will describe the occurrence frequency of the word in the comments. Basically, this process depends upon a few python packages: google-api-python-client for accessing the Youtube API, nltk for text processing, wordcloud for visualization, and matplotlib for showing the word cloud.



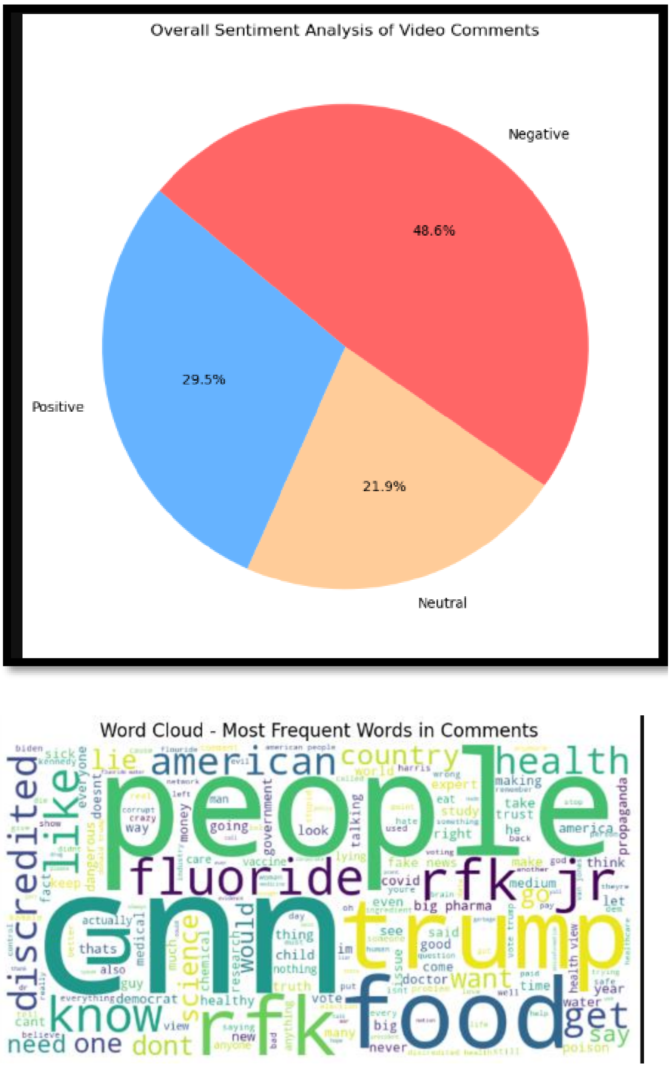


Fig. 3. Data visualizations for Youtube Platform Sentiment Analysis.

B. *Twitter Platform Sentiment Analysis*

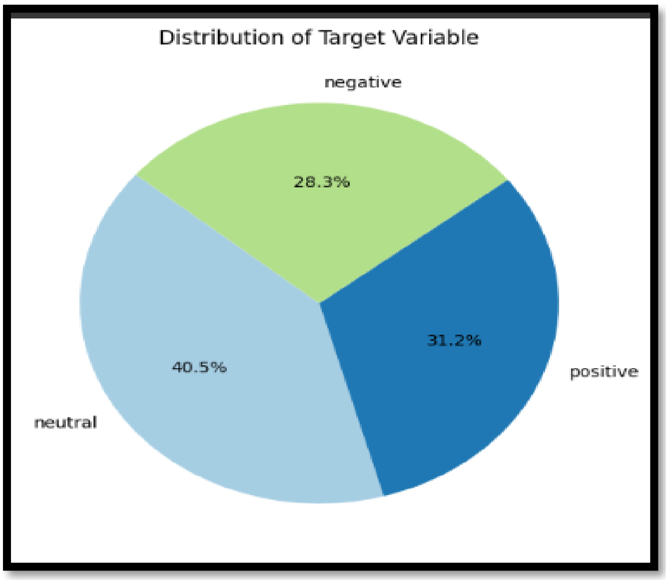
The experiment is conducted on a publicly available Twitter dataset containing historical data consisting of tweets and corresponding sentiment labels (e.g., positive, negative, neutral), having 27482 entries. Tweets were preprocessed to remove noise by removing URLS, user mentions, special characters/numbers and stopwords. To standardize data, the text was converted to lowercase.

The tweets were tokenized into individual words, and padding was applied to ensure that all sequences had the same length and then the text sequences were encoded into numbers. By trial and error, we had also implemented the TextBlob pre-trained sentiment analysis model was used to compute sentiment scores for each tweet.

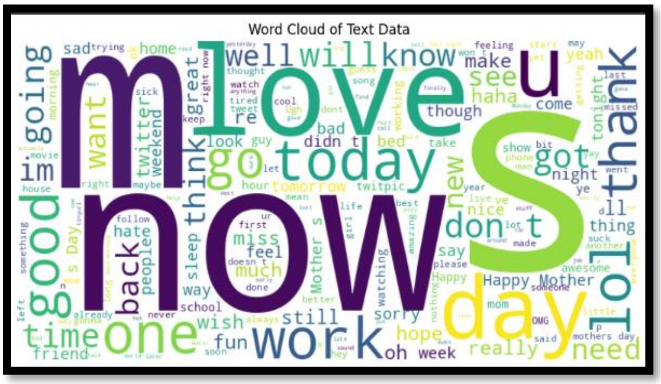
The dataset was split into training [80%] and testing [20%] sets and the ML models were trained on the processed data to classify tweet sentiment.

For model training, we had used four different machine learning techniques, LSTM, Logistic Regression, Random Forest and Support Vector machine [7]. These models were selected on the basis of their ability to capture word order-context in sequences, modeling sentiment probabilities, reducing overfitting and offering strong generalization for text classification [8].

The models were compared and evaluated using Train-Test accuracy, precision and recall scores. Fig. 4. Give the visual data representation of the twitter sentiment analysis. We have used word cloud, which represents the most frequently occurring word in the tweets.



(a)



(b)

Fig. 4. Data visualizations for Twitter Platform Sentiment Analysis.

The Fig. 5. shows the sentiment score frequency graph and we can infer that the neutral sentiment has the highest frequency, compared to the positive and negative sentiments.

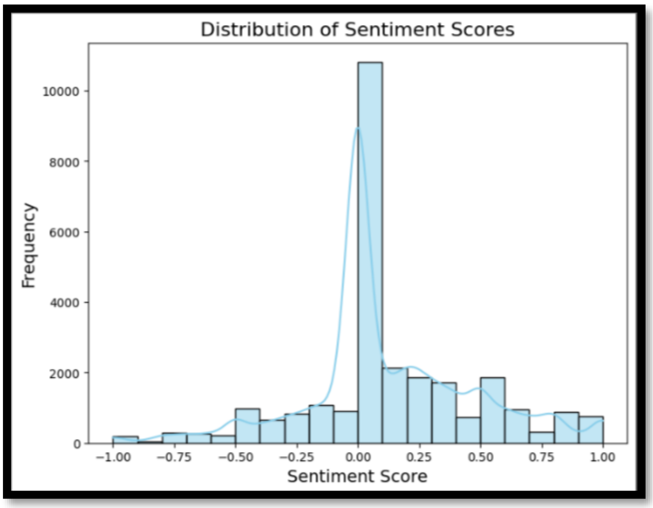


Fig. 5. Sentiment scores frequency distribution graph.

IV. RESULT ANALYSIS

The inference and illustration of the results after the Model evaluations on both platforms , Youtube and Twitter :

(1) Youtube

The code here has given a skeleton approach towards the analysis of YouTube comments using the YouTube Data API together with some NLP. In reality, there's no output for this script as it is, but the structure here provides the framework on which quite a number of analyses can stand delivering insight. More typical results from such analyses are word frequency distributions that allow the researcher or analyst to better understand the most common terms within the comments and provide a quantitative sense of what themes or topics are being discussed. In the case of a product review video, for example, common terms might be "price," "quality," or other features about which the audience would most be interested. Word clouds express these frequent words in a much more intuitive way and are often visually appealing; the relative size of the words shows their respective frequencies within the dataset. That gives at once the main subjects discussed and provides the general feeling or key debating point for the viewers. For instance, an adjective such as "amazing" might be larger in the center of the word cloud of positive responses, where technical terms can show areas of detailed discussion.

Analyses of this type highlight significant features in a variety of ways: they identify major themes and topics, showing prima facie those issues or ideas that have secured most attention. Second, they give insight into audience sentiment-or the predominance of positive/negative words. Without performing all of the sentiment analysis, it will show in the feedback how the video has been received. Third, they may detect novel trends or concerns. Through the frequency of unusual or new terms, areas of interest come to light that have not been anticipated. Moreover, content relevance analysis can be done through comparing the most occurring words with the subject of the video, which is important for analysts to determine whether the content of the video serves the viewers' expectations.

It all depends on the volume and quality of comments that come in, mainly; the more comments, generally speaking, the more reliable and meaningfully nuanced the insights. Further, preprocessing-stop-removal of punctuation must be very neat and clean so that useful contexts or major terms are not lost in this process. For example, development of stop word list based on the context of the video will bring considerable improvement in the quality of analysis. Lastly, though the word frequencies and word cloud result are enlightening, yet it cannot do away with human interpretations as far as finding out the underlying contexts and implications of results are concerned. Yet the present structure was no more than a skeleton that might be fleshed out at least in a couple of directions: basic sentiment analysis could classify comments into positive, negative, or neutral, to disaggregate the attitudes of an audience; topic modelling through LDA giving a much richer analysis digging out themes and subtopics not so clearly visible from the word frequencies.

This will provide the ability for network analysis to mine threads of comments and user interactions in support of identifying influential commenters and flows of information within the community. Content moderation mechanisms will be able to be devised for the automatic detection and flagging of harmful contents, including hate speech, spam, or misinformation.

Other probable extensions might include the provision of a comparative analysis, enabling the code to identify and compare comments between videos or channels in such a way as to highlight patterns of audience engagement and sentiment in regard to different content types. A time-series analysis could be implemented, for instance, understanding how comments and their sentiments change over time and how discussions change once the number of views for the video increases. Such an analysis of comments related to what was really happening in the video would provide such insight into the relations between given parts of the video and user responses. An interface, for instance, would let non-technical users, for example, content creators or marketers, do such an analysis on their own, which is going to make this tool all that more usable and practical for realistic applications. Though currently a limited framework, it is very useful for anyone interested in deep digging into comments on YouTube. It forms a very important basis for higher-level analysis in the provision of insight into online communities and audience engagement. With social media, such as YouTube, continuing to be at the heart of public opinion, consumer behavior, and setting cultural trends, this kind of tool is going to become increasingly important for a researcher, content creator, and marketer who wants to understand the beat of his audience and the shifting dynamics of online conversations.

```
Model Comparison - Accuracy Results:
Logistic Regression - Training Accuracy: 0.9340, Test Accuracy: 0.6198
SVM - Training Accuracy: 0.9497, Test Accuracy: 0.6198
Random Forest - Training Accuracy: 1.0000, Test Accuracy: 0.6172
Gradient Boosting - Training Accuracy: 0.8468, Test Accuracy: 0.5990
```

```
Evaluating Model: Logistic Regression
Accuracy: 0.6198
Class Negative: Sensitivity = 0.9309, Specificity = 0.9459
Class Neutral: Sensitivity = 0.3707, Specificity = 0.0435
Class Positive: Sensitivity = 0.2500, Specificity = 0.8600
--------------------------------------------------
Evaluating Model: SVM
Accuracy: 0.6198
Class Negative: Sensitivity = 0.8457, Specificity = 0.8933
Class Neutral: Sensitivity = 0.4138, Specificity = 0.1471
Class Positive: Sensitivity = 0.3875, Specificity = 0.8276
--------------------------------------------------
Evaluating Model: Random Forest
Accuracy: 0.6172
Class Negative: Sensitivity = 0.7553, Specificity = 0.8875
Class Neutral: Sensitivity = 0.3276, Specificity = 0.3544
Class Positive: Sensitivity = 0.7125, Specificity = 0.5846
--------------------------------------------------
Evaluating Model: Gradient Boosting
Accuracy: 0.5990
Class Negative: Sensitivity = 0.8564, Specificity = 0.8750
Class Neutral: Sensitivity = 0.4397, Specificity = 0.0645
Class Positive: Sensitivity = 0.2250, Specificity = 0.8793
--------------------------------------------------
```

(2) **Twitter**

As shown in Table 2, logistic regression model performed well compared to other models, showcasing a high train and test accuracy of 0.8178 and 0.7382, respectively.

From inference, we can also observe that the LSTM [Long Short Term Memory] Model has underperformed despite being a deep learning model, this is due to presence of insufficient data, and hence it is overfitting on a smaller dataset. To rectify it, we had performed parameter tweaking, increasing model complexity, adding dropout layers and decreasing no. of epochs.

The classification results say that the Logistic regression and random forest has performed well across all classes[9].

Parameters	Models			
	LSTM	Logistic Regression	Support vector machine	Random Forest
Train Accuracy	**0.9402**	0.8178	0.7861	0.9981
Test Accuracy	**0.65203**	**0.7382**	0.69996	0.7211
Precision	0.64	0.76	0.73	0.75
	0.62	0.67	0.64	0.65
	0.70	0.79	0.79	0.84
Recall [Sensitivity]	0.62	0.64	0.57	0.57
	0.61	0.78	0.78	0.83
	0.74	0.77	0.72	0.71

V. CONCLUSION

From our presented work it can be concluded that we can get an idea of what are sentiments of people in a certain scenario. Like in the YouTube comments analysis and for twitter analysis we get different perspectives on the same topic. We tried to compare how different age groups, gender, cultural differences etc have different kinds of opinions on the same topic. Using the word cloud, bar chart and histogram visualizations we were able to identify the emotions of how people are really feeling about a certain topic of discussion or product. We can also refine our project further to get multilingual inputs and integrate real-time sentiment analysis so as to give a more accurate and in-real time analysis. Our study focused on comparing various models like Random Forest, SVM, Gradient Boosting to get our output which explains that different kind of data worked properly with different machine learning models. It not only depicts practical application of machine learning concepts, but also showcases the potential of sentiment analysis in content moderation, businesses and product owners.

REFERENCES

[1] Shevtsov A, Oikonomidou M, Antonakaki D, Pratikakis P, Ioannidis S. What Tweets and YouTube comments have in common? Sentiment and graph analysis on data related to US elections 2020. PLoS One. 2023 Jan 31;18(1):e0270542. doi: 10.1371/journal.pone.0270542. PMID: 36719868; PMCID: PMC9888715.

[2] Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., & By, T. (2012). Sentiment Analysis on Social Media. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 919-926.

[3] Broklyn, Peter and Olukemi, Ayoolu and Bell, Chris, Social Media Sentiment Analysis for Brand Reputation Management (July 10, 2024). Available at SSRN: https://ssrn.com/abstract=4906218 or http://dx.doi.org/10.2139/ssrn.4906218

[4] 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*), Mar. 2021, doi: https://doi.org/10.1109/icais50930.2021.9396049.

[5] 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)*, Mar. 2021, doi: https://doi.org/10.1109/icais50930.2021.9396049.

[6] R. Pradhan, "Extracting Sentiments from YouTube Comments,*" IEEE Xplore*, Nov. 01, 2021. https://ieeexplore.ieee.org/document/9702561

[7] A. Sarlan, C. Nadam, and S. Basri, "Twitter sentiment analysis," in *Proceedings of the 2014 International Conference on Information Technology and Multimedia (ICIMU)*, Putrajaya, Malaysia, Nov. 2014, pp. 212-216, doi: 10.1109/ICIMU.2014.7066632.

[8] Y. Wang, J. Guo, C. Yuan, and B. Li, "Sentiment Analysis of Twitter Data," *Applied Sciences*, vol. 12, p. 11775, Nov. 2022, doi: 10.3390/app122211775

[9] V. A. Kharde and S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," *International Journal of Computer Applications*, vol. 139, no. 11, pp. 5-15, Apr. 2016, doi: 10.5120/ijca2016908625.

[10] Twitter Dataset: [Online]. Available: https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis (Accessed online- December 2024).