Manuscript Submitted	19.10.2023
Accepted	19.11.2023
Published	31.12.2023

Aspect Based Sentiment Analysis: Feature Extraction using Latent Dirichlet Allocation (LDA) and Term Frequency - Inverse Document Frequency (TF-IDF) in Machine Learning (ML)

Shakirah Mohd Sofi[1, 2]

[1] Jabatan Komputeran, Fakulti Multimedia Kreatif & Komputeran, Universiti Islam Selangor (UIS), 43000 Bandar Seri Putra, Kajang, Selangor

[2] Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra, Kuala Lumpur 54100, Malaysia

syakirah@kuis.edu.my

Ali Selamat[1, 2, 3]

[1] Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra, Kuala Lumpur 54100, Malaysia

[2] School of Computing, Faculty of Engineering, & Media and Games Center of Excellence (MagicX), Universiti Teknologi Malaysia, Skudai 81310, Johor Bahru, Malaysia

[3] Center for Basic and Applied Research, Faculty of Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, 50003 Hradec Kralove, Czech Republic

aselamat@utm.my

Abstract

The growth and development of social networks, blogs, forums, and e-commerce websites has produced a number of data, notably textual data, which has increased tremendously. Twitter is one of the most popular media social platforms; during the COVID-19 pandemic, people all around the world use social media to share their opinions or concerns about the pandemic that has changed their lives. It revealed a significant rise in tweets on coronavirus, including positive, negative, and neutral tweets about the virus's impact. Sentiment analysis faces challenges: sparse data limits understanding, while topic coherence and interpretability demand improvement for clearer insights. The primary goal of this paper is to improve the accuracy and effectiveness of sentiment analysis during the COVID-19 pandemic through the application of advanced techniques and classifiers. In this article, we experiment with such Support Vector Machines (SVM) and Naive Bayes (NB) on Twitter data for high-accuracy machine learning models. Using Latent Dirichlet Allocation (LDA)for feature extraction, we aim to capture comprehensive aspects and topics for sentiment analysis. Additionally, we explore Count Vectorizer and Term Frequency - Inverse Document Frequency (TF-IDF) as word embedding techniques. The main objectives are to extract topics, understand public concerns about Covid-19, and compare classifier performance in Aspect-Based Sentiment Analysis on Covid-19 tweets. This paper introduces advanced sentiment analysis techniques, such as LDA, Count Vectorizer, and SVM, enhancing nuanced sentiment analysis during the COVID-19 pandemic with notable 85% accuracy in SVM classification.

Keywords: *Aspect-Based Sentiment Analysis, Opinion Mining, Feature Extraction, Top Modeling, LDA, Count Vectorizer, TF-IDF, SVM, NB.*

1. Introduction

Nowadays, text data is disseminated through social media platforms such as Twitter, Facebook, and Instagram, as well as corporate platforms and e-commerce websites that the data is written in a brief or short and imprecise manner. Due to this situation, sentiment analysis of this data is challenging (Cambria et al., 2017), and a process that transforms the statement into something meaningful is required. Sentiment analysis can be categorized into three levels: document, sentence and entity or aspects (Hu & Liu, 2004). A focus on the document or sentence level makes the assumption that the topic is the only thing that is covered in that document or sentence, which is usually the case. For a more thorough study, it is consequently important to look into both entities and aspects. Aspect Based Sentiment analysis (ABSA) should be employed for a more detailed assessment. In ABSA, product characteristics, services, topics, issues, persons or events that prompt user opinions extraction and detection are also obtained in addition to sentiment polarity (Pontiki et al., 2016).

Text data extraction from sentiment analysis requires the precise use of Natural Language Processing (NLP) algorithms since text data is different from numerical data, visual data, or signal data. Regardless matter whether English, Arabic, Malay, Chinese, or Turkish is used for user opinions and expression that is written, sentiment analysis is utilized to discern the meaning of that sentence, whether it is a positive or negative. In this study, we worked with social media sites such as Twitter, which has seen an exponential increase in tweets on pandemics in short periods of time and written in English.

On March 11, 2020, the World Health Organization (WHO) declared COVID-19, also known as Novel Coronavirus 2019, a pandemic (World Health Organization, 2021). A pandemic is a widespread outbreak of a disease that extends well beyond national boundaries, affecting people globally. When the pandemic struck, a large portion of the population found themselves were confined to their homes, unable to engage in work activities, and restricted from going outside. Social media platforms such as Twitter were utilized extensively, irrespective of the hour or the age of the users. These platforms served as a means for the public to vent their sentiments, share their opinions, express their emotions, and convey their responses to the global pandemic that had enveloped the world.

In the case of the COVID-19 pandemic, extensive misinformation circulates on social media. False information related to the virus's transmission, prevention, vaccinations, patient isolation, mortality rates, and other aspects is frequently disseminated. The spread of inaccurate information significantly impacts countries, governments, international relations, medical teams, and the families of the affected individuals (Apuke & Omar, 2021). It is imperative to adopt a meticulous approach to ensure the acquisition of precise information. Employing analytical tools such as sentiment analysis is essential for gathering user opinions and evaluating them with precision.

In this article, we are conducting experiments using Liner Regression (LR), Support Vector Machines (SVM) and Naive Bayes (NB) classifiers on Twitter data to train machine learning models, aiming to achieve high accuracy rates. We employ Latent Dirichlet Allocation (LDA) as a top modeling approach for feature extraction, enabling us to capture all aspects and topics related to sentiment analysis. Additionally, we explore two widely used word embedding techniques: Count Vectorizer and Term Frequency - Inverse Document Frequency (TF-IDF). The primary objectives of this work are as follows:

i. To indicate the extraction of topics and present the prevailing discourse of public concern regarding Covid-19.

ii. To compare the performance of machine learning classifiers when coupled with word embedding techniques in the context of Aspect-Based Sentiment Analysis on Covid-19 tweets.

In conclusion, Sentiment Analysis (SA) is pivotal for discerning the sentiment in text, and its efficacy depends on a symbiotic relationship between feature extraction techniques, such as Latent Dirichlet Allocation (LDA) and TF-IDF, and machine learning classifiers like SVM and Naive Bayes. The selection of relevant features significantly influences the model's ability to accurately predict sentiment. LDA aids in understanding context and nuances, while SVM handles complex relationships.

The success of sentiment analysis hinges on the harmonious integration of feature extraction and machine learning classifiers, ensuring accurate and nuanced predictions in diverse textual contexts.

2. Related Work

Ever since the coronavirus outbreak in 2019, numerous researchers have been dedicated to investigating the virus's origins, transmission mechanisms, preventive measures, and its effects on the general population. This section will examine the utilization of NLP, with a particular focus on sentiment analysis, employing a variety of machine learning techniques. Several studies have been conducted regarding sentiment analysis related to pandemic outbreaks, utilizing social media platforms as a means for individuals to express their opinions, emotions, and a variety of viewpoints. These studies encompass a wide range of topics, including miss leading information (Priya & Kumar, 2021), online learning (Rapanta et al., 2020), vaccine sentiments (Yousefinaghani et al., 2021), the impact of infections on patients (Sayed et al., 2021), and much more. Chakraborty conducted a study that examined how individuals' behaviors while disseminating information on social media platforms influence the accuracy of the information being shared (Chakraborty et al., 2020). This research was conducted by collecting tweet data in English from 10 different countries. Its aim was to gain insights into how people from various countries affected by COVID-19 are coping with the situation (Kausar et al., 2021). (Naseem et al., 2021) conducted research utilizing the COVID Senti dataset, employing experimental approaches with word embeddings, machine learning, and deep learning classifications to achieve optimal results.

In this research, Latent Dirichlet Allocation (LDA) is employed to extract the relevance of topics during the COVID-19 pandemic. Two prominent topics that are explored include the origin of COVID-19, specifically concerning the virus's origins in China, and the outbreak of COVID-19 (Abd-Alrazaq et al., 2020). (Raza et al., 2021) conducted a study and found that Support Vector Machine (SVM) performed the best, achieving a 93% accuracy rate through a comparative analysis of our word embedding techniques, which included Count Vectorizer and TF-IDF. The future engineering techniques proposed for sentiment analysis of COVID-19 tweets have delivered noteworthy results. These results were obtained through a comparison of supervised machine learning with various feature extraction techniques, which encompass TF-IDF, Bag of Words (BoW), and concatenation (Rustam et al., 2021). (Avasthi et al., 2022) analyzes Twitter data to understand global and Indian sentiments during the COVID-19 crisis, employing sentiment analysis and LDA topic modeling on CORD-19 data to reveal insights into people's emotions, with potential implications for research and healthcare. By repeating the lemmatization process three times for optimal results, this research establishes it as the most effective approach, and in addition, uses the LDA model to identify and discuss eight crucial topics related to Coronavirus (Abdulaziz et al., 2021).

3. Methodology

Figure 1 illustrates a proposed experimental setup framework for this empirical study. The suggested framework comprises five stages: 1) data preparation, 2) data pre-processing, 3) feature extraction selection, 4) text classification methods, and 5) evaluation.

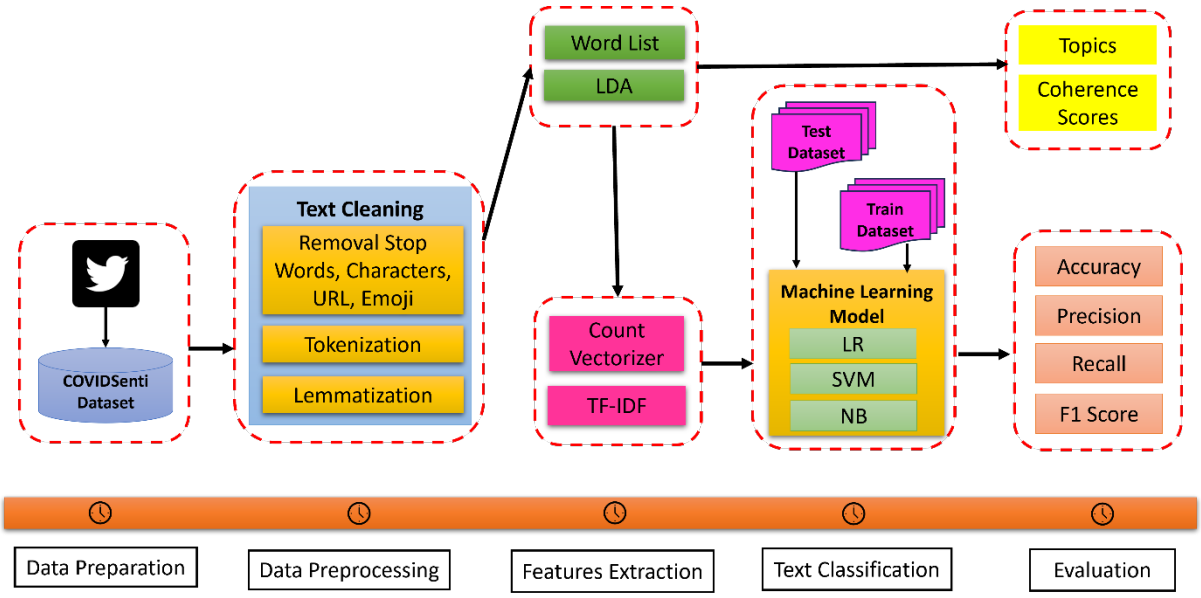

Figure 1 Overview of the proposed approach.

3.1 COVIDSenti Dataset and Data Collection

i. Dataset Preparation

In the initial stage, a publicly available dataset was obtained from a GitHub repository known as COVIDSENTI. This dataset consists of 2.1 million sentiment-related tweets spanning two months, from February 2020 to March 2020 (Naseem et al., 2021). For our study, we focused exclusively on tweets in the English language. The dataset was filtered using keywords to ensure that it pertained to COVID-19 and related issues. Subsequently, COVIDSENTI was partitioned into three equal subsets: COVIDSENTI-A for positive attitudes, COVIDSENTI-B for negative attitudes, and COVIDSENTI-C for neutral attitudes. An overview of the dataset distribution is presented in Table 1.

Table 1 Overview of the dataset distribution.

Dataset\ Label	**Positive**	**Negative**	**Neutral**	**Total**
COVIDSenti-A	1,968	5,088	22,949	30,000
COVIDSenti-B	2,033	5,471	22,496	30,000
COVIDSenti-C	2,279	5,781	21,940	30,000
COVIDSenti	6,280	16,335	67,835	90,000

In this research, Google Colaboratory and Python code were used to develop multiple machine learning models. While the development occurred on Google Colaboratory, the three machine learning algorithms were trained and executed on a local system.

ii. Data Pre-Processing

Information gathered from social media platforms is often noisy, unstructured, informal, and diverse. The initial stage of sentiment analysis involves data pre-processing, which aims to enhance the textual content's meaningfulness. This is achieved through the following sequential strategies:

a) *Removal Patterns:* The removal of stop words and unnecessary characters is a common method to reduce noise in textual data. Eliminating stop words and extraneous characters does not affect the interpretation of the sentiment aspect of a phrase. The program first converts uppercase letters to lowercase and then removes all special characters, emojis, hyperlinks, hashtags (e.g., #StaySafe, #COVID-19), stop words (such as "for," "the," and "is"), and URLs from the dataset tweets.

b) *Tokenization:* The second step is tokenization, which is the process of converting text into tokens before transforming it into vectors. Tokenization entails breaking raw text into words and phrases referred to as tokens. Tokenization aids in deciphering the text's meaning by analyzing the sequence of words.

c) *Lemmatization:* The final step involves lemmatization, which utilizes vocabulary and morphological analysis of sentences to return words to their root form. We employed lemmatization, which transforms terms to their base form (e.g., "killing" to "kill" or "stays" to "stay").

iii. Features Extraction

a) *Keywords or Words List Analysis*: After the data pre-processing, we conducted a keyword or word list analysis on our preprocessed corpus to identify the most frequently mentioned words. This analysis revealed that people consistently mention five keywords in relation to coronavirus cases, including the country of origin of the virus, the daily count of new cases, and the total number of deaths.

b) *Top Modeling with Latent Dirichlet Allocation (LDA*): LDA, or Latent Dirichlet Allocation, serves as a generative probabilistic model utilized to uncover the latent topics within a corpus of documents or text (Blei et al., 2003). It operates under the assumption that each document comprises a blend of various topics, and each individual word in a document is associated with one of these topics. In the realm of text analysis, LDA finds its utility in the automated categorization of documents into topics and the comprehension of the predominant themes embedded within a collection of documents.

In LDA, we typically specify the number of topics in advance, it was set to 10 topics. LDA learns these topics as distributions of words, and it also determines the topic distributions of the documents after training on the given dataset.

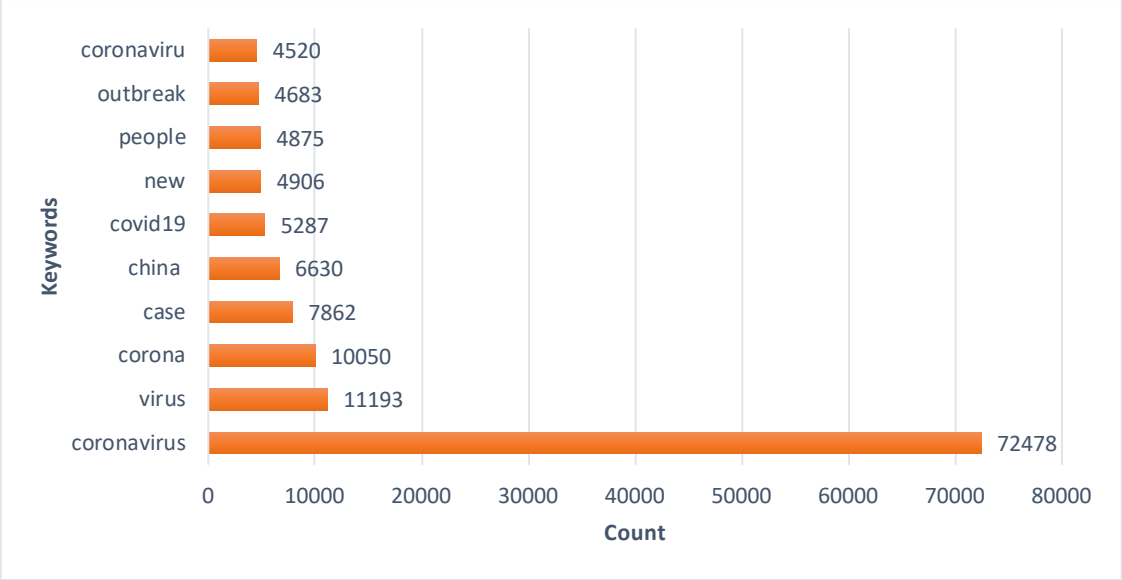

Figure 2 Top 10 frequently keywords mention in COVIDSenti dataset.

Table 2 Top 10 words in top 10 topics.

Topic 1:	coronavirus	update	fear	live	market	stock	iran	city	york	minister
Topic 2:	coronavirus	trump	say	fight	south	medium	bring	house	canaot	class
Topic 3:	coronavirus	covid	news	know	italy	need	panic	christ	mean	here
Topic 4:	coronavirus	china	outbreak	spread	death	health	world	cancel	concern	close
Topic 5:	coronavirus	virus	corona	people	trump	go	like	realdonaldtrump	kill	to
Topic 6:	coronavirus	case	covid	new	confirm	health	test	report	patient	positive
Topic 7:	coronavirus	covid	flu	youtube	pandemic	disease	risk	expert	rate	testing
Topic 8:	coronavirus	outbreak	quarantine	cruise	mask	ship	school	face	flight	ask
Topic 9:	coronavirus	people	coronaviru	like	get	think	work	iaom	hand	home
Topic 10:	coronavirus	covid	outbreak	spread	say	health	world	global	official	youtube

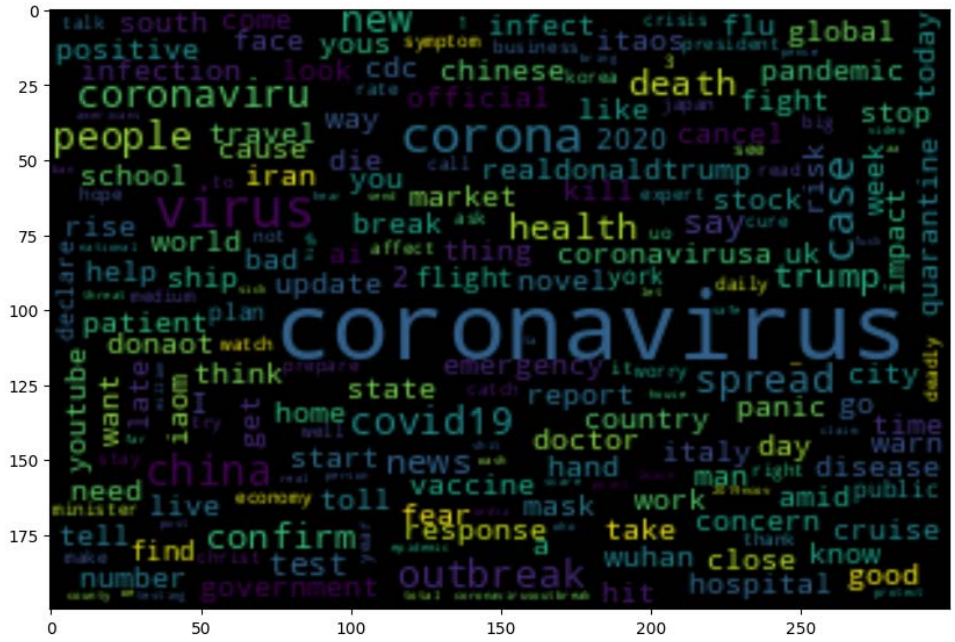

Figure 3 Word cloud in COVIDSenti dataset.

c) *Count Vectorizer and TF-IDF:* The experiments involved the utilization of vectorization methods and word embedding techniques, such as Continuous Bag of Words (CBOW), for feature extraction. Vectorization was accomplished using the term frequency-inverse document frequency (TF-IDF) approach. TF-IDF is a method for assessing the significance of a word within a document or corpus. The importance of a word increases with its frequency in the document but is balanced by the word's frequency in the entire corpus.

iv. *Text Classification*

To offer a comprehensive analysis, we employed machine learning classifiers to assess performance in the sentiment classification task. In our analysis, we utilized machine learning-based classifiers, including Support Vector Machine (SVM), Naive Bayes (NB), and Linear Regression (LR).

The next step in the process involves determining accuracy values. We employed a range of C values, specifically 0.1, 1, 10, and 100. The choice of C values plays a crucial role in the classification process. A smaller margin is preferred, which leads to higher accuracy in the classification results.

v. *Evaluation*

In the evaluation phase, the LDA model produces topic coherence values and scores, which can be visualized and analyzed. The results of the topic coherence assessment provide coherence values that are valuable for evaluating the quality of the Topic Modeling. A higher coherence value indicates a better model. To facilitate this evaluation, a coherence score for each topic is presented in Table 3, and a graphical representation is shown in Figure 4.

Table 3 Coherence Score for 20 Topics

Num Topics	Coherence Score	Num Topics	Coherence Score
1	0.1147	11	0.2895
2	0.1547	12	0.2784
3	0.277	13	0.2865
4	0.2744	14	0.2749
5	0.3059	15	0.252
6	**0.317**	16	0.279
7	0.3011	17	0.2701
8	0.3071	18	0.2849
9	0.2949	19	0.2728
10	**0.3158**	20	0.2699

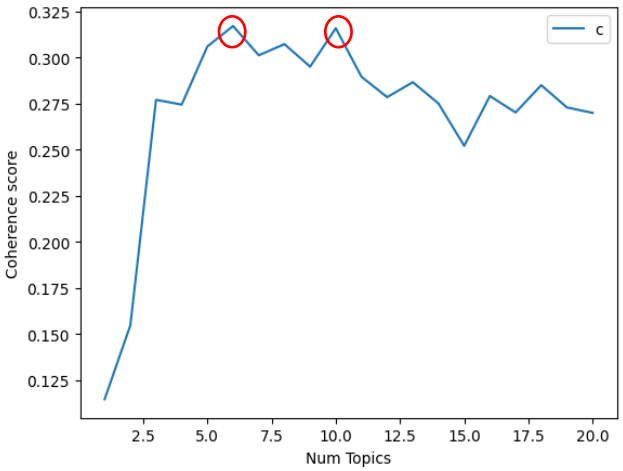

Figure 4: Visualization of Coherence Score for 20 Topics

Table 4 The results of LDA Model on 10 Topics with the highest Coherence Score

Num Topics	LDA Model
0	0.119*"coronavirus" + 0.029*"go" + 0.022*"hand" + 0.015*"declare" + 0.014*"people" + 0.013*"covid" + 0.012*"get" + 0.012*"to" + 0.011*"stop" + 0.011*"wash"'
1	0.091*"coronavirus" + 0.041*"people" + 0.030*"die" + 0.020*"coronaviru" + 0.020*"man" + 0.019*"infect" + 0.016*"day" + 0.015*"kill" + 0.010*"away" + 0.009*"life"'
2	0.121*"coronavirus" + 0.068*"covid" + 0.028*"health" + 0.026*"concern" + 0.025*"response" + 0.017*"trump" + 0.012*"public" + 0.012*"county" + 0.009*"worker" + 0.008*"who"'
3	0.119*"test" + 0.060*"positive" + 0.052*"coronavirus" + 0.028*"big" + 0.020*"covid" + 0.019*"south" + 0.017*"coronavirusoutbreak" + 0.016*"deal" + 0.015*"question" + 0.014*"cough"'
4	0.113*"coronavirus" + 0.020*"late" + 0.017*"covid" + 0.016*"news" + 0.015*"cause" + 0.014*"class" + 0.014*"thank" + 0.013*"iran" + 0.013*"youtube" + 0.011*"business"'
5	0.091*"coronavirus" + 0.020*"start" + 0.019*"take" + 0.018*"plan" + 0.018*"testing" + 0.013*"laugh" + 0.013*"mask" + 0.013*"not" + 0.011*"face" + 0.011*"coronaviru"'
6	0.119*"coronavirus" + 0.029*"go" + 0.022*"hand" + 0.015*"declare" + 0.014*"people" + 0.013*"covid" + 0.012*"get" + 0.012*"to" + 0.011*"stop" + 0.011*"wash"'
7	0.176*"virus" + 0.169*"corona" + 0.019*"know" + 0.016*"people" + 0.015*"like" + 0.015*"want" + 0.015*"itaos" + 0.014*"realdonaldtrump" + 0.013*"need" + 0.012*"think"'
8	0.140*"coronavirus" + 0.022*"cancel" + 0.020*"italy" + 0.018*"outbreak" + 0.016*"pandemic" + 0.015*"school" + 0.015*"amid" + 0.014*"fear" + 0.013*"spread" + 0.011*"thing"'
9	0.106*"coronavirus" + 0.042*"trump" + 0.021*"iaom" + 0.021*"get" + 0.018*"home" + 0.015*"tell" + 0.015*"work" + 0.013*"it" + 0.013*"say" + 0.012*"minister"'
10	0.064*"coronavirus" + 0.024*"bad" + 0.021*"let" + 0.020*"find" + 0.019*"economy" + 0.019*"total" + 0.019*"cure" + 0.018*"vaccine" + 0.017*"mean" + 0.017*"white"'

In the process of identifying the content for the 10 topics with the highest coherence scores, word associations are employed, as shown in Table 4. The word associations that emerge in topic 6 indicate that the content is related to an awareness about the coronavirus disease, it emphasizes the importance of people taking precautions to prevent the spread of the COVID virus, including the practice of regular handwashing. Meanwhile topic 10 associated with the negative impact of the coronavirus disease on the economy of a country. People are expressing a desire for a cure or preventive measures, emphasizing the importance of taking a vaccine.

The data that has been processed through LDA models will be stored in a new corpus and then utilized for the computation of vector weights and the TF-IDF process. The results of the dataset analysis using the chosen word embedding and feature extraction methods will be divided into test and train data. The classification process using machine learning baseline will be implemented to achieve optimal performance and obtain evaluation metrics.

Table 5 Comparison of Machine Learning Classifiers with TF-IDF

Machine Learning Classifiers with TF-IDF					
Models / Dataset		COVIDSenti-A	COVIDSenti-B	COVIDSenti-C	COVIDSenti
TF-IDF	SVM	0.852	0.847	0.828	0.857
	NB	0.768	0.754	0.741	0.760

Table 6 Comparison Performance of Machine Learning Result

Machine Learning Classifiers with TF-IDF								
Evaluation / Dataset	COVIDSenti-A		COVIDSenti-B		COVIDSenti-C		COVIDSenti	
	SVM	NB	SVM	NB	SVM	NB	SVM	NB
Precision	0.870	0.770	0.860	0.750	0.85	0.740	0.796	0.760
Recall	0.950	1.000	0.950	1.000	0.94	1.000	0.673	1.000
F1-Score	0.910	0.870	0.910	0.86	0.89	0.85	0.720	0.860

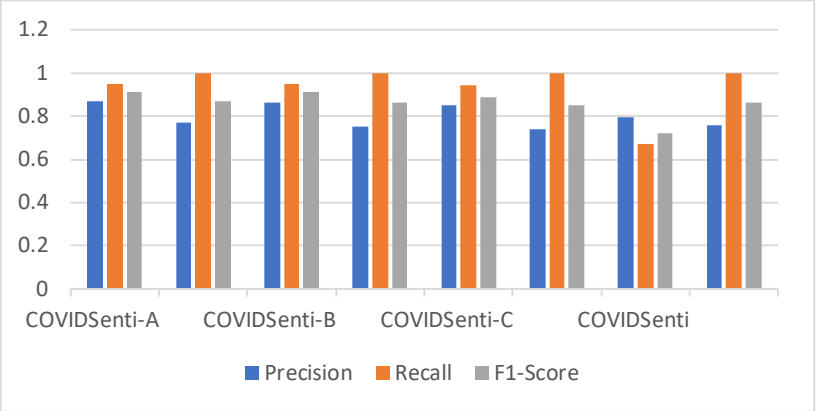

Figure 5: Comparison Performance of Machine Learning Result

4. Results and Discussion

This study involves various machine learning models applied to four COVIDSenti datasets, assessing their performance through predefined evaluation metrics. The top 10 frequently used keywords were gathered from the tweets in the COVIDSenti dataset, and the outcomes are showcased in a graphical representation in Figure 2. Furthermore, Figure 3 illustrates a word cloud highlighting the most commonly used words in those tweets within our corpus.

The most representative words for each of the 10 topics are typically displayed in a table, as shown in Table 2. This table provides insights into the prominent terms associated with each topic, aiding in the interpretation and labeling of the topics.

From the decisions of LDA models, it can be observed that the highest utilization of topics in this study is for topics number 6 and 10. The word associations within topic 6 indicate that the content is related to raising awareness about the coronavirus disease. It underscores the importance of individuals taking precautions to prevent the spread of the COVID virus, with an emphasis on practices such as regular handwashing. It appears that the words associated with topic 10 revolve around the negative impact of the coronavirus disease on the economy of a country. People are expressing a desire for a cure or preventive measures, emphasizing the importance of taking a vaccine.

The proposed method is employed to measure effectiveness in sentiment classification tasks. The table 5-6 showcase the highest accuracy achieved, emphasizing noteworthy results. The implementation of the proposed model, which combines LDA models with ensemble feature extraction involving Count Vectorizer and TF-IDF, demonstrates promising results in machine learning classification.

This study employs repeated lemmatization to enhance results and tackle challenges in obtaining accurate word roots, emphasizing its effectiveness over stemming to preserve feature quality. Additionally, LDA modeling identifies and explores ten essential topics related to Coronavirus, deepening the research's understanding.

It's great to see the positive impact of using an SVM classifier for COVIDSenti, with a notable 1.2% increase in accuracy compared to the model proposed by Naseem 2021 (Naseem et al., 2021) utilizing TF-IDF and machine learning classifiers. This underscores the significance of preprocessing methods in boosting performance accuracy. Additionally, the importance of correct feature extraction is highlighted, demonstrating its positive impact on achieving favorable results.

5. Conclusions

In the midst of the rise in COVID-19 conspiracy theories, social media has emerged as a widespread platform for both spreading and debunking misinformation and misconceptions. This article delves into the sentiments expressed on Twitter regarding COVID-19-related tweets. Within the context of COVID-19 sentiment, various sentiment analysis approaches are examined. As a result, there is an urgent requirement to establish a proactive and flexible governmental and public health presence to combat the dissemination of false information.

In subsequent research, three proposed future works include: 1) *Enhancement of LDA Models*: Advancing the development of LDA models to enhance their understanding of word linkages and, consequently, improve the accuracy of the classification process. This can be accomplished by strengthening data preprocessing efforts, with a specific focus on the creation of normalization dictionaries. 2) *Exploration of Alternative Word Embedding Methods*: Exploring other word embedding methods like Word2Vec, FastText, Glove, and Word2lda to investigate their impact on classification performance. 3) *Comparing the performance of Transformer Models*: Performing a comprehensive study that compares combined feature extraction utilizing Deep Learning classifiers such as CNN, BiLSTM, BERT, XLNET, and ALBERT. The objective is to analyze the impact on performance and derive insights into the effectiveness of these diverse approaches.

6. References

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. *Journal of Medical Internet Research*, *22*(4), e19016. https://doi.org/10.2196/19016

Abdulaziz, M., Alotaibi, A., Alsolamy, M., & Alabbas, A. (2021). Topic based Sentiment Analysis for COVID-19 Tweets. *International Journal of Advanced Computer Science and Applications*, *12*(1), 626–636. https://doi.org/10.14569/IJACSA.2021.0120172

Apuke, O. D., & Omar, B. (2021). Fake news and COVID-19: modelling the predictors of fake news sharing among social media users. *Telematics and Informatics*, *56*(March 2020), 101475. https://doi.org/10.1016/j.tele.2020.101475

Avasthi, S., Chauhan, R., & Acharjya, D. P. (2022). *Information Extraction and Sentiment Analysis to Gain Insight into the COVID-19 Crisis*. *January*, 343–353. https://doi.org/10.1007/978-981-16-2594-7_28

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment Analysis Is a Big Suitcase. *IEEE Intelligent Systems*, *32*(6), 74–80. https://doi.org/10.1109/MIS.2017.4531228

Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., & Hassanien, A. E. (2020). Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Applied Soft Computing Journal*, *97*, 106754. https://doi.org/10.1016/j.asoc.2020.106754

Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews*. https://doi.org/10.1145/1014052.1014073

Kausar, M. A., Soosaimanickam, A., & Nasar, M. (2021). Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak. *International Journal of Advanced Computer Science and Applications*, *12*(2), 415–422. https://doi.org/10.14569/IJACSA.2021.0120252

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., & Kim, J. (2021). COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. *IEEE Transactions on Computational Social Systems*, *8*(4), 976–988. https://doi.org/10.1109/TCSS.2021.3051189

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., & Eryiğit, G. (2016). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 19–30. https://doi.org/10.18653/v1/S16-1002

Priya, A., & Kumar, A. (2021). Deep Ensemble Approach for COVID-19 Fake News Detection from Social Media. *Proceedings of the 8th International Conference on Signal Processing and Integrated Networks, SPIN 2021*, 396–401. https://doi.org/10.1109/SPIN52536.2021.9565958

Rapanta, C., Botturi, L., Goodyear, P., Guàrdia, L., & Koole, M. (2020). Online University Teaching During and After the Covid-19 Crisis: Refocusing Teacher Presence and Learning Activity. *Postdigital Science and Education*, *2*(3), 923–945. https://doi.org/10.1007/s42438-020-00155-y

Raza, G. M., Butt, Z. S., Latif, S., & Wahid, A. (2021). Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models. *2021 International Conference on Digital Futures and Transformative Technologies, ICoDT2 2021*. https://doi.org/10.1109/ICoDT252288.2021.9441508

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLoS ONE*, *16*(2), 1–23. https://doi.org/10.1371/journal.pone.0245909

Sayed, S. A. F., Elkorany, A. M., & Mohammad, S. S. (2021). Applying Different Machine Learning Techniques for Prediction of COVID-19 Severity. *IEEE Access*, *9*, 135697–135707. https://doi.org/10.1109/ACCESS.2021.3116067

World Health Organization. (2021). WHO Coronavirus (COVID-19) Dashboard. In *WHO.int.*

Yousefinaghani, S., Dara, R., Mubareka, S., Papadopoulos, A., & Sharif, S. (2021). An analysis of COVID-19 vaccine sentiments and opinions on Twitter. *International Journal of Infectious Diseases*, *108*, 256–262. https://doi.org/10.1016/j.ijid.2021.05.059