Classification Of Societal Sentiments And Its Domain By Machine Learning Techniques

Preetha V[1], Sri Hari S[2], Yogeshwaran S T[3], Suriya N[4]
[1]Assistant Professor, Dept of CSE
[2, 3, 4]Dept of CSE
[1, 2, 3, 4] Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu, India

Abstract- *In the current digital era, social media and online platforms have developed into effective means for people to voice their thoughts, feelings, and opinions regarding a range of societal issues. Enhancing decision-making processes, promoting social well-being, and influencing public policy all depend on the analysis and comprehension of these popular attitudes. The goal of the project proposal is to develop and apply machine learning techniques for the classification of societal sentiments and the domains in which they are expressed. The proposed study intends to categorize online text data into sentiment categories, such as positive, negative, neutral, or specific emotions, by utilizing machine learning algorithms and natural language processing (NLP). In addition, the project aims to pinpoint the areas or subjects that these feelings are associated with, offering important perspectives on the most urgent social concerns. The project's identification of societal sentiments will find wide-ranging applications, such as sentiment analysis of public discourse, social trend tracking, public opinion monitoring on policy, and evaluating the influence of societal events on public sentiments. This research will provide researchers, policymakers, and social advocates with the necessary tools to make well-informed decisions and develop successful strategies for addressing societal concerns by precisely classifying sentiments and assigning them to particular domains. In the end, the project's proposal represents a significant advancement in the use of NLP and machine learning to better understand and address societal sentiments, making the world a more educated, understanding, and responsive place.*

Keywords- Machine Learning, Ensembled Algorithm, Sentiment Classification, Domain Classification

I. INTRODUCTION

A. *Motivation*

The introduction of social media has caused unheard-of changes in the ever-changing worlds of sports, movies, and technology. It has become common for enthusiasts to share updates, express emotions, and voice their opinions on virtual forums such as Facebook, Instagram, and X (formerly known as Twitter) [1] [5]. It is now crucial to glean valuable insights from user-generated content amid this digital noise [2]. Social media acts as a real-time data repository, providing insightful peeks into the pulse of society at critical junctures [3].

There has been a rise in online discussions about sports, movies, and technology, with fans, critics, and business insiders all adding to the story. Social media has emerged as the preferred forum for interaction, whether it is for discussing the most recent match results, analyzing new releases, or arguing about technological innovations [4] [5]. A dynamic forum for discussion and engagement, X is one of these platforms that sports, movie, and tech enthusiasts use frequently [6]. Examining sports, movies, and tech-related content on X provides a deep understanding of the views and opinions that are prevalent in the community [7].

The examination of content about sports, movies, and technology provides important insights into the attitudes and patterns that are dominant in the corresponding fields [8] [11]. Words such as "thrilling victory," "disappointing performance," "blockbuster hit," "cutting-edge technology," and "game-changing innovation" capture the wide range of opinions shared on social media [12]. Moreover, opinions and trends in the corresponding fields can be shaped by public conversation about news about sports, movies, or technology [11].

There is an urgent need for automated methodologies to analyze and classify sentiments within the sports, movie, and technology landscapes due to the large amount of social media data [9] [10]. Automating sentiment analysis with machine learning (ML) techniques is a promising way to gain insights into the attitudes and emotions that are prevalent in the community.

For improved classification accuracy, our method uses a hybrid framework that makes use of ensemble learning strategies like Support Vector Machines (SVM), Random Forest, and Naive Bayes algorithms.

Our ensemble model leverages the complementary advantages of Random Forest's ensemble learning methodology, SVM's ability to discriminate, and Naive Bayes' ease of use and effectiveness in classification tasks. We optimize hyperparameters to guarantee optimal performance through rigorous training and validation procedures, thereby establishing a new benchmark for automated sentiment and domain classification in the dynamic domains of sports, film, and technology.

B. Related Work

Sentiment analysis has advanced recently, driven by the use of social media data to gauge public opinion during the COVID-19 pandemic. Notably, Garcia, Klaifer, and Berton [2] carried out a thorough analysis of Twitter discourse related to COVID-19, focusing on sentiment trends in Brazil and the USA. Their research, which looked at topic identification as well as sentiment analysis, provided insightful information about the emotional dynamics surrounding the pandemic.

In a similar vein, Imran, Daudpota, Kastrati, and Batra [3] investigated emotion detection and cross-cultural polarity on COVID-19-related tweets by employing sentiment analysis and deep learning techniques. Their research provided a nuanced understanding of how sentiments vary across cultures and regions during the pandemic, illuminating the intricate relationship between cultural factors and emotional reactions. It was published in IEEE Access.

Additionally, new research highlights the value of infodemiology in tracking conversations about COVID-19 on social media sites like Twitter. For example, Jang et al. [4] carried out an infodemiology study in North America, using aspect-based sentiment analysis and topic modeling to investigate the changing conversation about COVID-19. Their research provided insightful information about the dynamics of sentiment and information flow in the area.

Furthermore, more complex sentiment analysis techniques can now be developed thanks to recent developments in machine intelligence. Whang and Vosoughi [9], for example, demonstrated how well BERT-based models classified instructional tweets about COVID-19. They were able to understand the subtle sentiment and informative content in these tweets by using cutting-edge natural language processing techniques. Interestingly, their research was carried out as a component of the WNUT-2020 task at Dartmouth CS.

New research shows how sentiment analysis is developing about the COVID-19 pandemic. By employing sophisticated machine learning models and conducting a thorough examination of social media data, scientists have achieved noteworthy progress in comprehending public sentiments, augmenting our comprehension of societal reactions to worldwide health emergencies [2].

Furthermore, the influence of COVID-19 on community sentiment dynamics has been the subject of recent research. For example, Zhou et al. [12] studied sentiment dynamics in a state in Australia through a case study to comprehend how community sentiment changed throughout the pandemic. Their research shed important light on the variables affecting community opinion as well as how opinions change over time in reaction to shifting conditions.

Furthermore, current studies are investigating novel methods of sentiment analysis in relation to the COVID-19 pandemic. Saranya and Usha [8] have created a machine learning technique for sentiment analysis on Twitter that makes use of Intelligent Word Net Lemmatize. Their goal is to improve sentiment classification's precision and effectiveness. Considering the ongoing difficulties caused by the pandemic, their methodology adds a great deal to the field by offering new perspectives on sentiment analysis methods.

C. Contribution

In our project context, this study advances the fields of sentiment analysis and domain classification through the introduction of an innovative hybrid classification pipeline tailored specifically for analyzing sentiments and domains expressed on microblogging platforms. Our primary contributions include the development of a comprehensiveframework focused on domains such as sports, movies, and technology. Our main contributions are listed as follows:

Architectural Innovation: We've developed a brand-new hybrid framework that combines cutting-edge sentiment and domain classification machine learning methods. This novel method offers a more flexible and adaptive solution for tweet classification by automating the sentiment labeling process without depending on lexicon-based techniques or tools like VADER.

Performance Evaluation: We used a variety of metrics, including accuracy, precision, recall, and F1 score, to conduct thorough performance evaluations to verify the efficacy of our hybrid framework. We further the broader understanding of AI applications in sentiment analysis [6] by comparing the performance of various machine learning techniques in sentiment and domain classification.

Enabling Public Sentiment Understanding: Our automated sentiment classification system is essential for enabling public sentiment understanding regarding a variety of topics in the film, sports, and technology domains. Stakeholders like marketers, analysts, and enthusiasts can learn a great deal about the prevalent attitudes and opinions in these particular areas of interest by classifying tweets according to their sentiments and domains.

An illustrative example would be the ability to classify tweets about technology, sports, and movies according to the sentiments expressed in them. This would enable customized marketing, audience engagement, and content creation strategies. Stakeholders can improve user satisfaction as well as engagement by effectively responding to audience preferences and concerns by having a thorough understanding of the sentiment dynamics within each domain.

By conducting research, we hope to contribute to the fields of sentiment analysis and domain classification and offer insightful information about public opinion in a variety of fields, such as technology, sports, and movies. We provide stakeholders with actionable insights for strategic planning and well-informed decision-making through our hybrid classification pipeline, which provides a strong framework for automated tweet classification.

II. AUTOMATED HYBRID TWEET CLASSIFICATION

A. Data Preprocessing

After obtaining the unprocessed textual tweet data, we initiated the preprocessing phase utilizing techniques from Natural Language Processing (NLP) to enhance the content by removing redundancies and anomalies. The subsequent steps involved:

Eliminating Missing and Null Rows: Before performing any additional processing, we thoroughly inspected the dataset to find and remove any missing or null rows, guaranteeing the accuracy and completeness of the data.

Stop Word Removal: We utilized the Python-based NLP toolkit NLTK to systematically eliminate stop words such as 'over,' 'under,' 'again,' 'further,' 'then,' 'once,' 'here,' and 'there.' Additionally, we excluded URLs and removed user mentions, hashtags, special characters like asterisks (*), dollar signs ($), exclamation marks (!), and applied regular expressions (RegEx) to cleanse the text.

Case Folding: We standardized the cases throughout the text because we knew that word case affected polarity (positive, negative, or neutral). Since tweets are primarily written in lowercase, we performed case-folding, which involved changing a small number of uppercase words to lowercase in every tweet.

Tokenization: We broke up each tweet sentence into individual words, or tokens, using the tokenization process, which made it easier for machine learning algorithms to understand the semantic meaning of each word.

Lemmatization: A fundamental aspect of Natural Language Processing, lemmatization involves identifying the root forms of words, commonly referred to as lemmas, to categorize various word forms into a single unit. For instance, the lemma "stop" groups together terms like "stopped," "stopping," and "stops," which are derived from the base word "stop." This technique enhances the efficiency of machine learning processes.

B. Feature Extraction

In tasks related to natural language processing (NLP) and machine learning, such as sentiment analysis and domain classification, feature extraction is an essential step. It entails converting unprocessed text data into numerical characteristics that are comprehensible and processable by machine learning algorithms. Within the framework of our project, feature extraction includes the following methods:

CountVectorizer: This method turns a set of text documents into a matrix of token counts, with a document for each row and a word for each column representing a distinct word in the corpus. The frequency of every word in the text is captured by CountVectorizer, which is used to create a "bag of words" representation.

Term Frequency-Inverse Document Frequency: This method compares a word's frequency within a document to its frequency throughout the corpus of documents to determine the significance of that word. By giving words that are common in one document but uncommon in the entire corpus more weight, this approach captures their uniqueness in context.

C. Sentiment Analysis

We preprocess the raw textual data extracted from tweets using natural language processing techniques in our sentiment analysis process [8]. This involves several steps, including the removal of stop words, special characters, URLs, user mentions, and hashtags. Additionally, we standardize text cases through case folding [1] and perform lemmatization [8]

to reduce words to their basic forms. Subsequently, the textual data is transformed into numerical representations suitable for machine learning models [8] using feature extraction techniques such as CountVectorizer [2]. We utilize supervised learning algorithms, such as Support Vector Machine (SVM) [3], Random Forest [3], and Naive Bayes [3] [8], for categorizing sentiments into positive, negative, and neutral categories.

D. Domain Analysis

Similar to this, to guarantee data cleanliness and integrity, we start our domain analysis process by preprocessing the raw textual data [8]. This entails actions such as text case standardization and noise reduction [1]. After preprocessing, we convert the textual data into numerical features appropriate for machine learning models [8] by utilizing feature extraction techniques such as CountVectorizer [2]. Next, we use machine learning algorithms to categorize tweets into their corresponding domains, such as sports, movies, and technology [8]. These algorithms include Support Vector Machines (SVM) [3], Random Forest [3], and Naive Bayes [3]. The selection of these algorithms is predicated upon their efficacy and versatility in domain classification assignments [3]. Furthermore, our framework employs ensemble techniques, such as combining SVM, Random Forest, and Naive Bayes models, to improve the robustness and accuracy of domain classification [3].

E. Data Labeling

The first step in our workflow is to label tweets appropriately for sentiment and domain classification. This labeling procedure simplifies the use of supervised machine-learning techniques later on. Our method ensures a smooth and effective process by automating this labeling without the need for conditional statements. Labels 0, 1, and 2 represent the positive, neutral, and negative classes into which tweets are classified for sentiment labeling. Similarly, tweets are labeled with corresponding categories to enable precise classification, such as movies (labeled 1), sports (labeled 2), and technology (labeled 4). Subsequent classification tasks are more reliable and effective as a result of this systematic labeling process.

F. Classifiers

We use a variety of machine learning classifiers in our approach to sentiment and domain classification, each one specifically designed for the task at hand and the features of the dataset. We use a range of algorithms for sentiment classification, such as Random Forest [5], Support Vector Machine (SVM) [9], Naive Bayes [1], and Logistic Regression [8]. Because these classifiers can handle high-dimensional data and capture intricate relationships between features and sentiment labels, they are a good fit for sentiment analysis tasks [11]. To improve sentiment classification's overall performance and robustness, we also use ensemble approaches that combine several classifiers, such as Random Forest, SVM, and Naive Bayes [7].

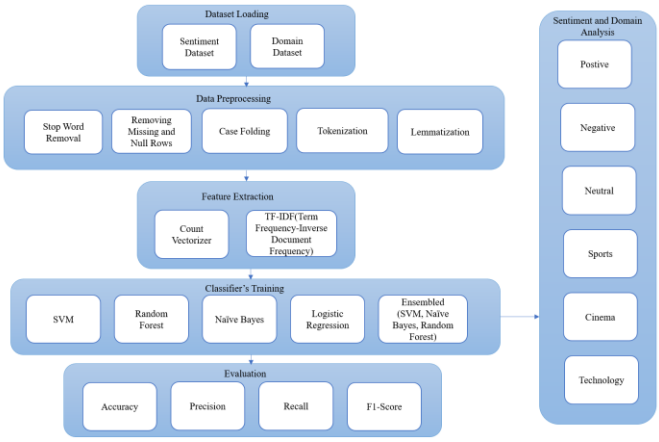

Fig 1: Hybrid Classification Framework

Similar machine learning algorithms, such as Support Vector Machines (SVM) [9], Random Forest [5], Naive Bayes [1], and Logistic Regression [8], are used for domain classification and are customized to the demands of domain classification tasks [2]. These classifiers are selected due to their capacity to handle textual data efficiently and their flexibility in solving multi-class classification problems [4]. Additionally, ensemble approaches are used to increase domain classification accuracy and dependability, with combined models offering complementary strengths and reducing individual classifier biases [6]. Our methodseeksto obtain precise and dependable sentiment and domain classification outcomes within the framework of our particular project specifications by employing these classifiers.

III. PERFORMANCE EVALUATION

A. Data Collection and Computional Environment

We use two different datasets in our model, one for sentiment and the other for domain classification. The sentiment dataset was obtained from Kaggle.com [10] and is made up of a wide range of tweets that convey different emotions. About 60,000 tweets that were taken from different sources make up this dataset, which offers a large corpus for sentiment analysis. Conversely, the domain dataset is curated by hand and only includes tweets about the tech, sports, and

movie domains. The source of these tweets is ChatGPT, which yielded a dataset of about 900 tweets that were classified into these domains.

Our model is developed and run on laptops that have an Intel Core i5 processor, a 256GB SSD and a 1TB HDD for storage, and NVIDIA graphics cards with 2GB of dedicated memory and 8GB of RAM. Although our laptops lack the processing power and storage capacity of server-grade infrastructure, they still have enough power to perform the sentiment and domain classification tasks effectively. We can perform experiments, train machine learning models, and assess classification performance in a portable and affordable way by utilizing the processing power of these laptops.

B. Performance Metrices

To evaluate the accuracy and reliability of our sentiment and domain classification models, we use a set of common performance metrics. These metrics offer important information about how well the models classify tweets according to their sentiment and domain categories.

TABLE I. EVALUATION MEASURE SCORES FOR MODEL PERFORMANCES

Models	Accuracy	Recall	Precision	F1 Score
Naive Bayes (Domain)	94%	94%	94%	94%
Random Forest (Domain)	94%	94%	94%	94%
SVM (Domain)	95%	95%	95%	95%
Logistic Regression (Domain)	93%	93%	93%	93%
Ensembled-Naive Bayes, Random Forest, SVM (Domain)	95%	95%	95%	95%
Naive Bayes (Sentiment)	75%	75%	76%	75%
Random Forest (Sentiment)	91%	91%	91%	91%
SVM (Sentiment)	85%	85%	85%	85%
Logistic Regression (Sentiment)	51%	51%	51%	50%
Ensembled-Naive Bayes, Random Forest, SVM (Sentiment)	93%	93%	93%	93%

The classification system's accuracy is computed as the proportion of correctly classified samples (including both positive and negative samples) to the total number of samples:

$$\text{Accuracy} = (TP + TN)/(TP + FP + FN + TN)$$

Precision measures the relevance of the classification outcomes, reflecting the percentage of accurately predicted positive samples among all classified as positive:

$$\text{Precision} = TP/(TP + FP)$$

Recall, or sensitivity, measures the proportion of true positive samples correctly identified among all actual positive samples. It assesses a model's ability to capture all relevant positive instances in a dataset:

$$\text{Recall} = TP/(TP + FN)$$

The F1-score combines precision and recall into a single metric to provide a fair evaluation of a model's performance. The harmonic mean of recall and precision is used to calculate it:

$$\text{F1 score} = (2(\text{precision} \times \text{recall}))/(\text{precision} + \text{recall})$$

C. Results

Sentiment Labeling: The distribution and character of sentiments expressed in tweets across various domains were clarified by our sentiment labeling analysis. We found that tweet lengths varied between domains and that tweets about movies typically had more in-depth information because of the complex conversations surrounding movies and related subjects. On the other hand, tweets about sports were typically shorter and better reflected the brisk nature of sports commentary. A wide variety of sentiments were found in the sentiment analysis, with a significant percentage displaying positive or neutral sentiments.

The tweet length distribution is depicted in Fig. 2, which shows how tweet lengths vary amongst different domains. This distribution shows possible variations in communication styles and content density and offers insights into the composition of tweets within each domain.

Moreover, word clouds corresponding to every sentiment class are shown in Fig. 3, where high-frequency words connected to positive, neutral, and negative sentiments are displayed. These word clouds show clear patterns in sentiment expression by providing a visual depiction of the most common themes and subjects within each sentiment class. This visualization offers important insights into the sentiment trends seen in the tweets that were analyzed and helps to understand the underlying sentiment dynamics within different domains.

Additionally, the word clouds from the domain and sentiment datasets are displayed in Figures 4 and 5, respectively. The aforementioned visualizations provide supplementary viewpoints on the salient themes and subjects present in each dataset, thereby augmenting our comprehension of the distribution of content and sentiment across diverse domains.

ML Performance: We assessed how well different machine learning models performed for tasks involving sentiment and domain classification. The domain classification accuracy scores of 94%, 94%, 95%, and 93% were attained by

the Naïve Bayes, Random Forest, SVM, and Logistic Regression classifiers. Interestingly, the ensemble model with Random Forest, SVM, and Naïve Bayes produced a high accuracy of 95%. The sentiment classification accuracy scores for the Naïve Bayes, Random Forest, SVM, and Logistic Regression models were 75%, 91%, 85%, and 51%,

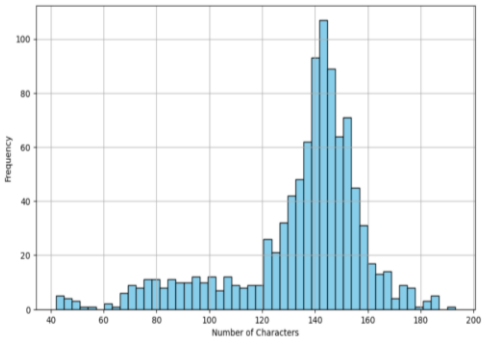

Fig 2: Tweets length distribution – Domain dataset

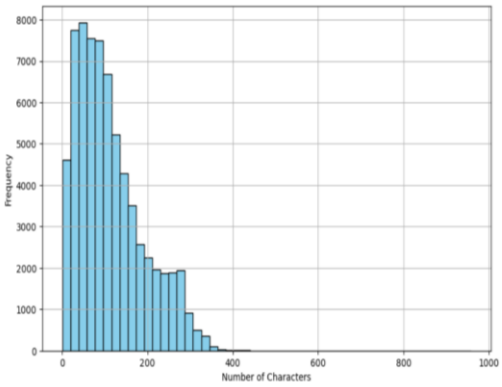

Fig 3: Tweets length distribution – Sentiment dataset

Fig 4: Word cloud for domain dataset

Fig 5: Word cloud for sentiment dataset

in that order. With an accuracy of 93%, the ensemble model that combined SVM, Random Forest, and Naïve Bayes showed competitive performance.

Table I provides an overview of these methods performance, demonstrating how well they work for automated sentiment and domain classification. The outcomes highlight how reliable machine learning algorithms are at correctly classifying tweets according to their respective sentiments and domains. SVM stood out as the best performer in domain classification, obtaining a remarkable 95% accuracy rate. On the other hand, Random Forest performed better in sentiment classification, with an accuracy of 91%. Using Random Forest, SVM, and Naïve Bayes, the ensemble model produced competitive and consistent results in both domain and sentiment classification tasks, proving the effectiveness of ensemble techniques in raising classification accuracy. These results highlight how important it is to use a variety of machine learning models and ensemble methods to efficiently analyze and categorize tweets based on their domains and sentiments.

IV. CONCLUSION AND FUTURE WORK

Finally, by utilizing machine learning techniques to precisely classify tweets into various domains and identify sentiments as positive, negative, or neutral, our project has demonstrated a strong framework for tweet domain and sentiment classification. We demonstrate the efficacy of ensemble methods in improving classification performance by achieving high accuracy scores with classifiers like Naïve Bayes, Random Forest, Support Vector Machines (SVM), and Logistic Regression.

Moreover, the integration of ensemble techniques, which merge the forecasts from various classifiers, has significantly enhanced the general precision and dependability of our classification system. Ensemble approaches have shown better results in capturing the subtleties of tweets across

various domains and sentiment classes by utilizing the advantages of individual classifiers and minimizing their drawbacks.

There are a number of interesting directions that future research and development could go. Taking our analysis a step further and incorporating domains other than technology, sports, and movies could yield a more thorough understanding of sentiment dynamics in a wider range of topics. Including fields like finance, politics, and health could provide insightful information about public opinion in a range of social contexts.

Furthermore, our next efforts will concentrate on implementing our classification framework in practical applications like market research and social media monitoring. Organizations can improve their customer engagement strategies and make data-driven decisions by incorporating sentiment analysis into their decision-making processes. This allows them to obtain insightful information about market trends, customer opinions, and brand perception.

Finally, research will continue to investigate how ensemble methods and model refinement can be further advanced to improve classification accuracy and adaptability to changing user behaviors and language trends. We can make sure that our sentiment analysis framework remains relevant and useful for identifying and analyzing the emotions expressed in online conversations by keeping up with the most recent developments in machine learning and natural language processing.

To sum up, our project on domain and sentiment classification establishes the groundwork for utilizing machine learning methods to obtain practical understanding from social media information. Our goal is to push the boundaries of sentiment analysis further via continued research and innovation, enabling institutions to make better decisions based on data when interpreting and addressing public opinion.

REFERENCES

[1] Bilgin, Metin, and İzzet Fatih Şentürk. "Sentiment analysis on Twitter data with semi-supervised Doc2Vec." 2017 international conference on computer science and engineering (UBMK). Ieee, 2017.

[2] Garcia, Klaifer, and Lilian Berton. "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA." Applied soft computing 101 (2021): 107057.

[3] Imran, Daudpota, Kastrati and Batra, "Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets," in IEEE Access, vol. 8, pp. 181074-181090, 2020, doi: 10.1109/ACCESS.2020.3027350

[4] Jang, Hyeju, et al. "Tracking COVID-19 discourse on twitter in North America: Infodemiology study using topic modeling and aspect-based sentiment analysis." Journal of medical Internet research 23.2 (2021): e25431.
Khan, Rijwan, Piyush Shrivastava, Aashna Kapoor, Aditi Tiwari, and Abhyudaya Mittal. "Social media analysis with AI: sentiment analysis techniques for the analysis of twitter covid-19 data." J. Crit. Rev 7, no. 9 (2020): 2761-2774.

[5] Kruspe, Anna, Matthias Häberle, Iona Kuhn, and Xiao Xiang Zhu. "Cross-language sentiment analysis of european twitter messages during the covid-19 pandemic." arXiv preprint arXiv:2008.12172 (2020).

[6] Rajput, Nikhil Kumar, Bhavya Ahuja Grover, and Vipin Kumar Rathi. "Word frequency and sentiment analysis of twitter messages during coronavirus pandemic." arXiv preprint arXiv:2004.03925 (2020). 38

[7] Saranya, S., and G. Usha. "A Machine Learning-Based Technique with IntelligentWordNet Lemmatize for Twitter Sentiment Analysis." Intelligent Automation & Soft Computing 36.1 (2023).

[8] Whang, Dylan, and Soroush Vosoughi. "Dartmouth CS at WNUT-2020 task 2: Informative COVID-19 tweet classification Using BERT." arXiv preprint arXiv:2012.04539 (2020)

[9] Xue, Jia, Junxiang Chen, Ran Hu, Chen Chen, Chengda Zheng, Yue Su, and Tingshao Zhu. "Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach." Journal of medical Internet research 22, no. 11 (2020): e20550.

[10] Z. Hu, Z. Yang, Q. Li, A. Zhang, and Y. Huang, "Infodemiological study on COVID19 epidemic and COVID-19 infodemic", to be published(2020), doi: 10.21203/rs.3.rs-18591/v1.

[11] Zhou, Jianlong, et al. "Examination of community sentiment dynamics due to COVID-19 pandemic: a case study from a state in Australia." SN Computer Science 2 (2021): 1-11.