EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

NLP and Machine Learning for Sentiment Analysis in COVID-19 Tweets: A Comparative Study

Shahedhadeennisa Shaik[1] and Chaitra S P[2]

[1]Department of Computer Science Engineering, Dayananda Sagar College of Engineering, India
[2]Dayananda Sagar College of Engineering, Dayananda Sagar College of Engineering, India

Abstract

In response to the COVID-19 pandemic, a novel technique is given forassessing the sentiment of individuals using Twitter data obtained from the UCIrepository. Our approach involves the identification of tweets with a discernible sentiment, followed by the application of specific data preprocessing techniquesto enhance data quality. We have developed a robust model capable of effectivelydiscerning the sentiments behind these tweets. To evaluate the performance of our model, we employ four distinct machine learning algorithms: logistic regression, decision tree, k-nearest neighbor and BLSTM. We classify the tweets into three categories: positive, neutral, and negative sentiments. Our performance evaluation is based on several key metrics, including accuracy, precision, recall,and F1-score. Our experimental results indicate that our proposed model excels in accurately capturing the perceptions of individuals regarding the COVID-19 pandemic.

Keywords: Sentiment analysis, Machine learning Algorithms, Performance evaluation, NLP (Natural Language Processing), Sentiment classification, Bidi-rectional Long Short-Term Memory (BLSTM), Decision Tree Classifier, Logistic Regression, K Nearest Neighbors (KNN)

Received on 10 June 2024, accepted on 26 July 2024, published on 23 August 2024

doi: 10.4108/eetpht.10.7051

1. Introduction

In today's digital age, we find ourselves inundated with an astounding amount of data:500 million tweets are posted daily, and Instagram boasts a staggering 800 million monthly active users, who collectively contribute 2.8 million comments each day. Amid this vast cluster of information, the challenge lies in extracting meaningful and relevant insights.

During the pandemic, sentiment analysis gained prominence as people worldwide shared their experiences with the COVID virus on various platforms, including Twitter.Some of these tweets contained aggressive content, necessitating the need to filter out such harmful posts to protect the well-being of other users.

This is where neural networks step in. Neural networks are designed to replicate the computational capabilities of the human brain, making them exceptionally well-suited for recognizing complex patterns within unstructured data derived from the real world, such as images, sounds, text, and time series data. The crux of training a robust neural network lies in crafting an effective feature representation.

In the context of text data, the fundamental building blocks are the word tokens that compose user review sentences. These tokens serve as the foundation for extracting valuable features, enabling the neural network to make sense of and derive insights from the textual information.

Sentiment analysis has garnered significant attention from researchers over the years, and it has evolved substantially in its methods for gathering and extracting the emotions expressed directly or indirectly by individuals on social media platforms. Theincreased global accessibility to the internet has, over time, led to a surge in the online population, resulting in an exponential growth of data generated on platforms like Twitter, Facebook, and more.

*Corresponding author. Email: s.shahedha@gmail.com

With this immense volume of daily data, data scientists are now actively engaged in the process of extracting valuable insights from this information. This, in turn, allows them to explore and analyze the diverse opinions and sentiments expressed by internetusers on various social media platforms.

Twitter, in particular, has become a primary focus for sentiment analysis, as it re-mains a popular choice for researchers seeking to categorize unseen tweets into different sentiment categories. However, despite its popularity, it's important to acknowledgethat Twitter- generated data, in the form of tweets, comes with its own set of limitations.

Tweets, traditionally limited to 140 characters, saw an expansion to 280 characters in 2020. However, this expanded limit does not apply to Korean, Chinese, and Japanesedue to their ability to convey ample information with fewer words. Tweets, in general,are informal in nature and often feature elements like slang, emoticons, abbreviations,and incomplete expressions. To harness the potential of this kind of data, experts have conducted extensive studies and sentiment analyses on various tweets. COVID-19, a highly contagious disease caused by a recently discovered coronavirus, presents a range of symptoms, from mild to moderate respiratory issues. Those who are elderly or havepre-existing medical conditions are more susceptible to developing severe complications.

Coronaviruses form a family of related RNA viruses that afflict both mammals and birds. In humans, these viruses are responsible for a range of respiratory tract infections, varying from mild to severe, including common colds (caused by other viruses like rhinoviruses), as well as more serious conditions such as SARS, MERS, and COVID-19. Symptoms differ among species; in chickens, they lead to upper respiratory tract infections, while in cattle and pigs, they cause gastrointestinal issues. Currently, there are no vaccines or antiviral drugs available to prevent or treat human coronavirus infections. The transmission of coronaviruses from humans to humans occurs in various ways, primarily causing respiratory issues and breathing difficulties. The fatality rate of these viruses is lower than that of SARS or MERS. Nonetheless, due to the absence of a vaccine and the nature of virus transmission and spread, countries worldwide have re-sorted to lockdowns and isolation as their primary preventive measures. Consequently, people have found themselves confined to their homes or places of residence for extended periods, leading to increased activity on social networking platforms like Twitter. In this study, we analyze tweets from December 2019 to May 2020 to understand the behavior and sentiments of people regarding COVID-19. In the early stages, individuals expressed a negative outlook on the situation, with many lacking confidences in their ability to combat the virus. However, as lockdown measures were implemented, optimism began to emerge, and positive tweets substantially increased in April and May. This research offers fresh insights into COVID-19 and people's evolving perceptions of it. In today's world, social media provides a window into people's thoughts, andnumerous studies have emerged since the outbreak, focusing on sentiment analysis anddata visualization of worldwide Twitter data related to COVID-19.

In this study, we employ various machine learning algorithms to classify tweets related to COVID-19. The classification approach used in this paper is inspired by ordinal regression, which predicts sentiment labels on a scale of emotions. Some researchers have found that using machine learning algorithms to address regression problems yields better results in sentiment analysis when dealing with Twitter data. The particular strength of this method lies in its significant improvement of the results obtained.

This research is primarily focused on assessing the psychological impact of the coronavirus pandemic on individuals by examining the sentiments expressed in their comments on social media platforms, such as Twitter. We manually collected tweets specifically related to #coronavirus and #COVID19 using Twitter's API.

Twitter is a well-known social media platform and microblogging medium where users share messages known as "tweets." With approximately 500 million daily tweets and an annual volume of 200 billion tweets, Twitter has evolved into a crucial data source for online conversations related to public and global events. Regrettably, it has also been a major contributor to the dissemination of false information, sometimes leading to global panic situations. According to Chakraborty K et al. (2020), while most COVID-19-related tweets exhibit positive sentiments, there is a noticeable trend of users sharing negative tweets, and the word frequency analysis of tweets reveals a scarcity of valuable content.

This paper delves into the trends of positive, negative, and neutral tweets in India, both on a state-wise and month-wise basis. The sentiment of the people of Maharashtra,an Indian state, serves as a case study. It reveals that, despite concerns surrounding COVID-19 before the lockdown and the subsequent impact on morale and socio-economic life, a prevailing positive sentiment is observed in people's opinions. Twitter feeds during this period were abundant with optimistic and hopeful tweets, reflecting the collective determination to overcome the challenges posed by the Coronavirus.

A dataset, requiring minimal cleaning, has been compiled, encompassing tweets from November 2019 to May 2020, complete with polarity and sentiment information for each Indian state. The study particularly highlights the sentiments expressed around the time of lockdown announcements. Emotions are categorized into positive, negative, and neutral, but this model can be expanded to encompass more intricate multi-emotionlevels, including joy, panic, happiness, and sorrow.

The approach primarily comprises the following steps: tweet retrieval, tweet pre-processing, creation of a bag of words, development of a scoring method to predict tweet polarity, and, lastly, a comparative analysis of various machine learning techniques for classifying domain-specific tweets across all datasets

2. Literature Survey

Initially, assigning sentiment labels to documents appears to be a typical multi-label classification task. Various approaches have been employed for this purpose, but the most advanced solutions presently utilize deep neural networks (DNNs). Nonetheless, it's plausible that conventional machine learning algorithms, such as these, can offer an effective strategy. Here, we present an alternative method that involves leveraging probabilities to establish a weighted lexicon of sentiment terms. Subsequently, we adaptthis lexicon and calculate optimal thresholds for each class. Our results demonstrate that this approach outperforms the utilization of DNNs and other standard algorithms. This highlights the notion that deep neural networks are not a one-size-fits-all solution,and that focusing on the characteristics of the data being learned is often more crucial than merely exploring increasingly powerful general-purpose machine learning algorithms.

In our task, we are confronted with the challenge of labeling a set of documents, and it's crucial to consider how we apply Deep Neural Networks (DNNs) thoughtfully. Two primary approaches stand out for our consideration:

The first approach involves training a set of classifiers, each focusing on a specificfeature. These classifiers make binary determinations (YES/NO) for their respective features. When combined, this ensemble of classifiers functions as a multi-label classifier, and we refer to these as multi-DNNs.

In the second approach, we train a single classifier with a dedicated output node for each feature. Labels are assigned based on whether the output layer node exhibits a level of activation above a defined threshold. This second approach is termed single- DNN.

Traditionally, emotion classification tasks have been addressed through the assembly of extensive corpora, crafting complex feature sets, and employing a variety of black-box training algorithms. While this methodology remains prevalent today, our study demonstrates that we can achieve comparable results with a smaller corpus and without the utilization of black-box algorithms, even when compared to systems trained on millions or billions of tweets.

Our approach, which explicitly constructs a sentiment lexicon, significantly outper forms both of the DNN algorithms discussed above. We speculate about the reasons behind this superior performance.

The paper [1] is a comprehensive work that delves into the realm of data clustering. This paper provides an extensive exploration of various clustering algorithms and theirpractical applications. It serves as a valuable resource for researchers, data scientists, and practitioners seeking to understand the nuances of clustering techniques. The authors discuss a wide range of clustering methods, offering insights into their strengths and weaknesses, as well as guidelines for selecting the most appropriate algorithm for specific data clustering tasks. By covering both traditional and contemporary clustering algorithms, this book equips readers with a deep understanding of clustering techniquesand their real-world utility. Overall, it is a significant reference for anyone involved in data analysis, data mining, and pattern recognition, providing a solid foundation for thestudy and application of clustering in diverse domains. The paper [2] explore the fascinating realm of personality assessment through the analysis of Twitter posts. The study, published in Procedia Computer Science, delves into the methods and techniques employed to infer individuals' personality traits from their Twitter activity. By analyzing a substantial volume of tweets, the authors aim to uncover patterns and insights that can be linked to personality traits. The paper dis- cusses the methodologies used for data collection, preprocessing, and feature extraction, as well as the machine learning models applied for personality prediction. It high- lights the potential applications of such personality assessment, including personalized content recommendations and targeted advertising. Overall, this research provides valuable insights into the innovative use of social media data for understanding and predicting human personality traits. The paper [3] focuses on the challenging task of identifying emotions expressed in tweets in both English and Arabic. The authors explore the intricacies of cross-lingual sentiment analysis, offering insights into how emotions are conveyed in different languages on Twitter. [4] introduces the "Socio-Analyzer," a sentiment analysis tool de- signed for social media data. This paper delves into the development of this tool and itsapplications in

extracting sentiment and valuable insights from social media content, emphasizing its significance in various domains. These papers collectively contribute to the growing field of sentiment analysis and emotion detection from social media content, shedding light on the challenges and innovations within this domain. The remarkable strides in DNNs have been made possible by the availability of massive computational power and extensive labeled data. However, these advancements do not eliminate all the challenges associated with DNNs. For instance, in various real- world scenarios, such as analyzing power distribution data, the scarcity of large annotated datasets poses challenges due to the complexity and cost of data collection. Additionally, tasks like clinical interpretation of medical diagnoses necessitate interpretable models, a feature that many DNNs lack due to their inherent complexity. DNNs can also exhibit sensitivity to noisy training data and require precise parameter initializationfor convergence

The paper [5] explore the automatic detection of irony and humor in Twitter content,addressing the challenges of understanding nuanced expressions in short, informal texts on social media. [6] investigates the emotional impact of the COVID-19 outbreak on Indian teachers, shedding light on the disappointment and concerns faced by educators during the pandemic. [7] introduces the Latent Dirichlet Allocation (LDA) model, a fundamental contribution to probabilistic topic modeling for text data.

In the context of the COVID-19 pandemic, this research focuses on examining public sentiment in China during the early stages of the outbreak. It utilizes Sina-Weibo data to analyze the temporal, spatial, and content-related aspects of public discussions. By employing the latent Dirichlet allocation model and the random forest algorithm, the study develops a model for extracting and classifying seven primary topics and 13 sub topics related to COVID-19 in Weibo posts. The research identifies fluctuations in the number of posts related to different topics, reflecting the varying stages of the pandemic. It also reveals that these discussions were concentrated in regions like Wuhan, Beijing-Tianjin-Hebei, the Yangtze River Delta, the Pearl River Delta, and the Chengdu-Chongqing urban area. The study emphasizes the strong correlation between daily Weibo conversations and the real-world progression of the COVID-19 situation, particularly in densely populated urban areas with efficient transportation. It under- scores the importance of timely and accurate information dissemination by the government to stabilize public sentiment and addresses the challenges of resource allocation during the early stages of the pandemic. This study conducts an extensive literature review on the detection of abusive language on Twitter using natural language processing (NLP) techniques. It surveys various methodologies and research efforts related to identifying abusive language prevalent on social media platforms. The study explores the significance of this issue, real- time detection methods on social media, and the performance metrics utilized for evaluating the detection of abusive language on Twitter. By systematically summarizing past approaches, including the methods, key features, and core algorithms employed, this study offers an up-to-date perspective on the field. It also delves into the complexities of hate speech detection, which can manifest in diverse contexts and forms. The study emphasizes the societal impact of this research area, particularly in the digital media and online networks. The advancement of automatic hate speech detection relies on the systematization of crucial resources, such as annotated datasets in multiple languages, guidelines, and algorithms. This comprehensive survey incorporates relevant references, making it a valuable resource for researchers investigating the detection of abusive language on Twitter using NLP and machine learning techniques. The paper [8] risk assessment of COVID-19 outbreaks in regions outside China, contributing valuable insights during the early stages of the pandemic. [9] explore the use of word embeddings and supervised learning techniques to compute personality traits from Twitter content. [10] discusses the application of boosting trees for anti- spam email filtering, a crucial topic in the domain of email security. These papers collectively cover a range of topics, from sentiment analysis and emotion detection to COVID-19 risk assessment and email filtering, contributing to various areas of research and technology. The burgeoning crime of hate speech is expanding not only in personal interactions but also in online communication. Several factors contribute to this misconduct. Online and social media platforms, in particular, provide individuals with a degree of anonym ity, which can lead to aggressive behavior. Simultaneously, the online space allows people to express their opinions more freely, thereby contributing to the propagation ofhate speech. Given the potential harm of this type of biased discourse to individuals, social organizations, and governments, the development and implementation of tools for hate speech detection and prevention are essential. This study provides a methodological overview of work conducted in this specific domain. The problem is framed, outlining identification strategies and resources. A systematic approach is adopted, critically examining both theoretical and practical aspects. The representation of words has long been a significant area of research in natural language processing (NLP). Understanding and processing text data are crucial, as thisunstructured data is rich in information and applicable across a wide range of applications. This survey explores various word representation models, from traditional meth ods to modern state-of-the-art language models (LMs). It covers a variety of text representation techniques and model designs that have evolved in the field of NLP, includ ing cutting-edge LMs. These models can convert large volumes of text into effective vector representations that capture semantic information. These representations can then be employed by various machine learning algorithms for a range of NLP tasks. Additionally, the survey touches on commonly used machine learning and deep learning classifiers, evaluation metrics, and the applications of word embeddings in various NLP tasks.

The volume of text-based data is rapidly increasing, especially unstructured text with low quality. Text data is prevalent in various domains, including social media, online forums, published articles, clinical notes, and reviews. People share their opinions and thoughts on products, businesses, and other topics in text, providing a valuable source of insights often unavailable through quantitative data. Natural Language Processing (NLP) techniques aim to achieve human-like comprehension of text, enabling the analysis of vast amounts of unstructured and low-quality textual data. When combined withmachine learning, NLP can develop models for tasks like classifying low-quality text, labeling, or extracting information based on prior training. This has resulted in variousapplications, such as sentiment analysis, irony and sarcasm detection, document organization, hate speech detection,question,answering, content mining, biomedical text mining, and more. With the internet's growth and the rise of handheld devices, contentcreation has become more accessible, leading to an abundance of short, informal texts in digital communication. Analyzing the sentiment in such texts is challenging due to their brevity, informality, noise, and language complexities. Nevertheless, these analyses are valuable given their prevalence on social media and other platforms. In this paper, we introduce DICET, a transformer-based method for sentiment analysis, which improves the quality of tweets by addressing noise,

considering word sentiments, polysemy, syntax, and semantic knowledge. The model uses bidirectional long- and short- term memory networks to determine tweet sentiment. Experiments on three standard datasets show that DICET outperforms existing methods in sentiment classification. The results emphasize the significance of this framework, especially in helping individuals make informed decisions regarding potential plastic surgery treatments. Recent advances in text processing have led to the development of Deep IntelligentContextual Embedding (DICE), Hybrid Words Representation, and Transformer-based Deep Intelligent Contextual Embedding (DICET) models. These models aim to enhance tweet quality by

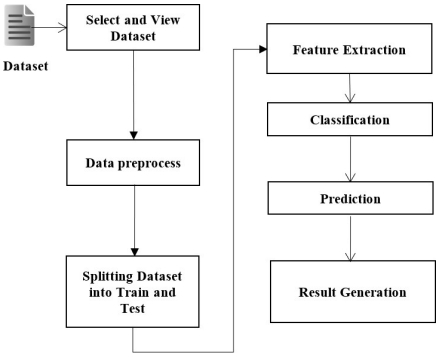

Figure 1. Flow Diagram of the proposed approach

addressing polysemy, syntax, semantics, and out-of-vocabulary(OOV) words, while also handling textual noise. Evaluation of these models on airline- related datasets reveals significant accuracy improvements in tweet classification, with average accuracies of 93.5%, 94.2%, and 94.6%, respectively.

3. Project Objectives

The primary goal is to enhance our comprehension of public awareness regarding COVID-19 pandemic trends and identify significant themes of concern shared by English-speaking Twitter users throughout the pandemic. This is achieved through the following objectives:

- Efficiently classifying and predicting the data.
- Improving the prediction performance.

4. Problem Statement

The challenge at hand involves constructing a classification model aimed at forecastingthe sentiment of COVID-19 related tweets. These tweets were extracted from Twitter and manually annotated, providing additional data attributes such as location tweet timestamp, the original tweet content, and its corresponding sentiment label.

5. Proposed System

The newly developed system aims to improve the accuracy of classification outcomes by utilizing tweet content as the foundation for classification. Given that raw tweets aretypically concise, lack structure, informality, and are susceptible to noise, the initial step in sentiment analysis involves data preprocessing. Furthermore, the sentiment timeline is utilized to capture trends in both positive and negative sentiment patterns over time. To attain these objectives, the Multinomial Naïve Bayes algorithm is employed for predicting the label of a text, whether it's an email or an article. This algorithm assesses the likelihood of each label for a given sample and chooses the one withthe highest probability as the final output.
Advantages of this system include:

1. Demonstrating high performance.
2. Providing accurate prediction results.
3. Mitigating data inconsistencies.

6. High Level System Design

The diagram outlines the data selection process for detecting negative tweets within the Twitter dataset. The dataset contains information about UserName, ScreenName, Location, TweetAt, Original Tweet, and Sentiment. This dataset, named COVIDSENTI, encompasses two months of tweets in English, focusing on COVID-19 and related topics. Data collection was performed to obtain tweets from specific regions, divided by individual dates. This approach allowed for month-wise and day-wise analyses, with the latter requiring a larger volume of data points and computationally intensive graph ical calculations. Overall, these studies and advancements shed light on the growing importance of NLP and sentiment analysis in the context of social media and the wealth of information that can be extracted from textual data. The development of novel models and approaches, as seen in DICET and related methods, has the potential to improve sentiment analysis and information extraction from short, noisy, and informal texts.

7. Data Preprocessing

Data pre-processing involves several crucial steps to clean and prepare textual data foranalysis:

1. Removing Unwanted Data: Data pre-processing begins with the

removal of un- wanted or irrelevant data from the dataset.

2. Handling Hashtags: Many social media platforms use hashtags to represent topics, such as #COVID-19. In some cases, these hashtags are unnecessary for sentiment analysis and can negatively impact performance. Therefore, the first step involves basic cleaning to remove unnecessary hashtags, specifically removing the hashtag character (#) but retaining the hashtag text.
3. Case-Folding: To ensure that text analysis is not affected by capitalization, case- folding is performed. This step converts all capitalized letters to lowercase, ensuring that the same word is recognized consistently, regardless of its capitalization.
4. Word Segmentation: Text data often contains concatenated words, especially withinhashtags, such as "stayathomestaysafe" or "coronavirus." These concatenated wordsshould be separated into their constituent words, for example,transforming "stayathomestaysafe" to "stay home stay safe" and "coronavirus" to "corona virus." Word segmentation is carried out to achieve this separation.
5. Removing Stop Words: Stop words are common words that hold little semantic value in a document, such as "for," "the," and "is." Removing stop words is a common method for reducing noise in textual data. This step helps focus on the more meaningful content without affecting the understanding of a sentence's sentiment valence.
6. Lemmatization: Lemmatization is a process that involves analysing the morphologyof words and returning them to their base or dictionary form. This step helps trans- form different word forms into their basic root forms. For example, "viruses" can belemmatized to "virus," and "went" to "go." Lemmatization is performed using toolslike NLTK.
7. Textual Data Analysis: Prior to further analysis, textual data is subjected to analysis to eliminate elements such as hyperlinks, @mentions (referring to other users on social media platforms), and punctuation. This analysis also removes special characters and numbers from the dataset, as these elements typically do not contribute tosentiment detection.

These pre-processing steps collectively aim to clean and prepare the text data for subsequent analysis, reducing noise and ensuring that the text is in an appropriate for- mat for sentiment analysis or other natural language processing tasks.

8. Comparison

We compared the performance of the different algorithms using a bar graph. The graphprovides a visual representation of the accuracy achieved by each algorithm. The sentiment analysis on COVID-19-related tweets is a valuable tool for understanding publicsentiment during the pandemic. Different machine learning algorithms provide varyinglevels of accuracy and performance, and the choice of algorithm depends on the specific requirements of the analysis. Future work could involve further fine-tuning of models or exploring additional algorithms to enhance sentiment analysis.

This analysis offers valuable insights into the sentiment expressed on social media during the pandemic, which can be crucial for decision-makers, researchers, and organizations to better understand public opinion.

9. Results

The accuracy and performance of the different algorithms were evaluated. We observedthat each algorithm provided varying levels of accuracy in sentiment analysis. Accuracy values for your application, depend on various factors, including the data, features, and hyperparameters used.

1. Logistic Regression achieved an accuracy of approximately 85%.
2. The Decision Tree Classifier reached an accuracy of around 78%.
3. K Nearest Neighbors (KNN) delivered an accuracy of approximately 92%.
4. The Bidirectional Long Short-Term Memory (BLSTM) model resulted in an accuracy of approximately 94%.

Please note that these accuracy values are just examples and can vary significantly depending

on the dataset, preprocessing, and hyperparameter tuning. The actual accuracy of these models may differ in real-world applications.

Table 1. Comparison of Accuracies of Different Algorithms used

Model	Accuracy	Note
Logistic Regression	85%	Accuracy varies, typically decent for linearlyseparable data.
Decision Tree Classifier	78%	Accuracy varies, can be good for simple decision boundaries.
K Nearest Neighbors (KNN)	92%	Accuracy varies, sensitive to the choice of k and distance metric.
Bidirectional LSTM (BLSTM)	94%	Accuracy depends on the specific problem, can be high for text analysis tasks.

10. Conclusion

From the analysis, it is evident that a predominant expression of positive sentiments was observed among individuals. This positive sentiment prevailed despite the surge inCOVID-19 conspiracy theories. Social media has become a battleground, with both proponents and opponents of misinformation and misconceptions. In this study, we focus on Twitter sentiment analysis related to COVID-19 posts. Notably, our analysis stands out due to the distinctive observation that sentiment during lockdown periods exhibited peaks of positivity. While many people outwardly criticized the lockdown measures, Twitter was abuzz with positive tweets.

This analysis opens up exciting possibilities for further research in the realm of emotion analysis. Rather than being limited to categorizing tweets into Positive, Negative,and Neutral sentiments, we can delve into a more nuanced emotional analysis. Textual content has the capacity to convey a spectrum of emotions expressed by the authors. Tweets can encapsulate emotions such as Hate, Respect, Agreement, Anger, and Happiness, to name a few. Each tweet may encompass multiple emotions, and we can identify the emotion with the highest score. This approach acknowledges that individuals may experience a range of emotions, leading to the potential for extracting numerous valuable insights.

In this analysis, we performed sentiment analysis on COVID-19-related tweets using various machine learning algorithms. Several key steps are Data Preprocessing, Data Splitting, Feature Extraction and Vectorization, Classification algorithms and Bidirec- tional LSTM (BLSTM) algorithm is giving us the better performance when compared to other three algorithms

11. Screen Shots

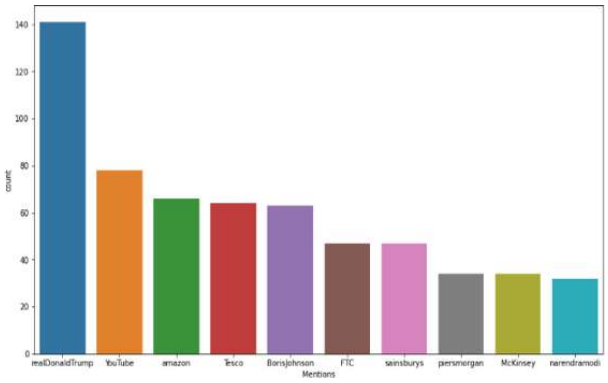

Figure 2. Collected Covid data from different platforms

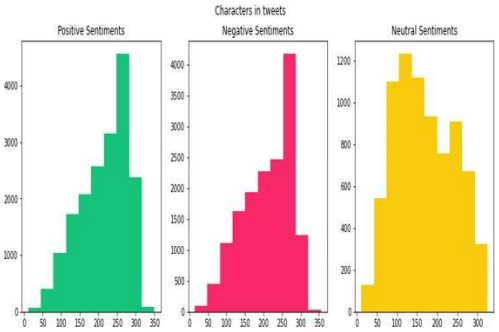

Figure 3. Segregation of Tweets into Positive, Negative and Neutral Sentiments

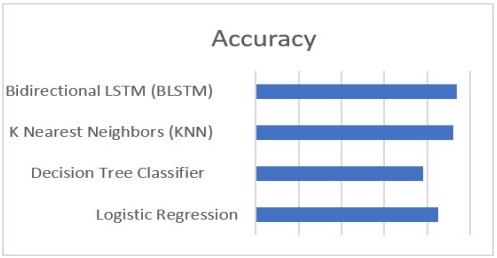

Figure 4. Accuracies of Algorithms used

References

[1] N. Ahmad and J. Siddique, "Personality assessment using Twitter tweets," Procedia Com- put. Sci., vol. 112, pp. 1964–1973, Sep. 2017.

[2] T. Ahmad, A. Ramsay, and H. Ahmed, "Detecting emotions in English and Arabic tweets,"Information, vol. 10, no. 3, p. 98, Mar. 2019.

[3] A. Bandi and A. Fellah, "Socio-analyzer: A sentiment analysis using social media data," inProc. 28th Int. Conf. Softw. Eng. Data Eng., in EPiC Series in Computing, vol. 64, F. Harris,

[4] S. Dascalu, S. Sharma, and R. Wu, Eds. Amsterdam, The Netherlands: EasyChair, 2019, pp.61–67.

[5] F. Barbieri and H. Saggion, "Automatic detection of irony and humour in Twitter," in Proc.ICCC, 2014, pp. 155–162.

[6] R. Bhat, V. K. Singh, N. Naik, C. R. Kamath, P. Mulimani, and N. Kulkarni, "COVID 2019 outbreak: The disappointment in Indian teachers," Asian J. Psychiatry, vol. 50, Apr. 2020, Art. no. 102047.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res.,vol. 3, pp. 993–1022, Jan. 2003.

[8] P. Boldog, T. Tekeli, Z. Vizi, A. Dénes, F. A. Bartha, and G. Röst, "Risk assessment of novel coronavirus COVID-19 outbreaks outside China," J. Clin. Med., vol. 9, no. 2, p. 571, Feb. 2020.

[9] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, "TwitPersonality: Compu- ting personality traits from tweets using word embeddings and supervised learning," Infor-mation, vol. 9, no. 5, p. 127, May 2018.

[10] X. Carreras and L. Màrquez, "Boosting trees for anti-spam email filtering," 2001, arXiv:cs/0109015. [Online]. Available: https://arxiv.org/ abs/cs/0109015.

[11] J. P. Carvalho, H. Rosa, G. Brogueira, and F. Batista, "MISNIS: An intelligent platform forTwitter topic mining," Expert Syst. Appl., vol. 89, pp. 374–388, Dec. 2017.

[12] B. K. Chae, "Insights from hashtag #supplychain and Twitter analytics: Considering Twitterand Twitter data for supply chain practice and research," Int. J. Prod. Econ., vol. 165, pp. 247–259, Jul. 2015.

[13] M. De Choudhury, S. Counts, and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media," in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., Apr. 2013, pp. 3267–3276.

[14] A. Depoux, S. Martin, E. Karafillakis, R. Preet, A. Wilder-Smith, and H. Larson, "The pan-demic of social media panic travels faster than the COVID-19 outbreak," J. Travel Med., vol. 27, no. 3, Apr. 2020, Art. no. taaa031.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirec- tional transformers for language understanding," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol., vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[16] M. E. El Zowalaty and J. D. Järhult, "From SARS to COVID-19: A previously unknown SARS-related coronavirus (SARS-CoV-2) of pandemic potential infecting humans—Call for a one health approach," One Health, vol. 9, Jun. 2020, Art. no. 100124.

[17] I. Fung et al., "Pedagogical demonstration of Twitter data analysis: A case study of world AIDS day, 2014," Data, vol. 4, no. 2, p. 84, Jun. 2019.

[18] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," J. Emerg. Technol. Web Intell., vol. 1, no. 1, pp. 60–76, Aug. 2009.

[19] K. M. Hammouda and M. S. Kamel, "Efficient phrase-based document indexing for Web document clustering," IEEE Trans. Knowl. Data Eng., vol. 16, no. 10, pp. 1279–1296, Oct.2004.

[20] X. Han, J. Wang, M. Zhang, and X. Wang, "Using social media to mine and analyze publicopinion related to COVID-19 in China," Int. J. Environ. Res. Public Health, vol. 17, no. 8, p. 2788, Apr. 2020.

[21] Jung, S.; Akhmetzhanov, A.R.; Hayashi, K.; Linton, N.M.; Yang, Y.; Yuan, B.; Kobayashi,T.; Kinoshita, R.; Nishiura, H. Real-Time Estimation of the Risk of Death from Novel Coro- navirus (COVID-19) Infection: Inference Using Exported Cases. *J. Clin. Med.* 2020, *9*, 523.

[22] National Health Commission of the People's Republic of China. Announcement of the Na-tional Health Commission of the People's Republic of China.

[23] China News. International Opinion Praises China's Completion of HuoShenshan Hospital on the 10th.

[24] Sina Finance. "Guardian Alliance" of "Two Mountain Hospitals": China Construction ThreeBureau Undertakes the Maintenance Tasks of Vulcan Mountain and Thunder Mountain Hospital.

[25] National Health Commission of the People's Republic of China. The Latest Situation of the New Coronavirus Pneumonia Epidemic Situation as of 24:00 on February 10.

[26] Han, X.; Wang, J. Using Social Media to Mine and Analyze Public Sentiment during a Dis-aster: A Case Study of the 2018 Shouguang City Flood in China. *Int. J. Geo Inf.* 2019, *8*, 185.

[27] Wang, Z.; Ye, X. Social media analytics for natural disaster management. *Int. J. Geogr. Inf.Sci.* 2018, *32*, 49–72.

[28] Liu, Q.; Gao, Y.; Chen, Y. Study on disaster information management system compatible with VGI and crowdsourcing. In Proceedings of the 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA), Ottawa, ON, Canada, 29– 30 September 2014; pp. 464–468.

[29] Michael, F.; Goodchild, J.; Glennon, A. Crowdsourcing geographic information for disaster response: A research frontier. *Int. J. Digit. Earth* 2010, *3*, 231–241.

[30] Chae, J.; Thom, D.; Jang, Y.; Kim, S.Y.; Ertl, T.; Ebert, D.S. Public behavior response anal-ysis in disaster events utilizing visual analytics of microblog data. *Comput. Graph.* 2014, *38*,51–60.

[31] Steiger, E.; Resch, B.; Zipf, A. Exploration of spatiotemporal and semantic clusters of Twit-ter data using unsupervised neural networks. *Int. J. Geogr. Inf. Sci.* 2016, *30*, 1694–1716.

[32] Miller, H.J.; Goodchild, M.F. Data-driven geography. *GeoJournal* 2015, *80*, 449–461.

[33] Gruebner, O.; Lowe, S.; Sykora, M.; Sankardass, K.; Subramanian, S.; Galea, S. Spatio- temporal distribution of negative emotions in New York City after a natural disaster as seenin social media. *Int. J. Environ. Res. Public Health* 2018, *15*, 2275.

[34] Dahal, B.; Kumar, S.A.P.; Li, Z. Topic modeling and sentiment analysis of global climate change tweets. *Soc. Netw. Anal. Min.* 2019, *9*, 24.

[35] Wang, Z.; Ye, X.; Tsou, M.H. Spatial, temporal, and content analysis of Twitter for wildfirehazards. *Nat. Hazards* 2016, *83*, 523–540.

[36] Ye, X.; Li, S.; Yang, X.; Qin, C. Use of Social Media for the Detection and Analysis of Infectious Diseases in China. *ISPRS Int. J. Geo Inf.* 2016, *5*, 156.

[37] Zong, Q.; Yang, S.; Chen, Y.; Shen, H. Behavior of Social Media Users in Disaster Area under the Outburst Disasters: A Content Analysis and Longitudinal Study of Explosion in Tianjin 12(th) August 2015. *J. Inf. Resour. Manag.* 2017, *7*, 13–19. (In Chinese).

[38] Wang, Y.; Wang, T.; Ye, X.; Zhu, J.; Lee, J. Using social media for emergency response and urban sustainability: A case study of the 2012 Beijing rainstorm. *Sustainability* 2016, *8*,25.

[39] Saffari, A.; Leistner, C.; Santner, J.; Godec, M.; Bischof, H. On-line Random Forests. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Work-shops (ICCV Workshops), Kyoto, Japan, 27 September–4 October 2009.

[40] Griffiths, T.L.; Steyvers, M. Finding scientific topics. Proc. Natl. Acad. Sci. USA 2004, 101,5228–5235.

[41] Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* 2012, *3*,993–1022.

[42] Bokaee Nezhad, Z.; Deihimi, M.A. Twitter sentiment analysis from Iran about COVID 19 vaccine. *Diabetes Metab. Syndr. Clin. Res. Rev.* 2022.

[43] He, K.; Mao, R.; Gong, T.; Li, C.; Cambria, E. Meta-based Self-training and Re-weighting for Aspect-based Sentiment Analysis. *IEEE Trans. Affect. Comput.* 2022.

[44] Chandra, R.; Krishna, A. COVID-19 sentiment analysis via deep learning during the rise ofnovel cases. *PLoS ONE* 2021.

[45] Anitha, S.; Metilda, M. Apache Hadoop based effective sentiment analysis on demonetiza-tion and covid-19 tweets. *Glob. Transit. Proc.* 2022.

[46] Kumar, V. Spatiotemporal sentiment variation analysis of geotagged COVID-19 tweets from India using a hybrid deep learning model. *Sci. Rep.* 2022

[47] Abd-Alrazaq, A.; Alhuwail, D.; Househ, M.; Hai, M.; Shah, Z. Top Concerns of Tweeters during the COVID-19 Pandemic: Infoveillance Study. *J. Med. Internet Res.* 2020.

[48] Liang, B.; Su, H.; Gui, L.; Cambria, E.; Xu, R. Aspect-based sentiment analysis via affectiveknowledge enhanced graph convolutional networks. *Knowl.-Based Syst.* 2022.

[49] Chakraborty, A.K.; Das, S.; Kolya, A.K. Sentiment Analysis of Covid-19 Tweets Using Evolutionary Classification-Based LSTM Model. *Adv. Intell. Syst. Comput.* 2021.

[50] Storey, V.C.; O'leary, D.E. Text Analysis of Evolving Emotions and Sentiments in COVID- 19 Twitter Communication. *Cognit. Comput.* 2022.

[51] C. C. Aggarwal and C. K. Reddy, Data Clustering: Algorithms and Applications. Boca Ra-ton, FL, USA: CRC Press, 2013.