Semiconductor Optoelectronics, *Vol. 42 No. 2 (2023), 1468-1480*
https://bdtgd.cn/

TWEET-BASED SENTIMENT ANALYSIS AND FORECASTING FOR THE COVID-19 PANDEMIC

Maninder Singh
Department of Computer Science, Punjabi University, Patiala

Abstract
The global impact of the novel coronavirus disease (COVID-19) has significantly affected people worldwide. Each nation has implemented necessary precautions against this highly contagious disease due to limited vaccine access and the absence of a straightforward, effective COVID-19 therapy. Consequently, individuals are increasingly turning to online social networking platforms (e.g., Facebook, Reddit, LinkedIn, and Twitter) to share their perspectives on COVID-19. This study focused on analyzing user sentiments related to COVID-19 using a Twitter dataset. For a period of 36 days (from 25 July to 29 August 2020), I acquired a dataset of COVID-19-related Twitter posts from Kaggle to conduct sentiment analysis. Multiple machine learning (ML) strategies were employed to classify user sentiments about COVID-19. The dataset was initially categorized into three sentiment ratings: positive, negative, and neutral, to train various ML algorithms for predicting user concerns regarding COVID-19. Feature extraction methods such as Word2Vec and TF-IDF were utilized in this study. Results indicated that Word2Vec, coupled with a random forest classifier, yielded superior outcomes.

Keywords: COVID-19; sentiment analysis; machine learning; neural network; natural language processing

1. Introduction

COVID-19, also known as the coronavirus illness, emerged recently in Wuhan, Hubei Province, China, and rapidly disseminated worldwide. On March 11, 2020, the World Health Organization declared it a pandemic due to its continued global spread. Despite the significant impact on millions of lives, numerous countries implemented lockdowns of varying durations to curb the virus's transmission [23]. Throughout this period, people extensively utilized social media platforms to express their thoughts, opinions, and feedback on COVID-19. Platforms like Twitter experienced a substantial surge in pandemic-related tweets within a short timeframe.

Amidst the isolation period, individuals express their emotions on social media, where real-time and valuable information about COVID-19 is available. However, social media data can sometimes be unhelpful or misleading. Encountering inaccurate or negative information exacerbates people's distress. With "staying at home," "work from home," and "isolation time" becoming the new norm, social networking platforms are widely used for sharing news, ideas, emotions, and advice. Instances of misinformation, where individuals attempt to confuse or mislead with false or irrelevant information, are prevalent. For instance, claims like "eating

bananas prevents COVID-19" contribute to misinformation. Individuals affected by the illness undergo various physical and mental changes, necessitating the swift application of logical techniques to understand informative data streams. While distinguishing relevant material from the vast and noisy data on platforms like Twitter and Facebook can be challenging, cleaning this data reveals human feelings, emotions, expressions, and thoughts. Thorough examination provides insights into the current state of mind, attitude, and nature of a significant human population. The growing number of social media users, relying on it for educational content, contributes to the expanding volume of data. This paper focuses on employing Natural Language Processing (NLP) with various AI algorithms to extract useful information effectively [7]. However, challenges in determining inherent importance using NLP strategies, such as contextual phrases and words, along with ambiguity in text or speech, necessitate the use of ML-based algorithms [12][9][5]. Utilizing sentiment analysis of Twitter data from Kaggle, this paper explores public attitudes to investigate the escalating concern about coronaviruses.

The paper unfolds in the following manner. The initial segment of Section 1 serves as an introduction. Section 2 offers a succinct overview of pertinent literature. In Section 3, a comprehensive elucidation of the entire process is presented. The section 4 delves into the discussion of experimental results. Section 5 encapsulates the findings and outlines potential avenues for future research.

2. Related Works

Since the onset of the coronavirus pandemic, researchers have extensively discussed its causes, consequences, and trends. This section presents the sentiment analysis of tweets conducted through various machine learning techniques. The difficulty lies in discerning valuable information from the noise in the data.

Another study focused on the subjects and emotions expressed by people on Twitter regarding COVID-19. Researchers collected tweets related to COVID-19 and classified them as positive, negative, or neutral. They employed various feature sets and classifiers to determine the prevailing sentiment in the tweets. The Bidirectional Encoder Representations from Transformers (BERT) model demonstrated the highest accuracy (94.80%) and served as the sole assessment metric in this study due to its precision. While classification accuracy is a common starting point, it alone may not adequately assess a model's efficacy. Therefore, precision, recall, and F1-score, in addition to accuracy, must be considered to validate the model's performance [8].

An alternative study delved into the psychological impact of COVID-19, aiming to scrutinize the intricacies of human behavior and mood [15]. The analysis indicated that news about COVID-19 has induced heightened anxiety and crisis feelings among individuals due to the ramifications of the coronavirus. Numerous studies have dissected the economic repercussions of the industrial crisis and the COVID-19 issue on diverse industries and nations [4]. Given the extensive data accumulated from various social media platforms in recent years, sentiment analysis based on tweets has found application across multiple domains, encompassing YouTube, Reddit, Facebook, and Twitter [1]. The research uncovers challenges associated with the gathered data [3]. Diverse machine learning (ML) and deep learning (DL) classifiers scrutinize information in both short and long texts. While logistic regression and

Naive Bayes yield average results of 74% and 91%, respectively, for short text evaluation, both models exhibit poor performance when testing on lengthy text passages [17]. The contemporary reliance on social media for news dissemination is evident, with individuals utilizing platforms to express their opinions and thoughts regarding this unusual infection [11].

Sentiment analysis, an efficient method for text analysis, autonomously extracts insights from unstructured data originating from sources such as social media, emails, and support requests. Machine learning (ML) techniques, including various data types, play a pivotal role in automating information harvesting [6],[2]. Jain et al. [6], in their exploration of ML systems for Twitter sentiment analysis, scrutinize diverse measures, presenting a detailed sentiment analysis procedure. Employing multinomial Naive Bayes and decision tree models as analytical tools, the decision tree achieves impeccable scores of 100% for accuracy, precision, recall, and F1-score. Researchers worldwide collaborate to compile and disseminate Twitter datasets for COVID-19 [2],[10]. In [17], the authors utilize three distinct Twitter datasets to conduct sentiment analysis on COVID-19-related tweets. After collecting and pre-processing the datasets, and creating TF-IDF vector representations, various ML models are employed to predict attitudes. The evaluation reveals that, compared to alternative methods, the decision tree attains the highest accuracy at 93%. Furthermore, in [18], a set of keywords is employed to mine Twitter for user sentiment. Subsequently, these sentiments are used to train a Naive Bayes classifier (NBC) algorithm for deciphering user attitudes.

Pokharel et al. [16] delineate the responses of Nepalese citizens to the coronavirus epidemic. Employing CORONAVIRUS and COVID-19 as search terms, they meticulously sift through Twitter for relevant mentions. The sentiment analysis encompasses tweets from Nepal within the period of May 21 to May 31, 2020. In demonstrating the correlation between the rise in COVID-19-positive patients and the progression of time, the authors of [19,20] utilize a mediative fuzzy correlation technique.

3. Methodology

This section elucidates the essential understanding necessary for the study, encompassing every resource and method employed. Following the compilation of raw data, a meticulous pre-processing stage was undertaken to eliminate any errors. The sentiment of each document was subsequently evaluated through sentiment analysis. A diverse array of methods was then applied to extract relevant characteristics. Ultimately, machine learning techniques were employed to classify user sentiments. Figure 1 illustrates the comprehensive strategy employed in performing sentiment analysis.

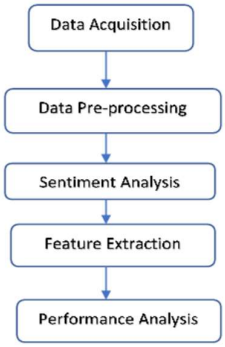

Figure 1: Proposed Methodology

3.1. Data Acquisition

In the scope of this research endeavor, the Covid19 dataset, diligently procured from Kaggle, unfolds as a comprehensive repository of tweet texts intricately linked to the domain of Covid19. This substantial dataset encapsulates a myriad of expressions and conversations circulating around the prevalent global health crisis, recognizing the fluid nature of the discourse. It notably encompasses a diverse array of terms such as "Corona," "Covid19," and "Coronavirus," with the recognition of case insensitivity ensuring a comprehensive coverage of relevant content.

Carefully curated for the explicit purpose of research exploration, this dataset unfolds over a finite yet crucial temporal expanse. Spanning a significant duration of 36 days, from July 25 to August 29, 2020, this temporal window captures a dynamic snapshot of the public's sentiment, concerns, and discussions during a pivotal phase of the ongoing pandemic. The inclusion of this specific timeframe serves to capture the evolving landscape of opinions and emotions within the context of Covid19, offering invaluable insights into the temporal dynamics of public discourse.

As the research progresses, delving into this rich repository promises to unveil nuanced patterns, sentiment shifts, and emergent themes, providing a holistic understanding of the collective voice echoing through the digital realm during this critical period. The meticulous curation of this dataset lays the foundation for a comprehensive analysis that transcends mere quantitative metrics, aiming to uncover the qualitative subtleties embedded within the vast tapestry of online conversations surrounding Covid19.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	user_nam	user_locat	user_descr	user_creat	user_follo	user_frien	user_favo	user_verifi	date	text	hashtags	source	is_retweet
2	á¾oáŽ¥â˜›	astroworlc	wednesda	########	624	950	18775	FALSE	########	If I smelled	the scent of hand sanitize	Twitter for iPhone	FALSE
3	Tom Basile	New York,	Husband, f	########	2253	1677	24	TRUE	########	Hey @Yankees	@YankeesPR and @N	Twitter for Androic	FALSE
4	Time4fistic	Pewee Val	#Christian	########	9275	9525	7254	FALSE	########	@diane34	['COVID19']	Twitter for Androic	FALSE
5	ethel mert	Stuck in th	#Browns #	########	197	987	1488	FALSE	########	@brookba	['COVID19']	Twitter for iPhone	FALSE
6	DIPR-J&K	Jammu an	ðŸ–Šï¸ Offic	########	101009	168	101	FALSE	########	25 July :	['CoronaVirusUpdates', 'C	Twitter for Androic	FALSE
7	ðŸŽ¹ Franz	ÐÐ¾Ð²Ð¾	ðŸŽ¼ #ÐÐ	########	1180	1071	1287	FALSE	########	#coronavi	['coronavirus', 'covid19']	Twitter Web App	FALSE
8	hr bartend	Gainesville	Workplace	########	79956	54810	3801	FALSE	########	How #COV	['COVID19', 'Recruiting']	Buffer	FALSE
9	Derbyshire	LPC		########	608	355	95	FALSE	########	You now have to wear face covering		TweetDeck	FALSE
10	Prathamesh	Bendre	A poet, rei	########	25	29	18	FALSE	########	Praying	['covid19', 'covidPositive']	Twitter for Androic	FALSE
11	Member o	ðŸ‘‡ðŸ»lo	Just as the	########	55201	34239	29802	FALSE	########	POPE AS	['HurricaneHanna', 'COVI	Twitter for iPhone	FALSE
12	Voice Of CBSE Students			########	8	10	7	FALSE	########	49K+		Twitter Web App	FALSE
13	Creativegr	Dhaka,Ban	I'm	########	241	1694	8443	FALSE	########	Order	['logo', 'graphicdesigner', '	Twitter Web App	FALSE
14	SEXXYLYPF	Hotel livin	My ink	########	0	8	32	FALSE	########	ðŸ‘‡ðŸ»@	['COVID19']	Twitter Web App	FALSE
15	Africa You	Africa	Official ac	########	830	254	3692	FALSE	########	Let's all	['COVID19']	Twitter Web App	FALSE

Figure 2: Snapshot of Dataset

3.2.Data Processing

Fundamentally, data pre-processing serves as a pivotal step in purifying raw data to ensure heightened accuracy in subsequent analyses. The distinctive traits of the linguistic model of Twitter present unique challenges. Raw tweets commonly harbor significant noise, misspelled words, and a plethora of acronyms and slang terms, all of which pose obstacles to the precision of our algorithm. The pre-processing of data is undertaken to enhance accuracy by eliminating noisy characteristics. The subsequent steps are implemented to meticulously prepare the dataset for analysis:

- Initially, all occurrences of symbols such as #, @, !, $, %, &, HTML elements, and integers were systematically removed from the entire dataset. The Python programming language's regular expression module was employed to execute these processes.
- Our acquired dataset encompasses both lowercase and uppercase letters. To ensure consistency, we uniformly convert all letters to lowercase.
- Subsequently, the entire text dataset underwent tokenization, a process involving the breakdown of extensive text into more manageable segments, such as individual words or phrases [24].
- Ultimately, the complete text collection underwent stemming to yield refined Twitter text. Stemming involves determining a term's root shape by removing its affixes [25]. Both tokenization and stemming operations were conducted using the Python NLTK module.

3.3.Sentiment Analysis

Sentiment analysis, a nuanced process for evaluating the emotional nuances within textual content, strives to distinguish whether user expressions convey positive, negative, or neutral sentiments. To facilitate this intricate task, the TextBlob library, renowned for its proficiency in handling these three classification types, was judiciously employed [11].

TextBlob furnishes pivotal polarity (P) and subjectivity (S) values for the purpose of categorization. Positivity is discerned when the polarity value surpasses 0 ($P > 0$), neutrality prevails when the value is precisely 0 ($P = 0$), and negativity is inferred otherwise. Subjectivity, denoted by a floating-point integer within the span of [0.0, 1.0], encapsulates a spectrum where 0.0 represents an elevated degree of objectivity, while 1.0 signifies a profound degree of subjectivity.

Upon the meticulous completion of these multifaceted steps, a nuanced spectrum of emotions is methodically assigned to each tweet. This comprehensive analysis not only provides a binary categorization but also offers a nuanced understanding of the sentiment dynamics encapsulated within the user-generated content.

3.4.Feature Extraction

Enhancing the performance of trained models is achievable through the process of feature extraction, which entails deriving features from input data. In this regard, I have employed Word2Vec feature extraction methods, which transform textual information into numerical vectors.

Word embeddings are generated through a suite of interconnected models known as Word2vec. These models, characterized by a two-layer, shallow neural network design, are trained to replicate word contexts from linguistic data. By utilizing an extensive corpus of text as input, Word2vec constructs a vector space, typically with numerous dimensions, assigning each unique word in the corpus a corresponding vector in the space. The arrangement of word vectors in the vector space ensures proximity if the words share a common context in the corpus. Developed by a team of Google researchers led by Tomas Mikolov, Word2vec has been subject to further exploration and documentation by other scholars. Notably, embedding vectors created using the Word2vec algorithm offer several advantages when compared to older techniques such as latent semantic analysis.

3.4.1 CBOW and skip grams

Word2Vec employs two distinct model architectures, namely Continuous Bag of-Words (CBOW) and continuous skip-gram, to craft a distributed representation of words. In the continuous bag-of-words architecture, the model predicts the current word based on a window of adjacent context words. Notably, the prediction remains unaffected by the order of the context words, adhering to the bag-of-words assumption.

Conversely, the continuous skip-gram architecture utilizes the current word to forecast the context words anticipated to appear within the surrounding window. This architecture assigns greater importance to adjacent context words over more distant ones [1] [4]. The authors observe [5] that CBOW exhibits swifter computational speed compared to skip-gram, which, although slower, proves more efficacious, particularly for infrequent words.

This binary choice between CBOW and skip-gram represents a crucial consideration in leveraging Word2Vec for distributed word representations. The trade-off between computational efficiency and effectiveness in capturing nuances, especially in the case of less frequent words, underscores the need for a judicious selection based on the specific requirements and nuances of the dataset at hand. As we delve into the application of Word2Vec in this study, understanding the nuances of these model architectures becomes imperative for extracting meaningful insights from the textual data within the Covid19 dataset.

3.5.Classifier Models

The examination of user sentiment within online social networks has traditionally leveraged various categorization techniques, with a predominant reliance on machine learning and deep learning classifiers. In the course of this study, seven distinct classification models were deployed, each delineated below.

- Logistic Regression (LR): LR utilizes a simplified logistic equation for analyzing data with a binary dependent variable. Despite its simplicity, numerous intricate variants exist [28]. In regression analysis, logistic regression is employed to ascertain the variables involved in a logistic procedure.

- Support Vector Machine (SVM): SVM, a plane-based classification algorithm, constructs a discrete hyperplane in the descriptive space of training data and

components. It categorizes examples or cases based on their placement relative to this hyperplane. In a linearly separable dataset, SVM identifies the hyperplane that effectively separates the two groups. The primary objective of SVM is to determine the optimal hyperplane in the training data between two datasets. This is achieved by solving an optimization problem using the following equation [29]:

$$\max Q(\alpha) = \sum_{i=1}^{n} \propto_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \propto_i \propto_j d_i d_j X_i^T X_j$$

- k-Nearest Neighbour (k-NN): The k-NN approach stands out as one of the most straightforward machine learning algorithms, grounded in the principles of supervised learning. Operating on the premise of retaining all existing data, it categorizes additional data points by identifying commonalities. This unique methodology ensures that newly acquired data can be seamlessly classified through the k-NN approach [30].
- Naïve Bayes: The Naïve Bayes method, a classification approach rooted in the Bayes theorem, operates as a probabilistic classifier. Its predictions are contingent on the likelihood of an event occurring [31]. Employing the following equation [32], it calculates the probability that an observation, X, belongs to the class Yk. For instance, if X represents a vector of word occurrences or word counts, the probability is determined accordingly.

$$P(Y_k/X) = \frac{P(Y_k)P(\frac{X}{Y_k})}{P(X)}$$

- Decision Tree (DT): A decision tree (DT) manifests as a tree structure reminiscent of a flowchart, delineating its core nodes through rectangles and leaf nodes through ovals [34]. Falling under the realm of supervised learning, the Decision Tree approach embodies methods designed for discerning patterns and making decisions based on training data.

- Random Forest (RF): Renowned for its prowess in addressing complex problems and elevating model performance, Random Forest (RF) stands as a prominent supervised learning technique employing an ensemble learning approach [35]. Operating as an ensemble classifier, RF strategically utilizes multiple decision trees derived from subsets of the training data, incorporating parameters selected at random .

- Extreme Gradient Boosting (XGBoost): A contemporary powerhouse in applied machine learning, Extreme Gradient Boosting (XGBoost) has emerged as a dominant force [36]. Functioning as an implementation of gradient-boosted decision trees, XGBoost is tailored for unparalleled speed and effectiveness. Optimal performance from XGBoost models demands a wealth of information and meticulous model refinement, setting it apart from approaches like random forest[36].

3.6.Evaluation Metrics

Performance evaluation in machine learning encompasses four key metrics: accuracy, indicating overall correctness; precision, revealing the precision of positive predictions; recall, assessing the model's ability to capture all positive instances; and F1-score, a balanced measure

considering precision and recall. These metrics collectively offer a nuanced understanding of a model's effectiveness, enabling comprehensive assessment and informed decision-making across diverse applications.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = 2* \frac{\text{Pr}\,ecision * \text{Re}\,call}{\text{Pr}\,ecision + \text{Re}\,call}$$

4. Results

This study aims to scrutinize the classification performance of a substantial Covid19 dataset, comprising 179,109 records sourced from Kaggle. Employing machine learning techniques such as Logistic Regression, k-NN, Naive Bayes, Decision Tree, Random Forest, and XGBoost, the research assesses performance through diverse evaluation metrics. Leveraging the Natural Language Toolkit (NLTK) package in Python for text processing and classification, this study utilizes NLTK's capabilities, including tokenization, tagging, and text manipulation. Integration of word2vec facilitates the transformation of tweets into numerical vectors, streamlining the construction of machine learning models. Post pre-processing, sentiment analysis of the Twitter dataset is conducted, categorizing user tweets into positive, neutral, or negative sentiments. The examination reveals that the majority of individuals express neutral sentiments toward COVID-19, providing valuable insights into public perceptions.

	text	scores	compound
0	smelled scent hand sanitizers today someone pa...	{'neg': 0.0, 'neu': 0.738, 'pos': 0.262, 'comp...	0.4939
1	hey yankees yankeespr mlb would made sense pla...	{'neg': 0.12, 'neu': 0.684, 'pos': 0.197, 'com...	0.2263
2	wdunlap realdonaldtrump trump never claimed ho...	{'neg': 0.0, 'neu': 0.794, 'pos': 0.206, 'comp...	0.2057
3	brookbanktv one gift give appreciation simple ...	{'neg': 0.0, 'neu': 0.53, 'pos': 0.47, 'compou...	0.7351
4	july media bulletin novel coronavirusupdates d...	{'neg': 0.0, 'neu': 0.753, 'pos': 0.247, 'comp...	0.3182

Figure 3: Sentiment Intensity Analyzer

Following the extraction of features, the dataset underwent a train-test-split, dividing it into two distinct sets. Opting for a 70:30 ratio, where 70% serves as the training dataset and 30% as the testing dataset, this step was crucial for model training and evaluation. Seven machine learning algorithms—logistic regression, support vector machine (SVM), decision tree, random forest, Naive Bayes, k-nearest neighbors (k-NN), and XGBoost—were employed for model training. The accuracy of each method on the test dataset was meticulously assessed. Comprehensive performance verification, including precision, recall, and F1-score values, is presented in Table 1.

Table 1: Measurement of various Machine Learching Approaches

Feature Extraction	**Algorithms**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Word2vec	Logistic Regression	89.12%	91.24%	82.81%	85.18%
	SVM	91.51%	92.90%	86.45%	89.45%
	Naïve Bayes	88.91%	90.90%	82.23%	85.13%
	k-NN	90.23%	92.15%	84.45%	86.03%
	Decision Tree	93.85%	94.37%	91.74%	92.45%
	Random Forest	95.51%	96.14%	93.32%	94.67%
	XGBoost	81.20%	85.23%	76.12%	79.25%
Feature Extraction	**Algorithms**	**Accuracy**	**Precision**	**Recall**	**F1-score**
TF-IDF	Logistic Regression	85.34%	90.14%	81.45%	83.41%
	SVM	90.51%	92.15%	85.20%	88.12%
	Naïve Bayes	87.99%	91.90%	81.20%	85.01%
	k-NN	88.20%	91.50%	83.41%	85.15%
	Decision Tree	92.58%	93.47%	90.58%	91.34%
	Random Forest	94.14%	95.89%	92.20%	94.10%
	XGBoost	80.25%	84.74%	75.45%	78.14%

Table 1 presents a comprehensive assessment of the performance of various machine learning algorithms under two distinct feature extraction methods: Word2Vec and TF-IDF. In the Word2Vec feature extraction category, the algorithms demonstrated noteworthy results. Logistic Regression achieved an accuracy of 89.12%, showcasing balanced precision (91.24%) and F1-score (85.18%). SVM displayed high accuracy (91.51%) and precision (92.90%), along with substantial recall (86.45%) and F1-score (89.45%). Random Forest outperformed others with an outstanding accuracy of 95.51%, high precision (96.14%), and superior F1-score (94.67%). However, XGBoost exhibited moderate performance with an accuracy of 81.20% and balanced precision (85.23%) and F1-score (79.25%). In the TF-IDF feature extraction category, similar trends were observed. Random Forest showcased exceptional performance with an accuracy of 94.14%, high precision (95.89%), and a superior F1-score of 94.10%. These results provide valuable insights into the strengths and weaknesses of each algorithm, aiding in informed decision-making for the selection of the most suitable approach based on specific requirements and priorities.

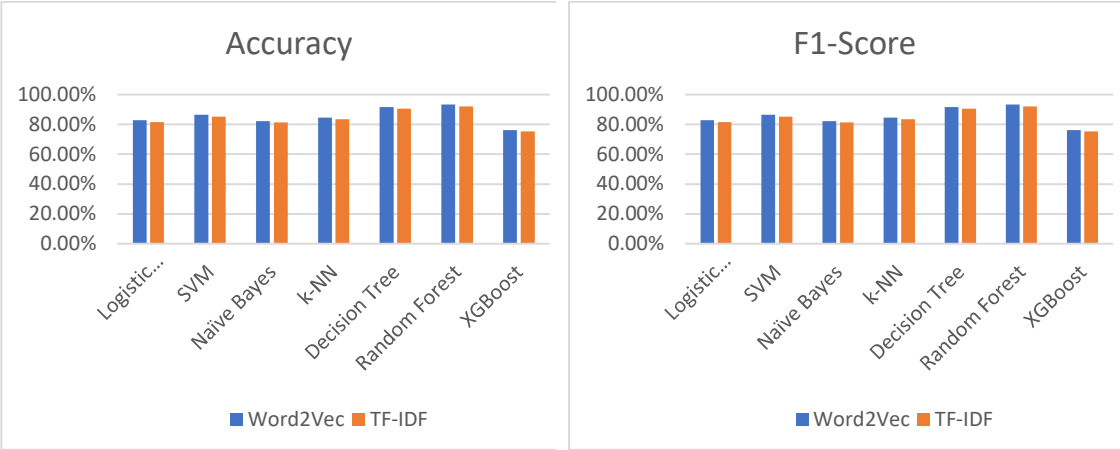

Figure 4: Accuracy and F1-Score with the Twitter Dataset

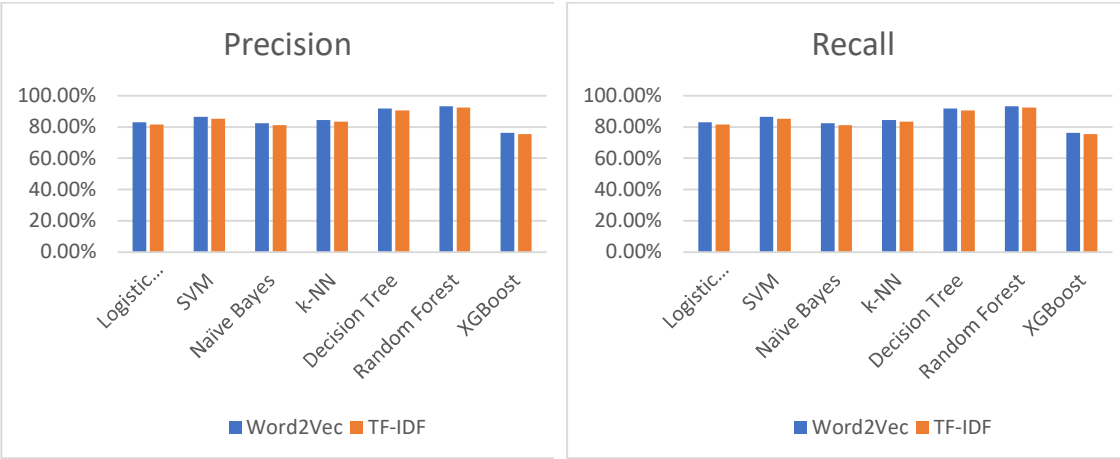

Figure 4: Precision and Recall with the Twitter Dataset

5. Conclusions

This study aims to delve into the nuanced realm of user sentiment by leveraging machine learning models for the accurate prediction of sentiments from a collection of COVID-19-related tweets acquired between July 25 and August 29, 2020. The employed feature extraction methods, TF-IDF and Word2Vec, exhibit commendable performance in categorizing user sentiments within the constructed machine learning models. Particularly, the Random Forest algorithm, in conjunction with both Word2Vec and TF-IDF, demonstrates outstanding efficacy. Notably, the Random Forest classifier paired with the Word2Vec feature extraction method consistently yields the most reliable and robust results. Across various machine learning algorithms, Word2Vec outshines TF-IDF, including Logistic Regression, SVM, Naive Bayes, k-NN, Decision Tree, and XGBoost.

Looking ahead, the future scope of this research involves extending the analysis to encompass diverse social networking platforms such as Facebook, Instagram, and LinkedIn. This expansion is anticipated to contribute to the development of a potent model capable of more precisely categorizing user sentiments across various online platforms.

6. References

1. Abbasi, A., Javed, A. R., Chakraborty, C., Nebhen, J., Zehra, W., & Jalil, Z. E.,"An Ensemble Learning Approach for Concept Drift Detection in Dynamic Social Big Data Stream Learning," IEEE Access, vol. 9, pp. 66408-66419, 2021.
2. E. Chen, K. Lerman, and E. Ferrara, "Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Dataset," JMIR Public Health and Surveillance, vol. 6, no. 2, p. e19273, 2020.
3. M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, and A. L. Schmidt, "The COVID-19 Social Media Infodemic," Scientific Reports, vol. 10, pp. 1–10, 2020.
4. N. Fernandes, "Economic Effects of Coronavirus Outbreak (COVID-19) on the World Economy," Available at SSRN 3557504, 2020.
5. S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, "An Ensemble Machine Learning Approach through Effective Feature Extraction to Classify Fake News," Future Generation Computer Systems, vol. 117, pp. 47–58, 2021.
6. A. Jain and P. Dandannavar, "Application of Machine Learning Techniques to Sentiment Analysis," in Proceedings of the 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, India, July 21–23, pp. 628–63, 2016.
7. H. Jelodar, Y. Wang, R. Orji, and S. Huang, "Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or Covid-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 10, pp. 2733–2742, 2020. [Online]. Available: https://doi.org/10.1109/JBHI.2020.3001216
8. A. Kumar, P. K. Roy, and J. P. Singh, "Working Notes of FIRE - 13th Forum for Information Retrieval Evaluation," Fire-WN, vol. 3159, pp. 1216–1220, 2021.
9. A. Kumar, G. S. Shankar, S. Gautham, P. K. Reddy, and G. T. Reddy, "A Two-Stage Text Feature Selection Algorithm for Improving Text Classification," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 20, pp. 1–19, 2021.
10. M. Kabir and M. S. CoronaVis, "A Real-Time COVID-19 Tweets Analyzer," arXiv, 2020, arXiv:2004.13932.
11. S. Loria, "TextBlob: Simplified Text Processing Release ver. 0.15.2," Available online. https://textblob.readthedocs.org/en/dev/index.html.
12. A. Mittal and S. Patidar, "Sentiment Analysis on Twitter Data: A Survey," in Proceedings of the 2019 7th International Conference on Computer and Communications Management, Bangkok, pp. 91–95, 2019.
13. A. Mondal, S. Mahata, M. Dey, and D. Das, "Classification of COVID19 Tweets Using Machine Learning Approaches," in Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task, Mexico City, pp. 135–137, 2021.
14. U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis," IEEE

Transactions on Computational Social Systems, vol. 8, no. 4, pp. 1003–1015, 2021. [Online]. Available: https://doi.org/10.1109/TCSS.2021.3051189

15. A. S. Neogi, K. A. Garg, R. K. Mishra, and Y. K. Dwivedi, "Sentiment Analysis and Classification of Indian Farmers' Protest Using Twitter Data," International Journal of Information Management Data Insights, vol. 1, p. 100019, 2021.
16. A. L. Pedrosa, L. Bitencourt, A. C. F. Fróes, M. L. B. Cazumbá, R. G. B. Campos, S. B. C. S. de Brito, and A. C. Simões E Silva, "Emotional Behavioral and Psychological Impact of the COVID-19 Pandemic," Frontiers in Psychology, vol. 11, p. 566212, 2020. [Online]. Available: https://doi.org/10.3389/fpsyg.2020.566212
17. B. Pokharel, "Twitter Sentiment Analysis During COVID-19 Outbreak in Nepal," Available at SSRN 3624719, 2020.
18. J. Samuel, G. Ali, M. Rahman, E. Esawi, and Y. Samuel, "Covid-19 Public Sentiment Insights and Machine Learning for Tweets Classification," Information Retrieval, vol. 11, p. 314, 2020.
19. M. Sethi, S. Pandey, P. Trar, and P. Soni, "Sentiment Identification in COVID-19 Specific Tweets," in Proceedings of the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, July 2–4, pp. 509–516, 2020. https://doi.org/10.1109/ICESC48915.2020.9155674.
20. R. B. Shamantha, S. M. Shetty, and P. Rai, "Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance," in Proceedings of the 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, pp. 21–25, February 23–25, 2019. https://doi.org/10.1109/CCOMS.2019.8821650
21. M. K. Sharma, N. V. Dhiman, V. N. Vandana, and V. N. Mishra, "Mediative Fuzzy Logic Mathematical Model: A Contradictory Management Prediction in COVID-19 Pandemic," Applied Soft Computing, vol. 105, p. 107285, 2021. doi: 10.1016/j.asoc.2021.107285.
22. M. K. Sharma, N. V. Dhiman, V. N. Vandana, and V. N. Mishra, "Mediative Fuzzy Logic Mathematical Model: A Contradictory Management Prediction in COVID-19 Pandemic," Applied Soft Computing, vol. 105, p. 107285, 2021. https://doi.org/10.1016/j.asoc.2021.107285
23. C. Shofiya and S. Abidi, "Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data," International Journal of Environmental Research and Public Health, vol. 18, no. 11, p. 5993, 2021. https://doi.org/10.3390/ijerph18115993
24. M. Straka and J. Straková, "Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with Udpipe," Association for Computational Linguistics, pp. 88–99, 2017. https://doi.org/10.18653/v1/K17-3009.
25. J. Lovins, "Development of a Stemming Algorithm," Mech. Transl. Computational Linguistics, vol. 11, pp. 22–31, 1968.
26. G. Stringhini, C. Kruegel, and G. Vigna, "Detecting Spammers on Social Networks," in Proceedings of the 26th Annual Computer Security Applications Conference, Austin, TX, United States, December 6–10, 2010, pp. 1–9. https://doi.org/10.1145/1920261.1920263
27. S. Tuli, S. Tuli, R. Tuli, and S. S. Gill, "Predicting the Growth and Trend of COVID-19 Pandemic Using Machine Learning and Cloud Computing," Internet of Things, vol. 11, p. 100222, 2020. https://doi.org/10.1016/j.iot.2020.100222

28. J. C. Stoltzfus, "Logistic Regression: A Brief Primer," Academic Emergency Medicine, vol. 18, no. 10, pp. 1099–1104, 2011. https://doi.org/10.1111/j.1553-2712.2011.01185.x
29. W. S. Noble, "What Is a Support Vector Machine?" Nature Biotechnology, vol. 24, no. 12, pp. 1565–1567, 2006. https://doi.org/10.1038/nbt1206-1565
30. S. Tan, "An Effective Refinement Strategy for KNN Text Classifier," Expert Systems with Applications, vol. 30, no. 2, pp. 290–298, 2006. https://doi.org/10.1016/j.eswa.2005.07.019
31. I. Rish, "An Empirical Study of the Naive Bayes Classifier," in IJCAI 2001 Workshop Empirical Methods Artificial Intelligence, pp. 41–46, 2001.
32. W. Dai, G. Xue, Q. Yang, and Y. Yu, "Transferring Naive Bayes Classifiers for Text Classification," AAAI, vol. 7, pp. 540–545, 2007.
33. A. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial Naive Bayes for Text Categorization Revisited," in Proceedings of 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, pp. 488–499, December 4–6, 2004. [Online]. Available: https://doi.org/10.1007/978-3-540-30549-1_40.
34. A. Priyam, A. Abhijeeta, R. Rathee, and S. Srivastava, "Comparative Analysis of Decision Tree Classification Algorithms," International Journal of Current Engineering and Technology, vol. 3, pp. 334-337, 2013.
35. B. Xu, X. Guo, Y. Ye, and J. Cheng, "An Improved Random Forest Classifier for Text Categorization," Journal of Computers, vol. 7, no. 12, pp. 2913–2920, 2012, https://doi.org/10.4304/jcp.7.12.2913-2920.
36. Z. Chen, F. Jiang, Y. Cheng, X. Gu, W. Liu, and J. Peng, "XGBoost Classifier for DDoS Attack Detection and Analysis in SDN-Based Cloud," in Proceedings of the IEEE International Conference On Big Data And Smart Computing (BigComp), Shanghai, China, pp. 251–256, 15–17 January 2018.