Exploring COVID-19 vaccine sentiment: a Twitter-based analysis of text processing and machine learning approaches

Ban Safir Khalaf, Hazlina Hamdan, Noridayu Manshor
Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Malaysia

Article Info

Article history:

Received Nov 10, 2023
Revised Apr 26, 2024
Accepted Aug 7, 2024

Keywords:

COVID-19
Feature extraction
Long short-term memory
Sentiment analysis
Text frequency-inverse document frequency

ABSTRACT

In the wake of the 2020 coronavirus disease (COVID-19) pandemic, the swift development and deployment of vaccines marked a critical juncture, necessitating an understanding of public sentiments for effective health communication and policymaking. Social media platforms, especially Twitter, have emerged as rich sources for gauging public opinion. This study harnesses the power of natural language processing (NLP) and machine learning (ML) to delve into the sentiments and trends surrounding COVID-19 vaccination, utilizing a comprehensive Twitter dataset. Traditional research primarily focuses on ML algorithms, but this study brings to the forefront the underutilized potential of NLP in data preprocessing. By employing text frequency-inverse document frequency (TF-IDF) for text processing and long short-term memory (LSTM) for classification, the research evaluates six ML techniques K-nearest neighbors (KNN), decision trees (DT), random forest (RF), artificial neural networks (ANN), support vector machines (SVM), and LSTM. Our findings reveal that LSTM, particularly when combined with tweet text tokenization, stands out as the most effective approach. Furthermore, the study highlights the pivotal role of feature selection, showcasing how TF-IDF features significantly bolster the performance of SVM and LSTM, achieving an impressive accuracy exceeding 98%. These results underscore the potential of advanced NLP applications in real-world settings, paving the way for nuanced and effective analysis of public health discourse on social media.

Corresponding Author:

Hazlina Hamdan
Department of Computer Science, Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
43400 UPM Serdang, Selangor Darul Ehsan, Malaysia
Email: hazlina@upm.edu.my

1. INTRODUCTION

The coronavirus, known as COVID-19, was declared a global pandemic in 2020, marking a significant turning point in modern public health history. The widespread impact of the virus, leading to over 5 million fatalities and more than 250 million cases globally [1], has compelled societies worldwide to adapt to new norms such as quarantine, social distancing, mask-wearing, and increased hygiene practices. This rapid emergence and ensuing global crisis have not only placed unprecedented pressure on healthcare systems and frontline workers but have also posed multifaceted challenges for scientists and policymakers [2].

The development and distribution of COVID-19 vaccines have been pivotal in the fight against the pandemic. As various vaccines were rapidly developed and deployed, public response to these vaccines became a crucial aspect of understanding and managing the pandemic effectively [3]. However, vaccine rollout phases faced significant challenges due to varying degrees of public acceptance, hesitancy, and resistance, influenced by a complex interplay of misinformation, societal beliefs, and individual health concerns [3]-[7]. This situation underscores the necessity of accurately gauging and understanding public sentiment towards vaccination efforts, which is essential for devising effective public health strategies and communication [4].

In this context, social media platforms, particularly Twitter, have emerged as vital arenas for public discourse, opinion formation, and information dissemination about COVID-19 vaccines [5]. The extensive use of these platforms for sharing experiences, opinions, and information about the pandemic has provided an unprecedented opportunity to analyze public sentiment in real-time [6], [7]. However, the sheer volume and complexity of this unstructured data make manual analysis impractical, thereby highlighting the need for automated techniques that can effectively process and analyze such data.

Existing research has demonstrated the efficacy of machine learning (ML) algorithms in analyzing social media data for various applications, including public health [7]-[11]. However, there remains a notable gap in fully harnessing the capabilities of natural language processing (NLP) techniques, particularly in the preprocessing stage, to enhance the accuracy and depth of sentiment analysis in the context of health communication [8]-[10]. This study aims to address this gap by exploring the use of text frequency-inverse document frequency (TF-IDF) and long short-term memory (LSTM) models to process and classify sentiments related to COVID-19 vaccines on Twitter [9]. The primary objectives of this research are twofold: firstly, to develop a robust predictive model utilizing advanced NLP techniques for identifying public sentiments and trends regarding COVID-19 vaccines on Twitter; and secondly, to evaluate the effectiveness of various ML algorithms, including K-nearest neighbors (KNN), decision trees (DT), random forest (RF), artificial neural networks (ANN), support vector machines (SVM), and LSTM, in sentiment analysis with a focus on LSTM's capability in handling time-series data.

This study contributes to the existing body of knowledge by providing insights into public opinion dynamics during a global health crisis and introducing a novel approach in sentiment analysis that combines TF-IDF preprocessing with LSTM classification. The methodology and results presented in this paper offer significant implications for public health communication, policy formulation, and crisis management, especially in leveraging social media data for informed decision-making.

The remainder of this paper is organized as follows: section 2 provides an overview of ML techniques and a review of the relevant literature focusing on the categorization of COVID-19 tweets. Section 3 details the methodologies employed in this study, including data collection, preprocessing, and analytical techniques. The outcomes are discussed in section 4, juxtaposing our findings with existing literature, and concluding with a summary of the study's contributions and potential directions for future research in section 5.

2. RELATED WORKS

In this section, various ML approaches applied to study COVID-19 vaccine-related issues are discussed [3], [10]. Some studies focused on predicting acceptance or hesitancy of individuals towards the vaccine [4]-[9], while others addressed misinformation on the vaccine [11] and side effects prediction [12]-[14]. Some studies also investigated mRNA degradation [15]-[17] and conducted cross-country comparisons [18]-[20]. Most of the datasets used in the relevant works were self-collected from social media websites, with Twitter being the primary source [21]-[24]. While a few studies collected data through surveys, the collected dataset size is generally smaller. Unfortunately, most studies did not make their collected data available to the public, which limits the ability to conduct further studies on the same dataset.

Researchers have conducted extensive experimentation with various ML models, revealing that there is no universally superior model. Performance is highly contingent on the dataset utilized, as well as the methodologies implemented for feature extraction and selection. A notable investigation into the impact of feature set complexity was conducted by Miikkulainen *et al.* [25], which found that the accuracy of the classification model diminished when more than nine feature permutations were employed in their specific dataset. It's important to recognize that each study has its unique emphasis and employs distinct ML algorithms accordingly. For instance, [9], [10] applied algorithms such as KNN, DT, RF, and ANN. Furthermore, Zaidi *et al.* [3] employed SVM, which emerged as the most efficacious algorithm in their analysis.

The collected datasets being self-collected and unique to each study limit the reproducibility of results over another study. Future research on this topic should utilize previously collected data and focus on improving ML and NLP models. Table 1 provides a summary of the studies covered and highlights the different ML algorithms used in each paper.

Table 1. Summary of the related work

Authors, year	Focus	Best algorithm	Accuracy (%)	Advantages	Limitation
[3], 2021	Prediction of trends of vaccine using RF, DT, RNN, SVM, and KNN	SVM	89	– Provided a future predictive study, including a voting classifier.	– Low performance with the date-wise dataset.
[4], 2022	COVID-19 vaccine hesitancy among American adults	Gradient boosting (GB)	91 overall, 97	– High accuracy in predicting vaccine hesitancy; incorporation of a wide range of variables including trust and knowledge about vaccines.	– Focused on a specific population (U.S. adults); online survey data may not be representative of the general population.
[5], 2022	Nepali COVID-19-related tweets classification	SVM+RBF	72.1	– Higher accuracy with hybrid features (syntactic and semantic information). Lower feature size (300-D).	– Limited to one kind of contextual information (semantic) from FastText. Traditional ML methods are used, not end-to-end.
[6], 2023	Arabic Tweets-based sentiment analysis during COVID-19 in KSA	CNN	92.80	– High accuracy in deep learning sentiment classification. Comprehensive dataset including before and during the pandemic.	– Study focused only on Saudi dialect and MSA, possibly limiting broader applicability. Comparison with other deep learning techniques, but not with other ML techniques.
[7], 2022	Prediction on vaccine acceptance, hesitation, and rejection using XLNet	XLNet	63	– A large dataset is used (over six million instances).	– The study considered one ML model only.
[8], 2021	mRNA degradation in the COVID-19 vaccine using G-CNN G-GRU	G-GRU	76	– The use of a hybrid DL model. – Representing mRNA in the graph for CNN and GRU.	– The hybrid DL model is computationally expensive.
[9], 2022	Side effects of COVID-19 vaccine using LDA	LDA	78	– Relatively good accuracy.	– The self-collected dataset is not available online.
[10], 2023	COVID-19 vaccine hesitancy sentiment analysis	TextBlob+TF-IDF+LinearSVC	96.75	– High accuracy with deep learning models. Detailed review of sentiment analysis techniques.	– Specific focus on Twitter data limits the generalizability to other platforms. Requires large datasets for optimal performance.
[11], 2021	COVID-19 vaccine sentiment analysis using NB DT	DT	96	– The study experimented with two feature extraction techniques. – Country-based datasets were used and compared.	– Only NB and DT were considered.
[12], 2021	COVID-19 vaccine acceptance and progress using DT KNN RF NB	DT	96	– The study uses and compares several datasets. – The study achieves high accuracy on the used dataset.	– Some datasets have lower accuracy.
[13], 2022	COVID-19 vaccine acceptance and hesitancy using ANN	ANN	82	– The study experiments with different sets of features and their impact on the classification process.	– The study considers the traditional ANN model only.
[14], 2024	Sentiment analysis on Omicron Tweets	BERT and RoBERTa	93.39 and 93.47	– Hybrid classifiers with multiple feature extraction techniques. Transformer-based models showing high accuracy.	– Focus on a specific variant of COVID-19 (Omicron), which may limit the application of the study to other contexts.
[15], 2024	Sentiment analysis with improved ASO and ReLU-GRU	ReLU-GRU	97.87	– High accuracy and effectiveness in feature selection with ASO and SA.	– Focused on Twitter data, limiting applicability to other social media.

3. METHOD

This section delineates the methodology and conceptual approach of our study, which is structured as both a re-implementation and an augmentation of the research conducted by [3], [18]. Initially, our approach involves utilizing the same dataset as the previous studies, aiming to replicate their methodology accurately. Subsequently, we introduce an advanced ML framework that integrates TF-IDF for text processing and LSTM for classification. This integration is designed to address and improve upon the limitations identified in the earlier studies.

The foundational research [3], which serves as the bedrock of our project, investigates the public's sentiment toward the COVID-19 vaccine through an analysis of Twitter data. The original study categorizes tweets into positive, negative, or neutral sentiments concerning the vaccine. Our research extends this work by scrutinizing a variety of algorithms, namely KNN, SVM, ANN, DT, and RF, as reported in [3]. The earlier findings highlighted ANN as the most efficacious algorithm for datasets organized by date. However, the overall accuracy achieved in date-based datasets was found to be suboptimal, likely due to the constraints of the algorithms in processing time-sequential data. Conversely, SVM emerged as the most potent model for analyzing randomly ordered datasets. Recognizing these insights, our study proposes the adoption of more sophisticated ML models, hypothesizing that alternative approaches, particularly those harnessing the capabilities of LSTM, could significantly enhance the accuracy and reliability of sentiment analysis in this context.

Through this methodology, we aim not only to replicate but also to build upon and refine the findings of previous research. By incorporating advanced NLP techniques and leveraging LSTM's proficiency in handling sequential data, our study seeks to offer a more nuanced and accurate interpretation of public sentiment toward COVID-19 vaccination as expressed on social media platforms.

3.1. Dataset description

The dataset underpinning this study is comprised of comprehensive comma-separated values (CSV) files harvested from Twitter, specifically curated to analyze and predict public sentiment towards the COVID-19 vaccine [3]. This sentiment is categorized into three primary opinions: positive, negative, or neutral. Our analysis bifurcates this dataset into two distinct categories for a more nuanced examination.

3.1.1. Random dataset

The first category, referred to as the 'random dataset', encompasses a diverse collection of tweet texts from users. This dataset is crucial for understanding the raw, unfiltered public sentiment as it spontaneously appears on the social media platform. It provides a broad view of public opinion, unanchored to specific temporal events, thereby offering a general sentiment landscape regarding the COVID-19 vaccine.

3.1.2. Date-wise dataset

The second category, known as the 'date-wise dataset', is methodically organized based on the timestamps of the tweets. This dataset is instrumental in tracing the evolution of public sentiment over time. By aligning the tweets with their respective dates of posting, we can observe and analyze shifts in public opinion in response to real-world events, health updates, and policy changes related to the COVID-19 vaccine.

Tables 2 and 3 in the following sections provide detailed descriptions of these datasets. Table 2 elucidates the random dataset, illustrating its composition and diversity, while Table 3 focuses on the date-wise dataset, shedding light on the chronological aspect of the data collected. This dual-dataset approach enables a comprehensive analysis of Twitter-based public sentiment, offering insights both in terms of overarching trends and time-specific reactions.

Table 2. Random dataset description

Dataset type	Dataset samples	Collection period	Attributes
.csv	228208	Dec 2020-April 2021	Tweet text

Table 3. Date-wise dataset description

Dataset type	Dataset samples	Collection period	Attributes
.csv	228208	Dec 2020-April 2021	Date, Tweet text

Building upon the framework set forth by Zaidi *et al.* [3], this study initiates its analytical journey with the raw, unlabelled dataset derived from Twitter. Recognizing the potential of noise and irrelevant data in these raw tweets, the first critical step involves a meticulous preprocessing phase. Here, extraneous symbols and superfluous elements within the text data are filtered out, setting a clean slate for ML processes. This preprocessing employs two principal techniques: tokenization, which is in line with the foundational methodologies suggested in [3], and the TF-IDF algorithm, an advanced approach to refine the textual data for better ML applicability.

Following this, the study progresses into the classification phase, leveraging various ML algorithms. Replicating the methods from the foundational literature, five traditional ML algorithms are utilized: KNN, SVM, DT, RF, and ANN. In addition to these established techniques, this study introduces a sixth, more advanced algorithm: LSTM. The LSTM algorithm is executed in two distinct stages. Initially, it processes the tokenized data, and subsequently, it is applied in conjunction with the TF-IDF method. This dual-phase implementation is strategically designed to explore and validate whether these methods, individually or in tandem, can enhance the accuracy and efficiency of sentiment classification.

The experimental design is bifurcated into two segments. The initial segment assesses the efficacy of the conventional ML algorithms in conjunction with the tokenization method. The latter segment, representing a proposed enhancement, applies the LSTM algorithm, integrating TF-IDF features to assess their impact on classification performance. Additionally, the study conducts a comparative analysis where the SVM algorithm is tested using the TF-IDF features, juxtaposed against the results yielded by the LSTM model. This comparative assessment aims to elucidate the degree to which the integration of LSTM and TF-IDF contributes to improved data extraction and heightened classification accuracy.

Figure 1 illustrates the comprehensive framework employed for classifying opinions on the COVID-19 vaccine, encompassing the journey from raw data acquisition to the nuanced process of sentiment analysis using these advanced ML techniques. This visual representation underscores the methodological rigor and the multi-layered approach of this study, aiming to yield insights with heightened precision and relevance in understanding public sentiment towards COVID-19 vaccination.

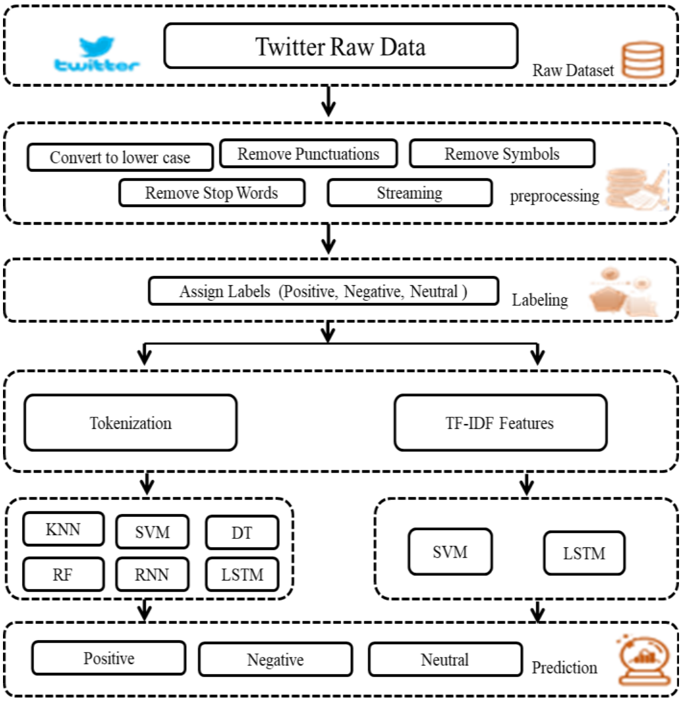

Figure 1. The framework of the COVID-19 vaccine opinion classification

3.2. Preprocessing

In the crucial preprocessing stage of our study, we meticulously refine the raw Twitter data to ensure the highest quality input for our ML algorithms. This process involves five key procedures, each designed to eliminate irrelevant or redundant elements from the text data. Figure 2 provides a visual representation, showcasing five examples of the input data (original tweet text) alongside the output data (processed text) following these preprocessing steps.

– Convert to lowercase: convert all the letters of the text into lowercase, we use the lowercase function from Python's string library (str).
– Remove punctuation: utilizing the regular expression library in Python, we strip away punctuation characters such as question marks (?), periods (.), and exclamation marks (!), among others. This step helps in simplifying the text for more accurate processing.
– Remove symbols: various symbols in tweets, including hashtags (#), asterisks (*), and symbols (@), are also removed.
– Remove stop words: our preprocessing also involves the removal of stop words-commonly used words in a language (like 'at', 'of', 'is') that offer little value in understanding sentiment. These words are filtered out to focus on more meaningful words that contribute significantly to the sentiment expressed in the text.
– Stemming: this process involves reducing words to their base or root form. For example, the plural word 'sides' is stemmed to its singular form 'side'.

Together, these preprocessing steps form the foundation of our data preparation, streamlining the text and paving the way for more accurate and effective sentiment analysis. The cleaned and processed data, as depicted in Figure 2, then serves as the input for the subsequent ML classification stages.

text	text_clean
Same folks said daikon paste could treat a cyt...	folk said daikon paste could treat cytokine st...
While the world has been on the wrong side of ...	world wrong side history year hopefully bigges...
#coronavirus #SputnikV #AstraZeneca #PfizerBio...	coronavirus sputnikv astrazeneca pfizerbiontec...
Facts are immutable, Senator, even when you're...	fact immutable senator even youre ethically st...
Explain to me again why we need a vaccine @Bor...	explain need vaccine borisjohnson matthancock ...

Figure 2. A snapshot of the input and output of the preprocessing

3.3. Labelling

The labeling is done using the TexBlob tool in Python. TextBlob offers a pivotal feature called 'polarity', which is instrumental in determining the sentiment orientation of a given text. This function evaluates the sentiment of text and assigns it a polarity score, categorizing it as positive, negative, or neutral.

It is essential to note that the polarity function typically yields floating-point numbers rather than integers. For instance, a score of 0.8 suggests that the text is likely to convey a positive sentiment. These float values provide a nuanced understanding of the sentiment's strength. However, for clarity and consistency in our analysis, we categorize these float values into their respective integer categories. Scores closer to 1 are categorized as positive, scores near -1 as negative, and scores around 0 as neutral. This approach ensures a standardized interpretation of the sentiment scores, facilitating a more straightforward and effective analysis. By utilizing the polarity scores provided by TextBlob, we can accurately label each tweet in our dataset with its corresponding sentiment, forming a foundational dataset for the subsequent ML classification process.

3.4. Feature extraction

Feature extraction stands as a critical process in ML and computer vision, pivotal in distilling meaningful features or patterns from a raw dataset. Its primary objective is to convert data into a more condensed, easily analyzable, and interpretable form. In the realm of image processing, this involves the discernment of distinctive patterns—such as edges, corners, shapes, textures, and colors—vital for object classification and recognition. In contrast, within NLP, feature extraction transforms textual data into a numerical format conducive to analytical computation. This encompasses a range of techniques, including but not limited to bag-of-words, word embeddings, and topic modeling, as noted by Zaidi *et al.* [3]. The significance of feature extraction in ML and data analysis is profound; it not only simplifies data complexity but also accentuates critical information, thereby potentially elevating the performance of analytical models.

In our text processing procedure, we explore two distinct methods. The first is tokenization, a foundational method established in preceding works, while the second is the TF-IDF algorithm, a more sophisticated technique proposed in this study. These methods are critically examined to assess their efficacy in feature extraction. Subsequently, we compare the classification results yielded by each algorithm to ascertain their relative effectiveness on our research datasets. This comparative analysis is aimed at determining the optimal strategy for feature extraction in sentiment analysis, particularly focusing on how these approaches can contribute to a more nuanced and accurate interpretation of public sentiment towards COVID-19 vaccines on Twitter.

4. RESULTS AND DISCUSSION

The experimental results are represented in the form of accuracy, precision, recall, and F-measure for both the random dataset and the date-wise dataset on each trained model can be seen in Table 4. A notable observation from the results is the superior performance of the LSTM model enhanced with TF-IDF features. This model, marked distinctly in the table for emphasis, exhibits the highest scores across all metrics for both datasets. This outcome signifies a substantial advancement in classification capability, particularly in the context of nuanced sentiment analysis in NLP applications. In a parallel comparative assessment, we observe a marked improvement in classification accuracy when incorporating TF-IDF features into the SVM model. This improvement is especially pronounced in the date-wise dataset, indicating the enhanced capability of the SVM model when supplemented with TF-IDF for temporal data analysis.

Figure 3 presented, provides a visual and consolidated overview of these results, mapping the performance achievements of the various models across the different datasets. This graphical representation allows for an immediate visual grasp of the comparative efficacies, further enhancing the understanding of the models' performances in different analytical scenarios.

Table 4. Summary of the results of all classification models

ML model	Random dataset				Date-wise dataset			
	ACC	Pre	Rec	F1	ACC	ACC	Pre	Rec
KNN	57.4	57.2	57	57.1	44.1	45.3	44.9	45.1
DT	85.2	85.6	85.3	85.4	45.5	46.2	46.3	46.2
RF	88.9	89.1	89.2	89.1	45.6	45	45	45.0
ANN	44.8	44.6	45.4	45.0	55.7	55.9	55.8	55.8
SVM	89.1	90.5	90.9	90.7	44.6	44.4	43.6	44.0
SVM TF-IDF	96.8	96.8	96.8	96.8	94.6	94.6	94.6	94.6
LSTM	97.1	97.3	97.3	97.3	94.7	95.3	95.1	95.2
LSTM TF-IDF	98.3	98.4	98.1	98.2	97.8	97.7	97.7	97.7

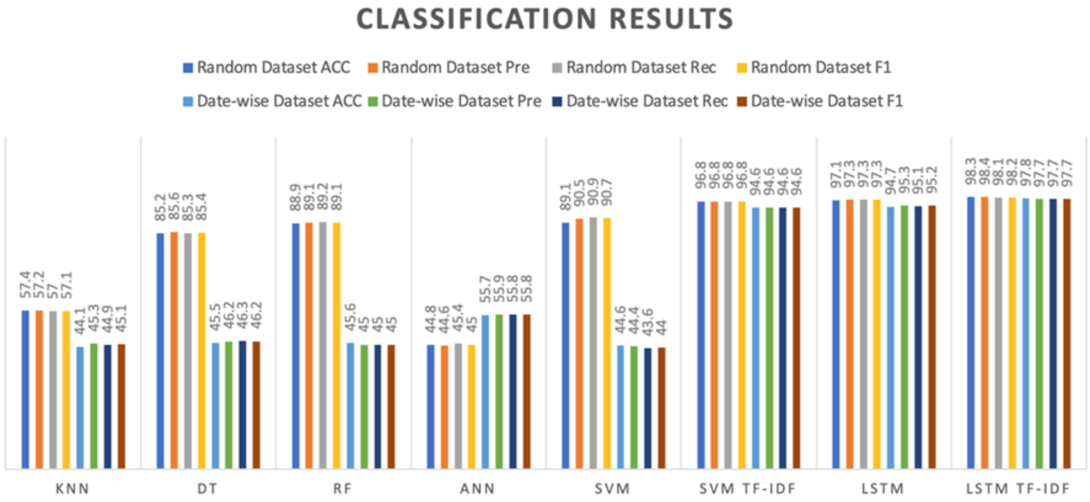

Figure 3. Comprehensive evaluation of model performance

In the realm of accuracy, our analysis revealed distinct patterns among the various models. The ANN and KNN models registered relatively lower performance metrics within the random dataset. Conversely, DT, and RF models exhibited comparably stronger performances, though DT trailed slightly behind RF in terms of accuracy. A standout observation was the notable efficacy of SVM, which not only outperformed the baseline models but also demonstrated remarkable enhancement when augmented with TF-IDF features. In this enhanced configuration, SVM's accuracy was paralleled by our LSTM model, with the LSTM-TFIDF variant ultimately emerging as the leading model in terms of accuracy.

When examining precision, a similar trend was observed. Both the ANN and KNN models showed the least precision, whereas the RF model exceeded the performance of DT. Significantly, the integration of TF-IDF features with SVM resulted in superior precision, outmatching the standard SVM configuration. The LSTM model's precision was comparable to that of the SVM-TFIDF, with the LSTM-TFIDF model achieving an outstanding precision rate of 98.4%.

The recall rates followed a similar trajectory as the accuracy and precision metrics. Both KNN and the standard SVM configurations recorded the lowest recall rates, while the ANN model showed a slight improvement. However, it was the SVM-TFIDF and LSTM configurations that demonstrated substantial increases in recall rates, with LSTM-TFIDF once again leading the pack. In terms of F-measure, which serves as a harmonized metric combining precision and recall, there was a clear trend of superiority with the SVM-TFIDF, LSTM, and LSTM-TFIDF models consistently outshining the other models. This pattern underscores the pivotal role of advanced feature engineering in enhancing model performance, especially in the context of NLP tasks.

Additionally, our focused efforts on refining the date-wise dataset, which initially showed underperformance, bore fruit. As demonstrated in Figure 3, models such as KNN, DT, RF, and standard SVM continued to show limited efficacy on this dataset. However, the ANN model exhibited a marked improvement in accuracy, though it remained at a moderate 55%. In sharp contrast, both the SVM-TFIDF and LSTM models demonstrated remarkable performances, each surpassing the 90% accuracy threshold. Notably, the LSTM-TFIDF variant exhibited marginally superior accuracy among these high-performing models.

In summary, Figure 3 captures a holistic view of the model performance across the different datasets used in our study. It distinctly highlights the importance of feature engineering, as evidenced by the enhanced performance of models employing TF-IDF features and LSTM in the context of NLP. This comprehensive analysis provides crucial insights into the effectiveness of various ML models in sentiment analysis, particularly in the challenging domain of social media sentiment towards COVID-19 vaccines.

5. CONCLUSION

This study embarked on a comprehensive journey to classify COVID-19 vaccination-related tweets into distinct sentiment categories: positive, negative, and neutral. Utilizing a dual-approach dataset—random and date-wise—we explored possible correlations between temporal factors and public opinions. Key to our methodology was the preprocessing of the dataset, involving the removal of extraneous symbols and stopwords, and the implementation of stemming techniques to refine the textual data. The employment of the TextBlob tool facilitated efficient labeling of tweets based on their sentiment. A focal point of our analysis was the testing and comparison of two advanced language processing algorithms. The integration of the TF-IDF algorithm proved instrumental in enhancing the feature quality and classification accuracy. Notably, the implementation of SVM and LSTM models yielded insightful results, with the LSTM-TFIDF combination demonstrating significant superiority, particularly in handling the date-wise dataset. This indicates LSTM's remarkable effectiveness in processing sequential data, marking a notable advancement in the field of NLP. The outcomes of this study underscore the potential of our proposed methodology in accurately classifying COVID-19 vaccination sentiments, offering valuable implications for public health communication strategies and policy-making in a digitalized world.

Reflecting on the insights and limitations encountered, this study paves the way for several promising research trajectories. Firstly, expanding the research scope to include tweets with diverse characteristics would offer broader validation of our methodology and generalize the findings. Additionally, incorporating factors like user attitudes toward COVID-19 could enhance the prediction models, aiding in understanding public responses to pandemic-related sanitary policies and practices in various settings. Moreover, the adaptability of our methodology could be tested against tweets focusing on other prevailing topics, thus broadening the horizon of understanding public sentiment in different contexts. This approach could unravel nuances in public opinions and sentiments on a range of issues, contributing to a more in-depth societal discourse analysis. Lastly, the exploration of more advanced ML techniques and algorithms stands as a fertile ground for future studies. Such advancements could further refine the accuracy and performance of sentiment analysis, particularly on social media platforms like Twitter. This pursuit would not only extend the boundaries of computational linguistics and social media analytics but also offer profound implications for data-driven decision-making in various sectors.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM) for supporting and funding this work (No. 6236500).

REFERENCES

[1] F. Erika and R. Nicole, "25.6% of the world has received a COVID vaccine. See how your country is doing," *Fortune*, 2021. https://fortune.com/2021/06/30/covid-vaccines-global-vaccination-rates-by-country/ (accessed Sep. 28, 2023).

[2] WHO, "COVID-19 symptoms," *Who.Int*, 2022. https://www.who.int/indonesia/news/novel-coronavirus/new-infographics/covid-19-symptoms (accessed Oct. 08, 2023).

[3] S. A. J. Zaidi, S. Tariq, and S. B. Belhaouari, "Future prediction of covid-19 vaccine trends using a voting classifier," *Data*, vol. 6, no. 11, p. 112, Nov. 2021, doi: 10.3390/data6110112.

[4] J. Mewhirter, M. Sagir, and R. Sanders, "Towards a predictive model of COVID-19 vaccine hesitancy among American adults," *Vaccine*, vol. 40, no. 12, pp. 1783–1789, Mar. 2022, doi: 10.1016/j.vaccine.2022.02.011.

[5] T. B. Shahi, C. Sitaula, and N. Paudel, "A hybrid feature extraction method for Nepali COVID-19-related Tweets classification," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–11, Mar. 2022, doi: 10.1155/2022/5681574.

[6] A. Alqarni and A. Rahman, "Arabic Tweets-based sentiment analysis to investigate the impact of COVID-19 in KSA: a deep learning approach," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 16, Jan. 2023, doi: 10.3390/bdcc7010016.

[7] K. Hayawi, S. Shahriar, M. A. Serhani, I. Taleb, and S. S. Mathew, "ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection," *Public Health*, vol. 203, pp. 23–30, Feb. 2022, doi: 10.1016/j.puhe.2021.11.022.

[8] N. Braig, A. Benz, S. Voth, J. Breitenbach, and R. Buettner, "Machine learning techniques for sentiment analysis of COVID-19-related Twitter data," *IEEE Access*, vol. 11, pp. 14778–14803, 2023, doi: 10.1109/ACCESS.2023.3242234.

[9] H. Lyu *et al.*, "Social media study of public opinions on potential COVID-19 vaccines: informing dissent, disparities, and dissemination," *Intelligent Medicine*, vol. 2, no. 1, pp. 1–12, Feb. 2022, doi: 10.1016/j.imed.2021.08.001.

[10] M. Qorib, T. Oladunni, M. Denis, E. Ososanya, and P. Cotae, "Covid-19 vaccine hesitancy: text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset," *Expert Systems with Applications*, vol. 212, p. 118715, Feb. 2023, doi: 10.1016/j.eswa.2022.118715.

[11] A. Muneer, S. M. Fati, N. A. Akbar, D. Agustriawan, and S. T. Wahyudi, "iVaccine-Deep: prediction of COVID-19 mRNA vaccine degradation using deep learning," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 9, pp. 7419–7432, Oct. 2022, doi: 10.1016/j.jksuci.2021.10.001.

[12] P. Sv, J. Tandon, Vikas, and H. Hinduja, "Indian citizen's perspective about side effects of COVID-19 vaccine – a machine learning study," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 15, no. 4, p. 102172, Jul. 2021, doi: 10.1016/j.dsx.2021.06.009.

[13] P. Chinnasamy, V. Suresh, K. Ramprathap, B. J. A. Jebamani, K. S. Rao, and M. S. Kranthi, "COVID-19 vaccine sentiment analysis using public opinions on Twitter," *Materials Today: Proceedings*, vol. 64, pp. 448–451, 2022, doi: 10.1016/j.matpr.2022.04.809.

[14] R. K. Godi *et al.*, "Sentiment analysis on Omicron Tweets using hybrid classifiers with multiple feature extractiontechniques and transformer based models," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 15s, pp. 257–275, 2024.

[15] N. Jacob and V. M. Viswanatham, "Sentiment analysis using improved atom search optimizer with a simulated annealing and ReLU based gated recurrent unit," *IEEE Access*, vol. 12, pp. 38944–38956, 2024, doi: 10.1109/ACCESS.2024.3375119.

[16] N. M. Abdulkareem, A. M. Abdulazeez, D. Q. Zeebaree, and D. A. Hasan, "COVID-19 world vaccination progress using machine learning classification algorithms," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 100–105, May 2021, doi: 10.48161/qaj.v1n2a53.

[17] S. Bhowmik, R. Sadik, W. Akanda, and J. R. Pavel, "Sentiment analysis with hotel customer reviews using FNet," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 13, no. 2, pp. 1298–1306, Apr. 2024, doi: 10.11591/eei.v13i2.6301.

[18] T. M. Lincoln *et al.*, "Taking a machine learning approach to optimize prediction of vaccine hesitancy in high income countries," *Scientific Reports*, vol. 12, no. 1, p. 2055, Feb. 2022, doi: 10.1038/s41598-022-05915-3.

[19] N. K, K. M, S. V. Easwaramoorthy, D. C R, S. Yoo, and J. Cho, "Hybrid approach of deep feature extraction using BERT–OPCNN & FIAC with customized Bi-LSTM for rumor text classification," *Alexandria Engineering Journal*, vol. 90, pp. 65–75, 2024, doi: 10.1016/j.aej.2024.01.056.

[20] S. Kumar, S. Kumar, A. Singh, and A. Raj, "COVID-19 data analysis and prediction using (machine learning) and vaccination update of India," *SSRN Electronic Journal*, 2021, doi: 10.2139/ssrn.3847564.

[21] T. Vaiyapuri *et al.*, "Sustainable artificial intelligence-based Twitter sentiment analysis on COVID-19 pandemic," *Sustainability*, vol. 15, no. 8, p. 6404, Apr. 2023, doi: 10.3390/su15086404.

[22] H. K. Obayes, K. H. Alhussayni, and S. M. Hussain, "Predicting COVID-19 vaccinators based on machine learning and sentiment analysis," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1648–1656, Jun. 2023, doi: 10.11591/eei.v12i3.4278.

[23] M. Kazemi, N. L. Bragazzi, and J. D. Kong, "Assessing inequities in COVID-19 vaccine roll-out strategy programs: a cross-country study using a machine learning approach," *Vaccines*, vol. 10, no. 2, p. 194, Jan. 2022, doi: 10.3390/vaccines10020194.

[24] R. Dutta, N. Das, M. Majumder, and B. Jana, "Aspect based sentiment analysis using multi-criteria decision-making and deep learning under COVID-19 pandemic in India," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 1, pp. 219–234, Mar. 2023, doi: 10.1049/cit2.12144.

[25] R. Miikkulainen *et al.*, "Evolving deep neural networks," in *Artificial Intelligence in the Age of Neural Networks and Brain Computing*, Elsevier, 2019, pp. 293–312. doi: 10.1016/B978-0-12-815480-9.00015-3.

BIOGRAPHIES OF AUTHORS

Ban Safir Khalaf completed her undergraduate studies in computer science at Ibn al-Haytham College, University of Baghdad, in 2013. Following her passion for advanced studies, she pursued and successfully obtained her master's degree in computer science with a focus on artificial intelligence from Universiti Putra Malaysia by 2023. During her master's program, her research was centered on the prediction of COVID-19 using machine learning algorithms. Her dedication and innovative approach led to significant advancements in algorithmic technology, marking her contributions as particularly noteworthy, especially in the challenging context of the global pandemic. She can be contacted at email: bansafir8989@gmail.com.

Hazlina Hamdan received a Ph.D. degree in artificial intelligence from the University of Nottingham, UK, in 2013. She is currently a senior lecturer with the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She has led several research projects funded by the Ministry of Higher Education (MOHE), and the research university grant scheme (GP). Her research areas include intelligent computing and applications, such as medical prognostic, pattern recognition, prediction systems, and optimization. She can be contacted at email: hazlina@upm.edu.my.

Noridayu Manshor received a bachelor's degree in computer science from UPM, a master's degree in computer science from UTM, and a Ph.D. degree in computational intelligence and systems science from USM. She is currently a Senior Lecturer with the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She has published numerous scientific journal articles and conference papers. Her research interests include pattern recognition, image processing, computer vision, and fuzzy sets. She is a regular reviewer of many journals. She can be contacted at email: ayu@upm.edu.my.