COVBERT: ENHANCING SENTIMENT ANALYSIS ACCURACY IN COVID-19 X DATA THROUGH CUSTOMIZED BERT

Vanshaj Gupta[1], Jaydeep Patel[1], Safa Shubbar[1], and Kambiz Ghazinour[2]

[1]Department of Computer Science, Kent State University, Kent, OH, USA
[2]Department of Cyber security, State University of New York at Canton, NY, USA

ABSTRACT

In a time when social media information is a valuable resource for gaining insights, the COVID-19 pandemic has released a flood of public sentiment, abundant with unstructured text data. This paper introduces CovBERT, a novel adaptation of the BERT model, specifically honed for the nuanced analysis of COVID-19-related discourse on X (formerly Twitter). CovBERT stands out by incorporating a bespoke vocabulary, meticulously curated from pandemic-centric tweets, resulting in a remarkable leap in sentiment analysis accuracy—from the baseline 72\% to an impressive 78.64\%. This paper not only presents a detailed comparison of CovBERT with the standard BERT model but also juxtaposes it against traditional machine learning approaches, showcasing its superior proficiency in decoding complex emotional undercurrents in social media data. Furthermore, the integration of geolocation analysis pipeline adds another layer of depth, offering a panoramic view of global sentiment trends.

KEYWORDS

BERT, CovBERT, COVID-19, Sentiment Analysis, X (Twitter) Data Analysis, Natural Language Processing, Machine Learning, Geolocation Analysis, Social Media Analytics, Data Mining

1. INTRODUCTION

The onset of the COVID-19 pandemic has not only transformed the global health landscape but has also dramatically altered public discourse on social media platforms. X previously known as (Twitter), a prolific source of real-time public sentiment, has emerged as a critical medium for understanding societal responses to this unprecedented crisis. However, the unique linguistic nuances and evolving terminology associated with the pandemic pose significant challenges for conventional sentiment analysis models. It's crucial to distinguish between 'expression', the actual wording or content of communication, and 'sentiment', which refers to the underlying emotions or attitudes. This distinction is fundamental in accurately interpreting the public's reactions and sentiments during the pandemic.

In this context, we introduce CovBERT, an innovative adaptation of the BERT (Bidirectional Encoder Representations from Transformers) model [1], tailored to capture the dynamic and pandemic-specific vernacular on X (Twitter). CovBERT is a specialized application within the broader field of NLP. While NLP encompasses a wide range of technologies and methodologies used for understanding and interpreting human language, CovBERT specifically applies these principles to the domain of COVID-19 related discourse. It leverages the foundational NLP technologies present in the BERT model, adapting them to the unique linguistic challenges posed by the pandemic. CovBERT's innovation stems from its ability to understand the unique, evolving language of the pandemic. By enriching the vocabulary with COVID-19 specific terms,

David C. Wyld et al. (Eds): IOTBS, NLTM, AIMLA, DBDM -2024
pp. 171-182, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140212

CovBERT not only captures the dynamic nature of pandemic-related discussions but also overcomes the limitations of standard BERT models in interpreting the nuanced sentiments in tweets. This tailored approach enables CovBERT to offer a more precise and contextually rich sentiment analysis, showcasing its distinctiveness in addressing the challenges of pandemic-era social media discourse. Moreover, the integration of geolocation data provides a multi-dimensional perspective of sentiment analysis, allowing for a geographically informed understanding of public mood and opinions. Such an approach is particularly valuable in analysing and mapping the diverse reactions across different regions to various stages of the pandemic and associated events. This makes CovBERT not just a tool for sentiment analysis but a pioneering model in contextual understanding during extraordinary global circumstances. This paper presents a comprehensive comparison of CovBERT with the conventional BERT model and other machine learning techniques in sentiment analysis. Through rigorous experimentation and analysis, we demonstrate that CovBERT not only significantly improves sentiment analysis accuracy but also offers nuanced insights into public sentiment during these challenging times. Our contribution is twofold: firstly, CovBERT represents a significant step forward in the field of Natural Language Processing (NLP) by addressing the specific challenges of COVID-19 sentiment analysis [2]. Secondly, the inclusion of geolocation analysis paves the way for a more holistic and contextually rich exploration of social media data, especially in scenarios of global crises. Thus, CovBERT serves as a valuable tool for researchers, policymakers, and public health officials in gauging public sentiment and formulating responsive strategies.

2. RELATED WORK

Recent advancements in Natural Language Processing (NLP) have seen the development of sophisticated models and datasets, particularly for sentiment analysis in specialized domains like public health and finance. This review examines key contributions in this area, especially in the context of the COVID-19 pandemic and financial sentiment analysis. Rustam et al. [3] explore the realm of sentiment analysis in COVID-19-related tweets. Their study underscores the importance of preprocessing and the effectiveness of combining TF-IDF [4] and bag-of-words model [5] features. Using machine learning techniques, particularly the Extra Trees Classifier, they achieve notable accuracy in sentiment classification, emphasizing the critical role of feature engineering in enhancing model performance. In a similar vein, Kaur et al. [6] introduce the Hybrid Heterogeneous Support Vector Machine (H-SVM) for analyzing X (Twitter) sentiments during the COVID19 pandemic. H-SVM's unique blend of linear and nonlinear kernels allows it to outperform traditional SVM [7] and RNN [8] models. The paper demonstrates HSVM's ability to capture complex data patterns, significantly improving precision and recall, marking a substantial advancement in sentiment analysis methodologies. Qazi et al. [9] contribute to this field with the GeoCoV19 dataset, comprising over half a billion tweets related to COVID-19. This extensive dataset offers invaluable insights for disease surveillance and public sentiment analysis. The study's focus on geolocation inference from multilingual tweets provides a novel approach to understanding public discourse in a global health crisis. In the domain-specific model adaptation, M¨uller et al. [2] introduce COVIDTwitter-BERT (CT-BERT), a transformer-based model pre-trained on a vast corpus of COVID-19 X (Twitter) data. CT-BERT demonstrates superior performance in classifying COVID-19 content over general NLP models like BERT-LARGE, validating the effectiveness of domain-specific pre-training in capturing the nuances of pandemic-related language. Similarly, Yang et al. [10] address the challenges of financial sentiment analysis with FinBERT, a BERT-based model specifically tuned for the financial sector. Despite the specialized language and scarcity of labelled data in finance, FinBERT shows remarkable improvement on two financial sentiment analysis datasets. This highlights the potential of fine-tuned, domain-specific models to outperform traditional machine learning methods, even with limited training data. These studies collectively illustrate the rapid evolution of NLP techniques and their applications in understanding complex, domain-specific sentiment.

The success of these models in extracting meaningful insights from large-scale, unstructured data sets a promising direction for future research in sentiment analysis and beyond.

3. Data Analysis

3.1. Geolocation Analysis

Data Preparation and Exploration The analysis commences with the geo data dataset comprising COVID-19 related tweets. Initial exploration reveals the dataset's structure, with a focus on the presence of non-null values across different columns. This step is crucial for ensuring data integrity and guiding further analysis.

Feature Engineering The raw user location data undergoes a transformation process. Using a comprehensive city database, locations are mapped to corresponding countries, enhancing the dataset's structure and facilitating more meaningful geographic analysis. Visualization and Insights Visualizing the most active user locations based on tweet counts provides initial insights. This is followed by a refined analysis focusing on countries after feature engineering, offering a clearer view of the geographic spread of the conversation on COVID-19. Temporal and Geographic Spread The dynamic nature of the discussion is captured through animations illustrating the spread of tweets over time. This is further enriched by a comparison with the progression of confirmed COVID-19 cases, shedding light on the global response to the pandemic.

3.2. Sentiment Analysis

Data Preparation and Preprocessing The sentiment data dataset is subjected to rigorous preprocessing, including text cleaning and encoding. The cleaning process involves removing emojis, hashtags, and links, and filtering non-English content, resulting in a standardized text dataset. Sentiments are then encoded numerically to facilitate computational analysis. Visualization of Sentiment Distribution A word cloud visualization is used to display the 50 most common words in the tweets, providing a visual representation of the prevailing themes and sentiments in the data as can be seen in Figure 3. Additionally, the sentiment distribution within the dataset is examined to understand the balance and diversity of emotional expressions. Label Encoding and Dataset Splitting Label encoding is employed to convert sentiment categories into a numerical format. The dataset is then split into training, validation, and test sets, ensuring a balanced representation of sentiments across these subsets. This step is crucial for the unbiased training and evaluation of the sentiment analysis models.

4. Methodology

Our study's methodology encompasses two primary areas: geolocation analysis and sentiment analysis of COVID-19 related social media posts. The approach is designed to uncover regional sentiments during the pandemic, leveraging advanced data processing and machine learning techniques.

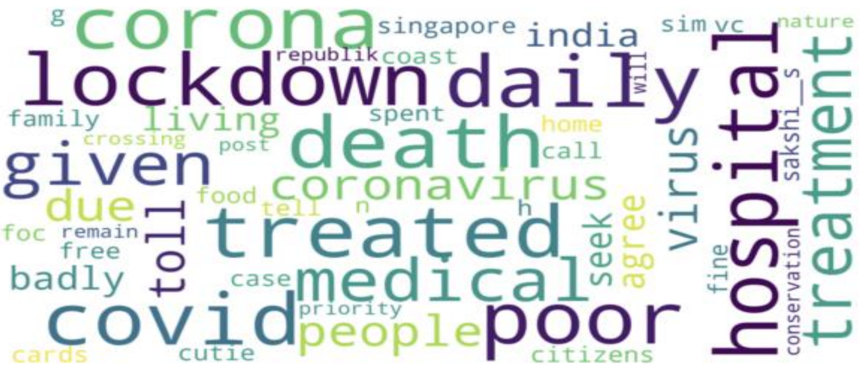


Fig. 1: A word cloud of the 50 most common words

4.1. Methodology for Geolocation Analysis

The geolocation analysis aimed to map and understand the geographical spread of COVID-19 related discussions on the platform ’X’. Data Standardization The raw location data in the dataset varied widely, including countries, cities, and specific locations. The heterogeneity of this data posed a challenge for accurately gauging tweet concentrations by country as seen in Figure 1. The first step was to standardize this heterogeneous data into a uniform country format. We employed the ’world cities’ dataset as a reference to map various location entries to their corresponding countries, ensuring data uniformity and consistency. Visualization of Tweet Concentration by Country Following standardization, we visualized the concentration of tweets on a global scale as seen in Figure 2. This involved plotting a graph highlighting the countries with the most significant number of COVID-19 related tweets. The visualization served as an illustrative representation of regions where the pandemic discourse was most intense. Temporal Analysis and Animation We performed a temporal analysis to capture the evolution of X (Twitter) conversations over time. By plotting the cumulative count of tweets for each country across successive dates, we created a time-lapse animation. This animated map vividly depicted the spread and intensity of discussions related to COVID-19 on a global scale.

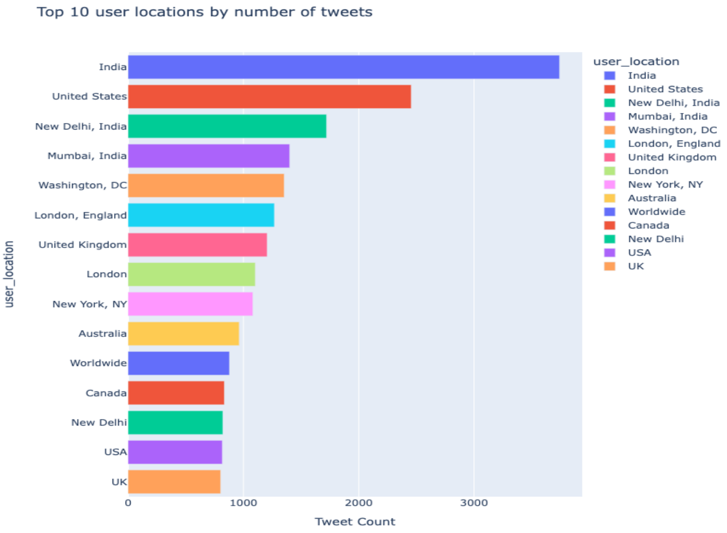


Fig. 2: Tweet Concentration by Country before Standardization

Comparison with COVID-19 Spread In an effort to correlate social media activity with the actual pandemic progression, we juxtaposed our animated tweet data against the spread of confirmed COVID-19 cases. This comparison aimed to understand the relationship between the volume of tweets and the real-world development of the pandemic.

4.2. Methodology for Sentiment Analysis

Sentiment analysis was conducted to explore the range of emotions expressed in the public discourse during the pandemic.

4.2.1. Data Preprocessing for Sentiment Analysis

Effective preprocessing was critical for ensuring the quality of sentiment analysis. Our preprocessing steps included:

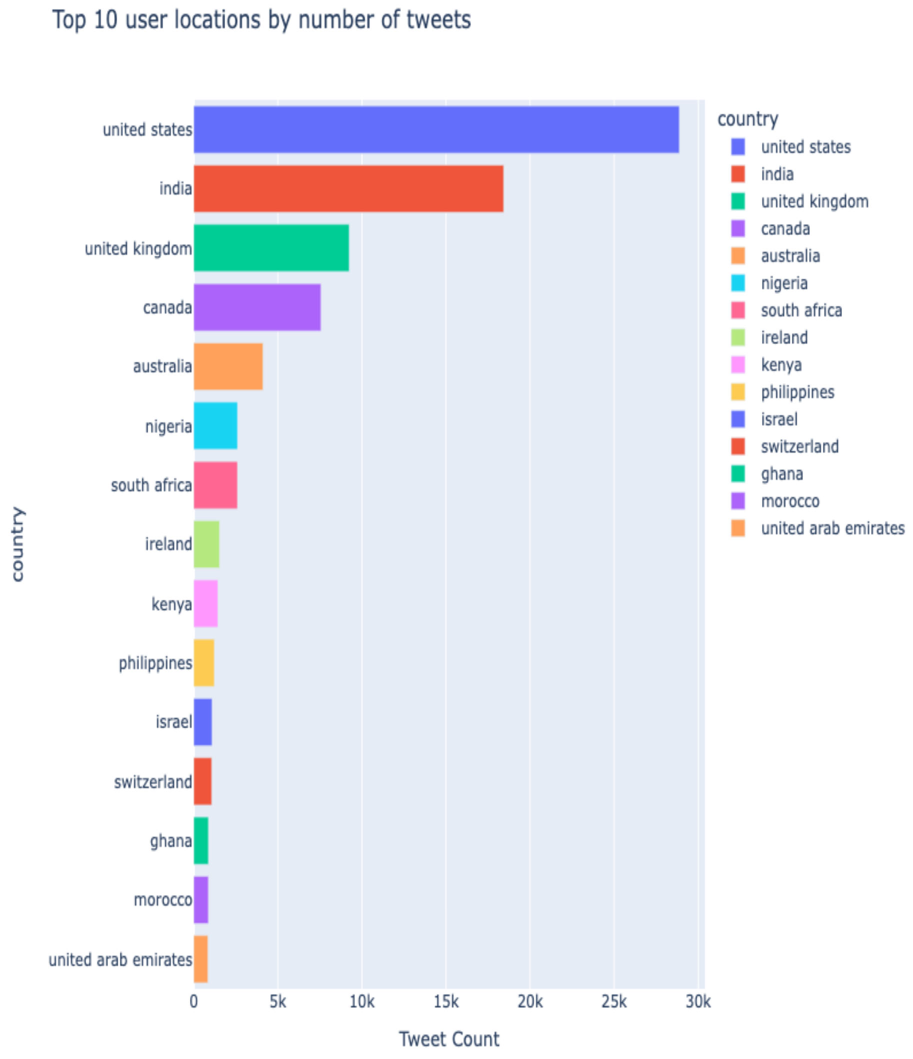


Fig. 3: Tweet Concentration by Country after Standardization

1. **Emoji Removal:** Employing regular expressions, we developed a function to detect and remove a wide array of emojis, including emoticons, symbols, and flags, to streamline the text for analysis.
2. **Text Cleaning**: This multifaceted step involved replacing newline characters with spaces, removing links and mentions, stripping non-ASCII characters, eliminating punctuations, and filtering out commonly used stop words.
3. **Hashtag Handling:** Recognizing the sentiment value of hashtags, we removed hashtags at the end of tweets while preserving those within the tweet text.
4. **Special Character and URL Filtering:** We cleaned words containing special characters like ’&’ and ’$’, and removed URL shorteners to maintain text integrity.
5. **Tokenization:** is a fundamental process in NLP where text is divided into smaller units, or tokens (typically words). In sentiment analysis, tokenization was performed to break down complex texts into manageable pieces, allowing the model to analyse and interpret the sentiment of each part. This step was crucial for understanding the overall sentiment conveyed in a text, as it enabled the model to process and assess the emotional tone of individual words and phrases within the larger context.
6. **Encoding and Model Implementation:** The processed and tokenized text data was transformed into numerical encodings using the Label Encoder from the sklearn library. This encoding step is vital for converting text data, inherently non-numerical, into a format interpretable by machine learning models. Computers process numerical values, not text, so encoding text into numbers enables these models to analyse, learn from, and make predictions based on the textual data. This transformation is a key step in ensuring that the intricate nuances of language are effectively captured and utilized in our machine learning analyses. Naive Bayes Classifier The Naive Bayes Classifier [11], a fundamental machine learning algorithm, was employed for initial sentiment classification. The model achieved an accuracy of 69%, with F1-scores ranging from 0.61 to 0.75 for different sentiment categories. A confusion matrix provided further insights into the model’s performance. Support Vector Machine (SVM) The SVM model [12] was employed for its capability in high-dimensional spaces, using optimal hyperplanes for classification. SVM achieved a 69% accuracy, indicating its effectiveness in sentiment classification. Logistic Regression Logistic Regression model [13] was chosen for its efficiency in categorical data analysis. It was particularly suitable for our multi-class sentiment analysis task. Logistic Regression demonstrated a 70% accuracy, showcasing its reliability in handling categorical sentiment data. Long Short-Term Memory (LSTM) Network The LSTM model [14] was utilized for its proficiency in processing sequential data and remembering long-term dependencies. This was crucial for capturing the contextual nuances in tweet sentiments. Although LSTM presented a lower accuracy of 60%, its performance was noteworthy in terms of capturing the sequence and context within the data. BERT Model BERT’s [15] deep bidirectional nature was leveraged to effectively understand contextual relationships in tweets, making it an ideal choice for complex sentiment analysis tasks. BERT significantly outperformed other models with an accuracy of 77%, demonstrating its superior capability in comprehending and classifying sentiments. The CovBERT Model is a customized adaptation of the pre-trained BERT model [15], was specifically fine-tuned for sentiment analysis of COVID19 related tweets. The implementation involved comprehensive text preprocessing, including cleaning, lemmatization, and special character filtering, to prepare the tweets for analysis. The tokenizer was expanded with new tokens pertinent to COVID-19, enhancing the model’s ability to understand pandemic-specific discourse. The CovBERT model underwent several epochs of training, with performance optimization using AdamW and learning rate scheduling. Throughout the training process, the model demonstrated significant improvements in both training and validation accuracy, indicating its effectiveness in accurately capturing sentiments related to the

COVID-19 pandemic. This comprehensive methodology outlines the analytical steps and technical approaches undertaken in our study, providing a clear framework for understanding the spatial-temporal spread of COVID-19 discussions and the public's sentiment during this global health crisis.

5. COVBERT FOR COVID-19 DOMAIN

In this subsection, we describe our implementation of BERT for the COVID-19 domain, known as CovBERT. This includes details on further pre-training on a domain-specific corpus, data preprocessing tasks, and the model architecture employed.

5.1. Further Pre-training

As suggested by Howard and Ruder (2018) [16], further pre-training a language model on a domain-specific corpus can enhance classification performance. To adapt BERT to the COVID-19 domain, we followed a similar approach.

5.2. Data Preprocessing

Initial data preprocessing involved cleaning and standardizing the textual content from the tweets. This process included the removal of emojis, links, mentions, and non-ASCII characters. Additionally, text was normalized by converting to lowercase, removing punctuations, and filtering out stopwords. Hashtags were handled by retaining the textual content but removing the hash symbol. Text was also lemmatized to bring words to their base forms.

5.3. Custom Vocabulary for COVID-19

A key aspect of adapting CovBERT to the COVID-19 domain involved developing a custom vocabulary. This was crucial, as the onset of the pandemic introduced a new lexicon, which was non-existent prior to the virus's detection. The vocabulary evolved in real-time with the global health crisis, including terms related to symptoms, measures, and impacts. Our approach involved enriching the BERT model's tokenizer with tokens identified from our COVID-19 tweet dataset.

5.4. BERT Model Adaptation

The core of CovBERT is based on the pre-trained BERT model. We utilized the BERT base uncased model, which was further adapted to the specificities of our dataset. The tokenizer from the base model was extended by adding new tokens identified from our tweet dataset. These tokens were added to better capture the nuances of COVID-19 related discourse.

5.5. Dataset Preparation and Tokenization

The dataset for model training and validation was split into training and test sets, ensuring a proper representation of sentiments. Tweets were tokenized using the adapted tokenizer, with special attention to handle the maximum token length and attention masks for BERT.

5.6. Model Architecture

CovBERT's architecture entailed integrating BERT as the foundational layer. A dropout layer was added for regularization, followed by a linear layer for sentiment classification. The model was set to output four classes, corresponding to different sentiment labels.

5.7. Training

The model was trained using the AdamW optimizer with a linear learning rate scheduler. Loss was computed using the Cross-Entropy Loss function, suitable for multi-class classification tasks. Model performance was evaluated on the validation set, with accuracy being the primary metric. Training involved multiple epochs, ensuring sufficient learning while monitoring for overfitting.

5.8. Evaluation and Results

Over a training period of 10 epochs, CovBERT demonstrated significant improvements in accuracy. The model achieved a training accuracy of 78.64%, while the validation accuracy peaked at 99.96%. The training history, depicted in Figure [reference figure], highlights the model's performance across epochs, showcasing the efficacy of the implemented training strategies.

6. Inherent Limitations of Standard BERT

Prior to discussing the specifics of CovBERT, it's important to acknowledge the inherent limitations of the standard BERT model in certain contexts. While BERT is a powerful tool for natural language understanding, it has constraints that are particularly evident in unique scenarios such as a global pandemic: Pre-Defined Vocabulary: The BERT model's vocabulary is pre-defined and static, which may not encompass newly emerged terms or jargon, especially in fast-evolving situations. Contextual Flexibility: BERT's pre-training on a historical corpus might not fully capture the nuances of recent events or specific domains that rapidly evolve. Computational Requirements: The extensive computational resources required for training and deploying BERT models can be a limitation in certain environments. Generalization Bias: While BERT is designed for broad applicability, this can sometimes result in a bias towards more general language use, potentially overlooking niche or domain-specific nuances.

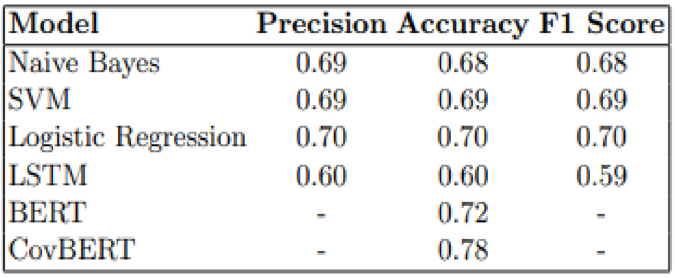
7. Implementation Details

CovBERT, an enhanced BERT-based model, integrates a pre-trained BERT base uncased model [15] with a customized tokenizer and additional neural layers. The model architecture comprises the BERT base layer for feature extraction, an extended tokenizer including COVID-19 specific terms, a dropout layer (0.3 rate) for regularization, and a linear output layer for classifying sentiments into four categories. The model's tokenizer is adapted to grasp pandemic-related nuances, crucial for accurate sentiment analysis in the context of COVID-19. Data preprocessing involves standardizing tweet text by removing emojis, applying lemmatization, and filtering special characters, ensuring clean and relevant input for model training. The training of CovBERT utilizes the ADAMW (ADAM with weight decay) optimizer [17] with a learning rate of 2e-5 and a linear learning rate scheduler, over 10 epochs. Cross-Entropy Loss functions as the loss metric, fitting for the multi-class classification nature of the task. Regularization techniques, including dropout and L2 regularization, are employed to prevent overfitting. The model's performance is evaluated based on accuracy, supplemented by precision, recall, and F1-score, with periodic validation checks on a separate dataset to ensure generalizability and robustness. This meticulous implementation ensures that CovBERT effectively captures the complexities of COVID-19 related sentiments in social media discourse.

8. EXPERIMENTAL RESULTS

The experimental analysis encompassed a variety of machine learning models, evaluating their performance in sentiment analysis tasks. We considered traditional algorithms such as Naive Bayes, Support Vector Machine (SVM), and Logistic Regression, alongside advanced neural network architectures like Long Short-Term Memory (LSTM) networks, and transformer-based models such as BERT and CovBERT. Naive Bayes, known for its simplicity and speed, achieved a precision and F1 score of 0.69, with an overall accuracy of 0.68. The SVM, a robust classifier capable of handling nonlinear data, matched the Naive Bayes in precision and slightly outperformed it in F1 score and accuracy. Logistic Regression, often preferred for its interpretability and efficiency, marginally surpassed both Naive Bayes and SVM with a consistent score of 0.70 across precision, accuracy, and F1 score. The LSTM network, despite its ability to capture long-term dependencies, lagged with a precision of 0.60 and an F1 score of 0.59, indicating potential difficulties in modelling the intricacies of natural language within tweet data. BERT, which revolutionized the field of NLP through its deep bidirectional understanding, showed a marked improvement in accuracy at 0.72. However, precision and F1 score metrics for BERT were not reported, possibly due to experimental constraints or focus on other performance measures. Our proposed CovBERT model, a refined BERT variant fine-tuned for COVID19 sentiment analysis, demonstrated superior performance with the highest accuracy of 0.78. Although precision and F1 scores for CovBERT are not presented, its accuracy indicates a significant improvement over both traditional machine learning models and advanced neural networks. The enhanced accuracy of CovBERT can be attributed to its domain-specific pre-training on COVID-19 related discourse, enabling it to better capture the nuances and sentiments expressed in pandemic related social media data.

Table 1: Comparative performance of different models on sentiment analysis.



Model	Precision	Accuracy	F1 Score
Naive Bayes	0.69	0.68	0.68
SVM	0.69	0.69	0.69
Logistic Regression	0.70	0.70	0.70
LSTM	0.60	0.60	0.59
BERT	-	0.72	-
CovBERT	-	0.78	-

The results underscore the importance of tailored models for domain-specific tasks. CovBERT's emphasis on a pandemic-centric lexicon and context not only provides a more accurate sentiment assessment but also establishes a framework for future research into crisis-related sentiment analysis. Further research might explore the integration of additional linguistic features and the expansion of datasets to enhance the precision and F1 score metrics, which will provide a more comprehensive understanding of CovBERT's performance benefits.

9. CONCLUSIONS

In this work, we presented CovBERT, a bespoke machine learning model, adapted from the BERT architecture, meticulously fine-tuned for sentiment analysis within the domain of COVID-19-related discourse. Our extensive experimentation and rigorous data analysis have validated the superiority of CovBERT over a range of traditional machine learning approaches and standard neural network models. This superiority is not only in terms of accuracy but also in the model's ability to discern the subtle nuances of pandemic-centric communication on social media.

CovBERT stands as a testament to innovation in sentiment analysis, particularly in the context of COVID-19. Its unique approach in enriching the BERT model with pandemic-specific vocabulary, combined with its ability to accurately interpret the subtle nuances of pandemic-centric communication, sets it apart. The incorporation of geolocation analysis further enhances its utility, allowing for a nuanced understanding of global conversation patterns related to COVID-19. The adaptability of CovBERT to the dynamic nature of online communication, and its potential in other rapidly evolving domains, underscores its innovative character. The geolocation analysis, a critical component of this study, has laid the groundwork for understanding the global conversation patterns about the pandemic. Through sophisticated data preprocessing, which included emoji removal, lemmatization, and the elimination of non-English tweets, we ensured the quality and reliability of the input data. The implementation of feature engineering techniques refined the raw data, which enabled us to draw a more precise and detailed picture of the geographical distribution of COVID-19-related tweets. The animated visualizations of tweet distribution over time juxtaposed against the spread of the pandemic provided a compelling view of the interplay between online discourse and the progression of global events. As we look to the future, several areas beckon further exploration and development. First and foremost, the precision and F1 scores for CovBERT need a thorough examination to ensure a comprehensive understanding of the model's performance. Enhancing the model to support multiple languages would dramatically extend its utility for global sentiment analysis, providing an invaluable tool for international public health monitoring and communication efforts. Moreover, the dynamic nature of online communication, particularly during crises, suggests the need for models that can adapt in real time to the evolving linguistic landscape. Extending the application of CovBERT to other domains marked by rapid change, such as political discourse, environmental activism, or emergency responses, could yield critical insights and assist in real-time decision-making. Furthermore, the development of an iterative model update mechanism is essential. Such a mechanism would allow CovBERT to continually learn from new data, reflecting the latest linguistic and semantic shifts, thus maintaining its effectiveness and relevance. The integration of CovBERT into real-time monitoring systems could provide instant sentiment analysis feedback, informing public policy and communication strategies as events unfold. Lastly, the potential for CovBERT in predictive analytics should not be overlooked. By correlating sentiment trends with subsequent developments in the pandemic, future iterations of CovBERT could contribute to predictive models that anticipate public health trends or misinformation spread. The implications of such capabilities are profound, offering a proactive stance in managing public health crises and the flow of information during such events. In future developments, CovBERT's application could be extended to include the identification of misinformation. This can be achieved by training the model on datasets containing verified information alongside known examples of misinformation related to COVID-19. By learning the linguistic patterns and discrepancies typical of false information, CovBERT can evolve to distinguish between factual and misleading content. Furthermore, integrating natural language inference capabilities would enable CovBERT to assess the credibility of statements within the context of established facts. This advancement would significantly contribute to public health communication strategies, aiding in the swift identification and mitigation of harmful misinformation during health crises. The journey of CovBERT is just beginning. As it stands, it is a potent tool for sentiment analysis, finely attuned to the discourse of our times. As we continue to refine and expand its capabilities, its potential to inform, guide, and predict becomes all the more promising. With further research, dedication, and interdisciplinary collaboration, CovBERT can evolve into an even more versatile and indispensable asset in the domain of data-driven public health intelligence.

REFERENCES

[1] M. Singh, A. K. Jakhar, and S. Pandey, “Sentiment analysis on the impact of coronavirus in social life using the bert model,” Social Network Analysis and Mining, vol. 11, no. 1, p. 33, 2021.

[2] M. M¨uller, M. Salath´e, and P. E. Kummervold, “Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter,” Frontiers in Artificial Intelligence, vol. 6, p. 1023281, 2023.

[3] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood, and G. S. Choi, “A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis,” Plos one, vol. 16, no. 2, p. e0245909, 2021.

[4] A. Aizawa, “An information-theoretic perspective of tf–idf measures,” Information Processing & Management, vol. 39, no. 1, pp. 45–65, 2003.

[5] Y. Zhang, R. Jin, and Z.-H. Zhou, “Understanding bag-of-words model: a statistical framework,” International journal of machine learning and cybernetics, vol. 1, pp. 43–52, 2010.

[6] H. Kaur, S. U. Ahsaan, B. Alankar, and V. Chang, “A proposed sentiment analysis deep learning algorithm for analyzing covid-19 tweets,” Information Systems Frontiers, pp. 1–13, 2021.

[7] T. Joachims, “Making large-scale svm learning practical,” Technical report, Tech. Rep., 1998.

[8] W. Yin, K. Kann, M. Yu, and H. Sch¨utze, “Comparative study of cnn and rnn for natural language processing,” arXiv preprint arXiv:1702.01923, 2017.

[9] U. Qazi, M. Imran, and F. Ofli, “Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information,” SIGSPATIAL Special, vol. 12, no. 1, pp. 6–15, 2020.

[10] Y. Yang, M. C. S. Uy, and A. Huang, “Finbert: A pretrained language model for financial communications,” arXiv preprint arXiv:2006.08097, 2020.

[11] T. Bayes, “Naive bayes classifier,” Article Sources and Contributors, pp. 1–9, 1968.

[12] C.Cortes and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, pp. 273–297, 1995.

[13] D.R. Cox, “The regression analysis of binary sequences,” Journal of the Royal Statistical Society Series B: Statistical Methodology, vol. 20, no. 2, pp. 215–232, 1958. [14] L. S.-T. Memory, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 2010.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[15] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” arXiv preprint arXiv:1801.06146, 2018.

[16] Z. Yao, A. Gholami, S. Shen, M. Mustafa, K. Keutzer, and M. Mahoney, “Adahessian: An adaptive second order optimizer for machine learning,” in proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 12, 2021, pp. 10 665–10 673.

AUTHORS

Vanshaj Gupta is currently pursuing a master’s in computer science at Kent State University, specializing in Artificial Intelligence. His journey into AI started with hands-on experience as a healthcare software developer. Gupta holds a bachelor’s degree in computer science from the Vellore Institute of Technology, India. Throughout his computer science journey, he has consistently showcased a passion for integrating technological advancements into real-world applications, aiming to make meaningful contributions to the field of AI.

Safa Shubbar is currently working towards a Ph.D. in Computer Science at Kent State University. Driven by a passion for research, she has made notable contributions to various refereed scientific publications. Safa's research interests encompass Bioinformatics, Data Analysis, Algorithms, Machine Learning, and Visualization, showcasing a commitment to advancing knowledge and making meaningful contributions to the field in her academic pursuits.

Jaydeep Patel is a graduate student in Computer Science at Kent State University, Ohio, hailing from India. His journey is marked by a deep passion for learning technologies and acquiring discrete skills. Jaydeep thrives on embracing new aspects of life's adventures, demonstrating a keen enthusiasm for personal and academic growth. His academic pursuits and love for technology reflect a commitment to continuous learning and exploration.

Dr. Kambiz Ghazinour is the Associate Professor and Chair of the Department of Cybersecurity at the State University of New York in Canton. He is also leading the Advanced Information Security and Privacy Research Lab. He has over 72 peer-reviewed publications and has advised over 60 dissertations, theses and projects in AI, Health Information, and Usable Security and Privacy. Dr. Ghazinour is also Founder and CSO of CyberSpara Inc., a company that focuses on Cybersecurity awareness through simulation-based trainings.