Ensemble Learning for Sentiment Analysis on Twitter Data Related to Covid-19 Preventions

By Sulistyo Sancoko

Ensemble Learning for Sentiment Analysis on Twitter Data Related to Covid-19 Preventions

Sulistyo Dwi Sancoko
Faculty of Science and Engineering,
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
sulistyo.dwisancoko@staff.uty.ac.id

Saucha Diwandari
Faculty of Science and Engineering
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
saucha.diwandari@staff.uty.ac.id

Muhammad Fachrie
Faculty of Science and Engineering
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
muhammad.fachrie@staff.uty.ac.id

***Abstract*— The World Health Organization (WHO) declared the 2019 Coronavirus disease outbreak (Covid-19) as a pandemic and made it a trending topic on social media platforms, such as Facebook and Twitter. Unfortunately, news and opinions shared on social media affect people's mentality and create panic situations in society, but in the other hand, these opinions can be analyzed using sentiment analysis approach to generate knowledge and insight for the local government to monitor people reaction to the policies that have been issued to prevent the outbreak of Covid-19 virus. Therefore, this work aimed to propose an ensemble learning model that can classify the sentiment inside the people's opinions from Twitter. The ensemble model used Naïve Bayes Classifier, C4.5, and k-Nearest Neighbors as base learners with voting mechanism to generate the final decision. For learning, the ensemble model used a dataset containing 3884 clean data that was successfully downloaded using Twitter API related to Covid-19 outbreak prevention and processed using TF-IDF method. The dataset has two classes, i.e., 'positive' and 'negative' to represent the sentiment of the opinion in each data. The proposed model got 80.61% of accuracy, 79.49% of recall, and 81.20% of precision, after being evaluated using 10-fold Cross Validation. It also performed better when compared to several learning models using only single Machine Learning algorithm.**

Keywords*—*sentiment analysis, text mining, ensemble learning, covid-19, naïve bayes, decision tree, k nearest neighbors

I. Introduction

The Covid-19 pandemic that spread throughout 2020 to 2022 is a new disease that first appeared in Wuhan, China. The Covid-19 pandemic has forced many countries around the world to take steps to control the spread of the virus by encouraging people to follow preventive procedures recommended by World Health Organization (WHO), such as wearing face masks, washing hands regularly, and practicing social distancing. Several countries have imposed quarantines to slow the spread of the virus. According to the latest statistics published by WHO, currently more than 500 million people worldwide have been infected with the COVID-19 virus and more than 6 million in Indonesia as of July 2022. The current situation has prompted people from all fields to provide quick solutions to overcome the pandemic.

The effectiveness of preventive procedures must be measured to ensure that the policies issued by the government in tackling the spread of the virus are appropriate. This is not an easy task, because ensuring that people remain obedient to health protocols is not just a one-time job, but a continuous work. On the other hand, every policy must be issued based on data and logical reasons to make it effective. In this part, the challenge is how to collect data and extract them to generate useful knowledge and insights for decision makers to make effective policies. One solution that can be conducted is making use of social media data, such as Twitter or others, which contain a lot of public opinions on certain topics that is currently trend, then convert the data into useful knowledge related to preventive procedures during Covid-19 pandemic.

Social media, e.g., Twitter, Facebook, or Instagram, plays significant role in shaping public opinion on various issues through the internet. Every user can post their opinion through social media platforms, and at the same time, thousands of other users can view the post and provide feedback via comments or instant emoticons. Indirectly, social media contributes to forming new habits to human daily activities, especially in the way people share ideas or opinions. Therefore, it is reasonable that in every second there are more than 10.000 of tweets are posted to Twitter and more than 1.000 picture are posted to Instagram, according to internetlivestats.com.

Related to data analytics or data mining discipline, social media has become a new source of data with various topics that can easily accessed through certain procedures. The task that is mostly conducted using social media data is sentiment analysis. Sentiment analysis is currently a very popular research topic within natural language processing field [1]. It can process texts from human and classify the sentiment extracted from them into positive, negative, or neutral, depending on the set of of words in the text. Although it is a challenging task to process human natural language which is usually unstructured, contains various styles, has different dialects, and is influenced by temporary social trend, it can generate useful insight for human in analyzing a lot of comments and reviews from various sources, such as social media, online shop platform, online news portal, etc. It is widely utilized for many purposes, such as for customer service experience [2]–[7], product review [8]–[14], stock market prediction [15]–[21], and even general election prediction [22]–[34].

Since the outbreak of Covid-19 pandemic in 2019, governments in all affected countries have imposed lockdown regulations to cut the people's mobility to stay at home. The regulation has forced the public to spend more time using social media and shared their thoughts and opinions regarding to Covid-19 pandemic. This has encouraged many researchers to use sentiment analysis to analyze people opinion related to people reaction to the outbreak of the virus. The insights obtained from sentiment analysis can be useful for the government or other relevan agencies to find out the current situation in society and make certain decisions based on public opinion.

This work aims to propose an ensemble learning model to classify sentiments from Twitter data according to prevention of Covid-19 outbreak. This approach is expected to perform better than classification model that uses single learning algorithm. The proposed model is evaluated using 10-fold Cross Validation and the performance is also compared to classification model with single learning algorithm.

II. Related Works

During the Covid-19 pandemic, many researchers worked on sentiment analysis research on various topics related to Covid-19 issue using Twitter data. The most popular topic is about public reaction to the outbreak of the virus which was conducted by [35]–[43]. The purpose of the works is to find out the public sentiment related to the spread of Covid-19 virus. As the spread of the virus escalated, governments in the affected countries enforce lockdown regulations to area with high infection rate and restricted visitors from abroad to control the spread of the virus. Many works has been conducted to analysis the sentiment of Twitter user related to lockdown regulation as in [44]–[49]. The pandemic has also forced students to study online and almost never attend school meetings. Of course, this online study has an impact on many things for students, such as learning motivation, imperfect absorption of knowledge, also disruption in financial support. Several works in this topic has been conducted as in [50]–[53]. In addition, the researcher also conducted a sentiment analysis on public opinion regarding the Covid-19 vaccination as in [54]–[58].

Previous works on sentiment analysis mostly conducted based on Naïve Bayes classifier since it has efficient performance compared to other Machine Learning algorithms [59], [60]. Naïve Bayes uses the concept of probability to classify data, and it works better in sentiment analysis since the extraction feature, i.e., TF-IDF, also work in the basis of probability measurement. According to [61], Naïve Bayes is also considered as a simple learning algorithm and easy to implement. It does not require long training time to learn data patterns as in Support Vector Machine (SVM), Multilayer Perceptron (MLP), or Deep Learning. In [50], [52], Naïve Bayes was successfully used in classifying sentiments according to Covid-19 related tweets. In general, Naïve Bayes shows good performance in small dataset, but it has poor accuracy when used in large dataset.

As Deep Learning is becoming more popular with significant improvement in classification tasks, many previous works according to sentiment analysis in Covid-19 issues used Deep Learning technique to obtain better performance. Long Short Term Memory (LSTM) is one of the most used Deep Learning architecture for sentiment analysis as conducted in [38], [43], [46], [56]. In general, LSTM and other Deep Learning architectures outperforms conventional Machine Learning algorithm in sentiment analysis, such as Naïve Bayes classifier, Support Vector Machine, and Neural Networks [62], [63]. Unfortunately, LSTM and other Deep Learning architectures require long training times to achieve the best performance. In addition, hyperparameter tuning is also needed to find the best LSTM architecture which is usually conducted through several observations to find out the most optimal one.

The other approach to obtain better performance is using ensemble learning model which is an integrated model that combines several "individual learner", such as Naïve Bayes, Support Vector Machine, Decision Tree, Neural Networks, etc. into a new model [61]. One of the basic strategies in ensemble model is doing voting from each individual learner and make decision based on majority result from all individual learners. This approach can provide better classification performance compared to basic model that only uses a single ML algorithm, but on the other hand it does not require long training times as in Deep Learning. In [37], an ensemble learning model consisting of Multilayer Perceptron and Naïve Bayes classifier was proposed for classifying sentiments on Covid-19 related tweets.

In this paper, an ensemble learning model is proposed with voting mechanism. There are three ML algorithms used as individual learner, i.e., Naïve Bayes classifier, k-Nearest Neighbors, and Decision Tree (C4.5). These three algorithms are chosen because they classify data in a very different ways where Naïve Bayes works based on probability measurement, k-Nearest Neighbors works based on distance measurement, while Decision Tree uses rule based (IF-THEN) classification model. These methods were performed in vary fields research such medical image [64]–[66] and plants [67]. The proposed model has been evaluated using 10-fold Cross Validation and compared to models with a single learning algorithm using Naïve Bayes classifier, k-Nearest Neighbors, and Decision Tree (C4.5), respectively.

III. Methodology

In this research, sentiment analysis is utilized to classify Covid-19 related tweets into two different classes, i.e., positive and negative using an ensemble learning model to achieve better performance. There are several works that was conducted as described in Fig. 1: data collection, data preprocessing, feature selection, data classification, and model evaluation. Data collection is the initial step to collect raw data from Twitter using Twitter API, then process them to obtain clean data. At the data preprocessing stage, data labeling is also performed manually by humans. This part took quiet long time to create good dataset. The feature from labeled dataset is extracted using TF-IDF method. To ensure that the dataset has good quality, a feature selection occurs to remove some misspelling words and useless words that rarely appears in documents.

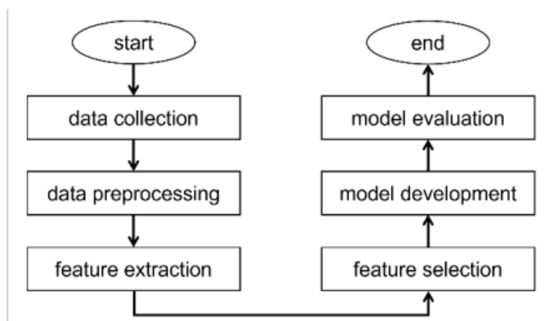

Fig. 1. Research methodology used in this work

Data classification is performed using ensemble learning model with voting mechanism where Naïve Bayes classifier, k-Nearest Neighbors, and Decision Tree (C4.5) are used as base learners. Model performance is measured based on accuracy, precision, and recall. For comparison purpose, the proposed model is compared to accuracy of single classifier model.

A. Data Collection

The dataset was obtained from Twitter using Twitter API through Rapidminer. Several keywords in Indonesian were used to collect the tweets according to Covid-19 prevention procedures, i.e., 'cegah covid', 'omicron', 'cuci tangan covid', 'covid 19', 'pakai masker', 'hindari kerumunan', 'jaga jarak covid', 'kurangi mobilitas', 'vaksin', and 'omicron BA4'. There are total of 9269 records of raw data that was successfully downloaded from 22nd of July until 28th of July 2022. These raw data consist of several attributes, such as 'user-id', 'created-at', 'from-user', 'to-user', 'language', text, etc., and still have noise that must be cleaned before it can be used for analysis. For sentiment analysis purpose, only text attribute from raw data that is used.

B. Data Preprocessing

The raw data can not be used immediately for sentiment analysis. Therefore, several data preprocessing steps should be applied to raw data to produce a clean dataset. As described in Fig. 2, the data preprocessing is performed by removing the hashtag symbol '#...', retweet symbol 'RT:', URL 'http…', account mentions '@...', and other symbols. In this stage, duplicate records are also removed, and only 3884 clean data remained. Each data is then labeled manually based on its sentiment, i.e., positive, or negative.

Tokenization is performed once the clean dataset is obtained. It is conducted by separating each word in the text as single token to be processed individually. Every token is then transformed to lowercase to make each letter uniform. Several stopwords are also removed to make the data more efficient and only contain important words to analyze. An additional process is also performed to filter tokens by length. In this work, tokens that contain less than 4 characters will be removed. This task aims to eliminate useless words outsite the stopword list. Finally, stemming is used to normalize each token by converting the word to its base form without any prefixes or affixes. Once the stemming done, the features from clean dataset can be extracted using TF-IDF method.

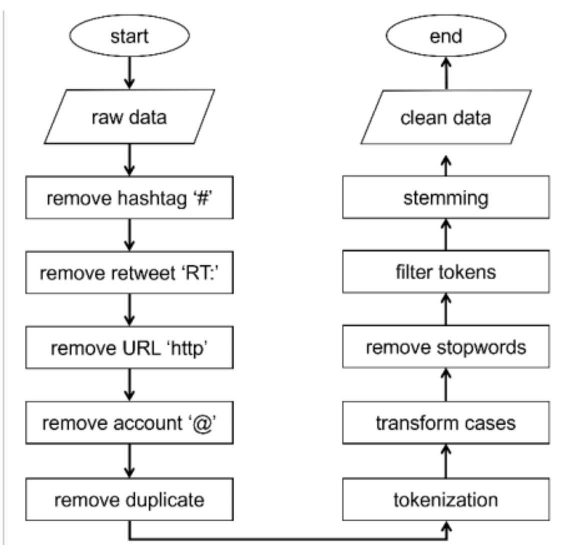

Fig. 2. Flowchart of data preprocessing for sentiment analysis

C. Feature Extraction

The clean dataset can not be used immediately for sentiment analysis. It needs to be further processed by extracting the feature from dataset using TF-IDF method. TF-IDF is popular method to extract feature from text data based on the frequency of word (term) in each sentence. TF-IDF for each term is calculated using (1).

$$TF-IDF(t,d) = TF(t,d) \times IDF(t,D) \qquad (1)$$

with

$$TF(t,d) = \frac{f_{t,d}}{N_d} \qquad (2)$$

and

$$IDF(t,D) = \log\left(\frac{N}{1+df_t}\right) \qquad (3)$$

Where $f_{t,d}$ is number of times the term t appears in the document d, N_d is total number of terms t in document d, N is the total number of documents, and df_t is the number of documents that contain term t. The addition by 1 in denominator of (3) aims to avoid the division by zero.

By using this method, the resulting dataset is a TF-IDF matrix where each column represents a single term from all documents (data) and each row represents the weight or TF-IDF score for each term. If the weight of term t in document d is 0, it means that t does not appear in d. The size of TF-IDF matrix can be very large, depending on the number of terms and documents. In this work, the TF-IDF matrix has 6453 columns with 3884 rows. However, the size of this matrix can be reduced to make the system works more efficiently.

D. Feature Selection

In this work, there is an additional stage to enhance the TF-IDF matrix by removing useless term (feature). This can help to make the system works more efficiently by using only essential terms instead of using all the terms from TF-IDF processing. Feature selection is simply conducted manually by checking which term appears in less than 2 documents (data) then excluding them from the TF-IDF matrix. This mechanism can reduce the number of terms significantly. In addition, terms that occur too often are also omitted to avoid bias in data classification because terms that appear in many documents indicate that these terms are not unique and have no significant effect on data classification. After performing feature selection to TF-IDF matrix resulted from the feature extraction step, there are only 1886 terms left. It is much more efficient than using the original TF-IDF matrix.

E. Model Development

Text mining is simply a classification task. In case of sentiment analysis, the text is classified into posstive or negative class. In many previous works, Machine Learning (ML) algorithms such as Naïve Bayes Classifier, Support Vector Machine, or Decision Tree, show the good result in sentiment analysis. In this work, an ensemble learning model is utilized using voting mechanism. This ensemble model combines several ML algorithms as base learners, to perform individual classifications from given dataset, then generates the final decision by taking the most decisions (voting) from each ML base learner. There are three base learners used in this work, i.e., Naïve Bayes Classifier, Decision Tree (C4.5), and k-Nearest Neighbors (kNN) as illustrated in Fig. 3.

The final decision of the model is gained from voting mechanism which is performed equally for each base learner. This ensemble model is expected to have better performance than using classification model with single ML algorithm.

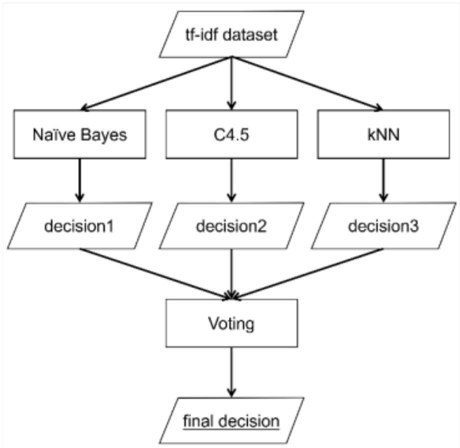

Fig. 3. The proposed ensemble learning model using voting mechanism

F. Model Evaluation

Instead of using split validation to train and test the ensemble learning model, this work used 10-fold Cross Validation to perform model evaluation, which is illustrated in Fig. 4. It aims to measure the performance comprehensively using several variations of training and testing data. Then, confusion matrix is used to evaluate the performance score resulted from Cross Validation stage. There are three metrics used to measure the performance of the model, i.e., accuracy, precision, and recall. The performance of proposed model is also compared to classification model that uses only single ML algorithm. This mechanism aims to evaluate whether the ensemble model performs better than the single ML learning model.

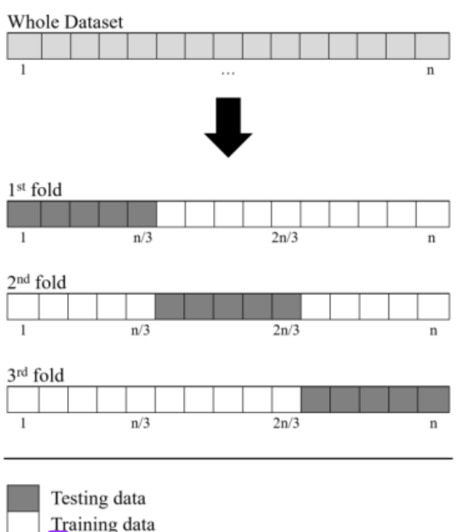

Fig. 4. The illustration of k-fold Cross Validation mechanism with k=3

IV. Result and Discussion

There are two stages of experiment carried out in this work. The first stage is training and testing single learning models, and the second stage is training and testing the ensemble learning model to be compared with the single model. Both stages were performed using 10-fold Cross Validation with stratified sampling.

A. Dataset Description

The experiments were performed using TF-IDF dataset that has been filtered in feature selection stage. There are total of 3884 records with 1886 features (terms). The dataset is considered as a balanced dataset since the number of records labeled 'positive' is slightly equal to the number of records labeled 'negative', i.e., 1735 and 1370 records respectively. All the features are in real number type.

B. Single Learning Model Result

Sentiment analysis using Naïve Bayes Classifier with Laplacian Correction get a good performance result with accuracy of 79.16%, average recall of 78.53%, and average precision of 79.02%. These results show the reason why the Naïve Bayes Classifier is popular in sentiment analysis tasks, as used in many previous works. As shown in confusion matrix in Table I, both 'positive' and 'negative' classes have good results in precision and recall.

TABLE I. Confusion Matrix of Naïve Bayes Learning Model

	True Positive	True Negative	Class Precision
Pred. Positive	1456	368	79.82%
Pred. Negative	279	1002	78.22%
Class Recall	83.92%	73.14%	

Compared to Naïve Bayes Classifier, C4.5 algorithm get worse result with much lower accuracy of 68.08%, average recall of 64.05%, and average precision of 78.75%. As seen in Table II, the 'negative' class has very low recall. Fortunately, the class precision of 'negative' class is high. The TF-IDF dataset with high-dimensional features does not seem to fit the Decision Tree technique in this case.

TABLE II. Confusion Matrix of C4.5 Decision Tree Model

	True Positive	True Negative	Class Precision
Pred. Positive	1707	963	63.93%
Pred. Negative	28	407	93.56%
Class Recall	98.39%	29.71%	

KNN algorithms gives moderate result with accuracy of 75.65%, average recall of 76.74%, and average precision of 76.77%. This result is better than C4.5, but slightly lower than Naïve Bayes Classifier. Based on several observations, the best k neighbors value is 11. Table III shows the confusion matrix of kNN model evaluation. The 'positive' class has quite low score compared to the 'negative' class.

TABLE III. Confusion Matrix of k-Nearest Neighbors Model

	True Positive	True Negative	Class Precision
Pred. Positive	1171	192	85.91%
Pred. Negative	564	1178	67.62%
Class Recall	67.49%	85.99%	

C. Ensemble Learning Model Result

The proposed ensemble learning model obtained better result than any single learning model that is mentioned before, with higher accuracy of 80.61%, higher average precision of 79.49%, and higher average recall of 81.20%. Based on Table IV, the combination of Naïve Bayes Classifier, C4.5, and kNN produces good class recall and precision.

TABLE IV. CONFUSION MATRIX OF ENSEMBLE LEARNING MODEL

	True Positive	True Negative	Class Precision
Pred. Positive	1545	412	78.95%
Pred. Negative	190	958	83.45%
Class Recall	89.05%	69.93%	

D. Performance Comparison

Naïve Bayes Classifier gives better result than C4.5 and kNN algorithms. However, the ensemble learning model still performed better than any single learning model as given in Table V. The proposed ensemble model scored higher in all evaluation metrics i.e., accuracy, precision, and recall. Based on this result, C4.5 has the worst performance, but it scored high in class recall. However, the performance is still far from 90% of accuracy, so it needs improvement in the next work. This can be realized using several approaches, such as automatic feature selection or automatic parameter optimization based on Genetic Algorithm (GA). The other strategy is by adding weight to each base learner.

TABLE V. COMPARISON BETWEEN ENSEMBLE LEARNING MODEL AND SINGLE LEARNING MODEL

Metrics	Naïve Bayes	C4.5	kNN	Ensemble
Accuracy	79.16%	68.08%	75.65%	**80.61%**
Recall	78.53%	64.05%	76.74%	**79.49%**
Precision	79.02%	78.75%	76.77%	**81.20%**

The proposed model is also evaluated using TF-IDF dataset that does not use feature selection procedure. The purpose is to find out the effect of feature selection system performance. As shown in Table VI, ensemble learning model still outperformed any single learning model. The performance of the model that using feature selection is slightly better than the model without feature selection. This result prove that feature selection has positive impact on improving the performance of proposed model by reducing the dataset dimension (removing useless features) and increasing the system accuracy, recall, and precision.

TABLE VI. EVALUATION OF USING FEATURE SELECTION

Metrics	Naïve Bayes	C4.5	kNN	Ensemble
Accuracy	77.07%	68.12%	74.69%	**80.42%**
Recall	77.74%	78.68%	75.90%	**80.13%**
Precision	77.94%	64.09%	75.81%	**80.24%**

V. CONCLUSION

Sentiment analysis is a great tool that can be used to extract people's opinions from social media platforms. This can be useful for many purposes. In case of Covid-19 pandemic, sentiment analysis can provide information about how people response to government's policy on preventing the outbreak of Covid-19 viruses. By knowing this information, government can take decision, for example to extend the policy period or to end the policy.

In this work, an ensemble learning model with voting mechanism is proposed to classify sentiment from Twitter data related to Covid-19 prevention procedures. The ensemble learning model combines three different Machine Learning (ML) algorithms that act as base learners, i.e., Naïve Bayes Classifier, Decision Tree (C4.5), and k-Nearest Neighbors (kNN). Based on the experiments, the proposed ensemble model achieved higher performance in accuracy, precision, also recall compared to any single learning model.

However, the proposed model still needs improvement to achieve higher performance up to 90% or even higher. This can be attempted using some stragies, such as doing automatic feature selection or automatic learning parameter optimization using Genetic Algorithm (GA). Moreover, adding weight to each base learner also can be an effective strategy to improve the proposed model.

ACKNOWLEDGMENT *(Heading 5)*

We would like to thank to Direktorat Riset, Teknologi, dan Pengabdian Masyarakat, Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi fiscal year of 2022, number SP DIPA-023.17.1.690523/2022 2nd revision on April 22nd, 2022.

REFERENCES

[1] J. Yang and J. Yang, *Aspect Based Sentiment Analysis With Self-Attention And Gated Convolutional Networks*, in *2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)*, 2020, pp. 146–149.

[2] A. Rane and A. Kumar, *Sentiment Classification System Of Twitter Data For US Airline Service Analysis*, in *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, 2018, vol. 01, pp. 769–773.

[3] K. Zvarevashe and O. O. Olugbara, *A Framework For Sentiment Analysis With Opinion Mining Of Hotel Reviews*, in *2018 Conference on Information Communications Technology and Society (ICTAS)*, 2018, pp. 1–4.

[4] P. Porntrakoon and C. Moemeng, *Thai Sentiment Analysis For Consumer's Review In Multiple Dimensions Using Sentiment Compensation Technique (SenseComp)*, in *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2018, pp. 25–28.

[5] S. Vanaja and M. Belwal, *Aspect-Level Sentiment Analysis On E-Commerce Data*, in *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2018, pp. 1275–1279.

[6] V. Ikoro, M. Sharmina, K. Malik, and R. Batista-Navarro, *Analyzing Sentiments Expressed On Twitter By UK Energy Company Consumers*, in *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2018, pp. 95–98.

[7] N. Srivats Athindran, S. Manikandaraj, and R. Kamaleshwar, *Comparative Analysis Of Customer Sentiments On Competing Brands Using Hybrid Model Approach*, in *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*, 2018, pp. 348–353.

[8] L. Rong, Z. Weibai, and H. Debo, *Sentiment Analysis Of Ecommerce Product Review Data Based On Deep Learning*, in *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, 2021, vol. 4, pp. 65–68.

[9] Pankaj, P. Pandey, Muskan, and N. Soni, *Sentiment Analysis On Customer Feedback Data: Amazon Product Reviews*, in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 320–322.

[10] S. Smetanin and M. Komarov, *Sentiment Analysis Of Product Reviews In Russian Using Convolutional Neural Networks*, in *2019 IEEE 21st Conference on Business Informatics (CBI)*, 2019, vol. 01, pp. 482–486.

[11] C. Chauhan and S. Sehgal, *Sentiment Analysis On Product Reviews*, in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 2017, pp. 26–31.

[12] Z. A. Guven, *The Effect Of BERT, ELECTRA And ALBERT Language Models On Sentiment Analysis For Turkish Product Reviews*, in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, 2021, pp. 629–632.

[13] S. N. Singh and T. Sarraf, *Sentiment Analysis Of A Product Based On User Reviews Using Random Forests Algorithm*, in *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2020, pp. 112–116.

[14] S. Wladislav, Z. Johannes, W. Christian, K. André, and F. Madjid, *Sentilyzer: Aspect-Oriented Sentiment Analysis Of Product Reviews*, in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2018, pp. 270–273.

[15] Y. Deng, Q. Xie, and Y. Wang, *Research On Investor Sentiment And Stock Market Prediction Based On Weibo Text*, in *2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, 2018, pp. 1–5.

[16] M. V. D. H. . Malawana and R. M. K. T. Rathnayaka, *The Public Sentiment Analysis Within Big Data Distributed System For Stock Market Prediction– A Case Study On Colombo Stock Exchange*, in *2020 5th International Conference on Information Technology Research (ICITR)*, 2020, pp. 1–6.

[17] R. Srivastava, S. Agarwal, D. Garg, and J. C. Patni, *Capital Market Forecasting By Using Sentimental Analysis*, in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 2016, pp. 9–12.

[18] H. Bourezk, A. Raji, N. Acha, and H. Barka, *Analyzing Moroccan Stock Market Using Machine Learning And Sentiment Analysis*, in *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 2020, pp. 1–5.

[19] M. Pouromid, A. Yekkehkhani, M. A. Oskoei, and A. Aminimehr, *ParsBERT Post-Training For Sentiment Analysis Of Tweets Concerning Stock Market*, in *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, 2021, pp. 1–4.

[20] R. Ahuja, H. Rastogi, A. Choudhuri, and B. Garg, *Stock Market Forecast Using Sentiment Analysis*, in *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2015, pp. 1008–1010.

[21] D. Shah, H. Isah, and F. Zulkernine, *Predicting The Effects Of News Sentiments On The Stock Market*, in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 4705–4708.

[22] S. Salari, N. Sedighpour, V. Vaezinia, and S. Momtazi, *Estimation Of 2017 Iran's Presidential Election Using Sentiment Analysis On Social Media*, in *2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, 2018, pp. 77–82.

[23] L. Wang and J. Q. Gan, *Prediction Of The 2017 French Election Based On Twitter Data Analysis*, in *2017 9th Computer Science and Electronic Engineering (CEEC)*, 2017, pp. 89–93.

[24] R. Rezapour, L. Wang, O. Abdar, and J. Diesner, *Identifying The Overlap Between Election Result And Candidates' Ranking Based On Hashtag-Enhanced, Lexicon-Based Sentiment Analysis*, in *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, 2017, pp. 93–96.

[25] P. Sharma and T.-S. Moh, *Prediction Of Indian Election Using Sentiment Analysis On Hindi Twitter*, in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 1966–1971.

[26] F. A. Wenando, R. Hayami, Bakaruddin, and A. Y. Novermahakim, *Tweet Sentiment Analysis For 2019 Indonesia Presidential Election Results Using Various Classification Algorithms*, in *2020 1st International Conference on Information Technology, Advanced Mechanical and Electrical Engineering (ICITAMEE)*, 2020, pp. 279–282.

[27] F. J. J. Joseph, *Twitter Based Outcome Predictions Of 2019 Indian General Elections Using Decision Tree*, in *2019 4th International Conference on Information Technology (InCIT)*, 2019, pp. 50–53.

[28] D. K. Nugroho, *US Presidential Election 2020 Prediction Based On Twitter Data Using Lexicon-Based Sentiment Analysis*, in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2021, pp. 136–141.

[29] O. Oyebode and R. Orji, *Social Media And Sentiment Analysis: The Nigeria Presidential Election 2019*, in *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2019, pp. 140–146.

[30] M. Ibrahim, O. Abdillah, A. F. Wicaksono, and M. Adriani, *Buzzer Detection And Sentiment Analysis For Predicting Presidential Election Results In A Twitter Nation*, in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015, pp. 1348–1353.

[31] M. Rodríguez-Ibáñez, F.-J. Gimeno-Blanes, P. M. Cuenca-Jiménez, C. Soguero-Ruiz, and J. L. Rojo-Álvarez, *Sentiment Analysis Of Political Tweets From The 2019 Spanish Elections*, vol. 9, pp. 101847–101862, 2021

[32] S. Singh and G. Sikka, *YouTube Sentiment Analysis On US Elections 2020*, in *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, 2021, pp. 250–254.

[33] B. Joyce and J. Deng, *Sentiment Analysis Of Tweets For The 2016 US Presidential Election*, in *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, 2017, pp. 1–4.

[34] R. Jose and V. S. Chooralil, *Prediction Of Election Result By Enhanced Sentiment Analysis On Twitter Data Using Word Sense Disambiguation*, in *2015 International Conference on Control Communication & Computing India (ICCC)*, 2015, pp. 638–641.

[35] S. Andhale, P. Mane, M. Vaingankar, D. Karia, and K. T. Talele, *Twitter Sentiment Analysis For COVID-19*, in *2021 International Conference on Communication information and Computing Technology (ICCICT)*, Jun. 2021, pp. 1–12.

[36] J. Huang, *Prediction Of Emotional Tendency Of COVID-19 Speech Based On Pre-Training Model*, in *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, 2021, vol. 4, pp. 345–349.

[37] M. H. Tsai and Y. Wang, *A New Ensemble Method For Classifying Sentiments Of COVID-19-Related Tweets*, in *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, Dec. 2020, pp. 313–316.

[38] A. Tareq and N. Hewahi, *Sentiment Analysis Of Tweets During COVID-19 Pandemic Using BLSTM*, in *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, 2021, pp. 245–249.

[39] Imamah and F. H. Rachman, *Twitter Sentiment Analysis Of Covid-19 Using Term Weighting TF-IDF And Logistic Regresion*, in *2020 6th Information Technology International Seminar (ITIS)*, 2020, pp. 238–242.

[40] I. Gupta, I. Chatterjee, and N. Gupta, *Sentiment Analysis Of COVID-19 Tweets*, in *2022 1st International Conference on Informatics (ICI)*, 2022, pp. 229–231.

[41] Z. Tariq Soomro, S. H. Waseem Ilyas, and U. Yaqub, *Sentiment, Count And Cases: Analysis Of Twitter Discussions During COVID-19 Pandemic*, in *2020 7th International Conference on Behavioural and Social Computing (BESC)*, 2020, pp. 1–4.

[42] G. Matošević and V. Bevanda, *Sentiment Analysis Of Tweets About COVID-19 Disease During Pandemic*, in *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2020, pp. 1290–1295.

[43] M. S. A. Pran, M. R. Bhuiyan, S. A. Hossain, and S. Abujar, *Analysis Of Bangladeshi People's Emotion During Covid-19 In Social Media Using Deep Learning*, in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2020, pp. 1–6.

[44] D. Kumar and U. Bansal, *Emotion Sentiment Analysis Of Indian Twitter-Data Of COVID-19 After Lockdown*, in *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, May 2021, pp. 421–426.

[45] P. Tyagi, N. Goyal, and T. Gupta, *Analysis Of COVID-19 Tweets During Lockdown Phases*, in *2021 9th International Conference on Information and Education Technology (ICIET)*, Mar. 2021, pp. 471–475.

[46] K. Saini, D. K. Vishwakarma, and C. Dhiman, *Sentiment Analysis Of Twitter Corpus Related To COVID-19 Induced Lockdown*, in *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, May 2021, pp. 465–470.

[47] P. Gupta, S. Kumar, R. R. Suman, and V. Kumar, *Sentiment Analysis Of Lockdown In India During COVID-19: A Case Study On Twitter*,

vol. 8, no. 4, pp. 992–1002, Aug. 2021

[48] N. Afroz, M. Boral, V. Sharma, and M. Gupta, *Sentiment Analysis Of COVID-19 Nationwide Lockdown Effect In India*, in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Mar. 2021, pp. 561–567.

[49] L. Sandra and S. Aritonang, *Lockdown Countdown: Lockdown Sentiment Analysis On Twitter Using Artificial Neural Network*, in *2021 International Conference on Data Science and Its Applications (ICoDSA)*, 2021, pp. 198–202.

[50] G. M. I. Alcober and T. F. Revano, *Twitter Sentiment Analysis Towards Online Learning During COVID-19 In The Philippines*, in *2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, Nov. 2021, pp. 1–6.

[51] E. Mondragon-Estrada and C. Camacho-Zuniga, *Undergraduate's Perspective On Being An Effective Online Student During Lockdown Due To COVID-19 Pandemic: An Educational Data Mining Study*, in *2021 Machine Learning-Driven Digital Technologies for Educational Innovation Workshop*, Dec. 2021, pp. 1–5.

[52] O. AlZoubi, F. Shatnawi, S. Rawashdeh, M. B. Yassein, and I. Hmeidi, *Detecting COVID-19 Implication On Education And Economic In Arab World Using Sentiment Analysis Techniques Of Twitter Data*, in *2022 13th International Conference on Information and Communication Systems (ICICS)*, Jun. 2022, pp. 352–357.

[53] S. Daulatkar and A. Deore, *Post Covid-19 Sentiment Analysis Of Success Of Online Learning: A Case Study Of India*, in *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2022, pp. 460–465.

[54] N. Bhoj, M. Khari, and B. Pandey, *Improved Identification Of Negative Tweets Related To Covid-19 Vaccination By Mitigating Class Imbalance*, in *2021 13th International Conference on Computational Intelligence and Communication Networks (CICN)*, Sep. 2021, pp. 23–28.

[55] İrfan Aygün, B. Kaya, and M. Kaya, *Aspect Based Twitter Sentiment Analysis On Vaccination And Vaccine Types In COVID-19 Pandemic With Deep Learning*, vol. 26, no. 5, pp. 2360–2369, 2022

[56] N. Wang and X. Lv, *Research On Emotional Analysis Of Netizens And Topic Distribution Under Public Health Emergencies : ——A Case Study Of COVID-19*, in *2020 International Conference on Public Health and Data Science (ICPHDS)*, 2020, pp. 76–80.

[57] R. Dumre, K. Sharma, and K. Konar, *Statistical And Sentimental Analysis On Vaccination Against COVID-19 In India*, in *2021 International Conference on Communication information and Computing Technology (ICCICT)*, 2021, pp. 1–6.

[58] H. Xu, R. Liu, Z. Luo, M. Xu, and B. Wang, *COVID-19 Vaccine Sensing: Sentiment Analysis From Twitter Data*, in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2021, pp. 3200–3205.

[59] L. Zhang, S. Wang, and B. Liu, *Machine Learning For Sentiment Analysis: A Survey*, vol. 8, no. 4, 2018

[60] E. Aydo and M. A. Akcayol, *A Comprehensive Survey For Sentiment Analysis Tasks Using Machine Learning Techniques*, 2016.

[61] P. Lin and X. Luo, *A Survey Of Sentiment Analysis Based On Machine Learning*, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12430 LNAI. pp. 372–387, 2020.

[62] S. Sukheja, S. Chopra, and M. Vijayalakshmi, *Sentiment Analysis Using Deep Learning - A Survey.Pdf.*

[63] J. Joseph, S. Vineetha, and N. V Sobhana, *A Survey On Deep Learning Based Sentiment Analysis*, vol. 58, pp. 456–460, 2022

[64] E. I. Sela, R. Pulungan, R. Widyaningrum, and R. R. Shantiningsih, *Method For Automated Selection Of The Trabecular Area In Digital Periapical Radiographic Images Using Morphological Operations*, vol. 25, no. 3, pp. 193–200, 2019

[65] E. I. Sela and R. Pulungan, *Osteoporosis Identification Based On The Validated Trabecular Area On Digital Dental Radiographic Images*, vol. 157, pp. 282–289, 2019

[66] E. I. Sela and Sutarman, *Extracting The Potential Features Of Digital Panoramic Radiograph Images By Combining Radio Morphometry Index, Texture Analysis, And Morphological Features*, vol. 14, no. 2, pp. 144–152, 2017

[67] A. F. Aji *et al.*, *Detection Of Palm Oil Leaf Disease With Image Processing And Neural Network Classification On Mobile Device*, vol. 5, no. 3, pp. 528–532, 2013

Ensemble Learning for Sentiment Analysis on Twitter Data Related to Covid-19 Preventions

ORIGINALITY REPORT

15%
SIMILARITY INDEX

PRIMARY SOURCES

1	jiki.cs.ui.ac.id Internet	55 words — 1%
2	www.ncbi.nlm.nih.gov Internet	53 words — 1%
3	Muhammad Fachrie, Farida Ardiani. "Predictive Model for Regional Elections Results based on Candidate Profiles", 2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2021 Crossref	36 words — 1%
4	mafiadoc.com Internet	35 words — 1%
5	Pinar Duygulu. "Searching for repeated video sequences", Proceedings of the international workshop on Workshop on multimedia information retrieval - MIR 07 MIR 07, 2007 Crossref	26 words — 1%
6	Dwi Juniati, I Ketut. "The Influence of Cognitive and Affective Factors on the Performance of Prospective Mathematics Teachers", European Journal of Educational Research, 2022 Crossref	23 words — 1%

7	Rianto, Arief Hermawan, P. Insap Santosa. "Knowledge and Prevention of Repetitive Strain Injury Among Computer Users", 2018 International Conference on Orange Technologies (ICOT), 2018 Crossref	20 words — < 1%
8	getjson.sid.ir Internet	19 words — < 1%
9	journalofbigdata.springeropen.com Internet	19 words — < 1%
10	www.researchgate.net Internet	19 words — < 1%
11	www.jsoftware.us Internet	18 words — < 1%
12	"Smart Trends in Information Technology and Computer Communications", Springer Science and Business Media LLC, 2016 Crossref	16 words — < 1%
13	"Sentimental Analysis and Deep Learning", Springer Science and Business Media LLC, 2022 Crossref	15 words — < 1%
14	Kybernetes, Volume 42, Issue 3 (2013-05-27) Publications	14 words — < 1%
15	María Elena Brenlla, Guadalupe Germano, Mariana S. Seivane, Rocío Fernández da Lama, Ruth Ogden. "Experiences of distortions to the passage of time during the Argentinian Covid-19 pandemic", PLoS ONE Internet	14 words — < 1%

16 Toshihide Yokouchi, Minoru Kondo. "LSTM-based Anomaly Detection for Railway Vehicle Air-conditioning Unit using Monitoring Data", IECON 2021 – 47th Annual Conference of the IEEE Industrial Electronics Society, 2021
Crossref
14 words — < 1%

17 Adal A. Alashban, Yousef A. Alotaibi. "Speaker Gender Classification in Mono-Language and Cross-Language Using BLSTM Network", 2021 44th International Conference on Telecommunications and Signal Processing (TSP), 2021
Crossref
12 words — < 1%

18 Sara Laghmati, Bouchaib Cherradi, Amal Tmiri, Othmane Daanouni, Soufiane Hamida. "Classification of Patients with Breast Cancer using Neighbourhood Component Analysis and Supervised Machine Learning Techniques", 2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet), 2020
Crossref
12 words — < 1%

19 coek.info
Internet
12 words — < 1%

20 nopren.ucsf.edu
Internet
12 words — < 1%

21 Abdullahil Kafi, M. Shaikh Ashikul Alam, Sayeed Bin Hossain, Siam Bin Awal, Hossain Arif. "Feature-Based Mobile Phone Rating Using Sentiment Analysis and Machine Learning Approaches", 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019
Crossref
11 words — < 1%

22	Kudakwashe Zvarevashe, Oludayo O. Olugbara. "A framework for sentiment analysis with opinion mining of hotel reviews", 2018 Conference on Information Communications Technology and Society (ICTAS), 2018 Crossref	10 words — < 1%
23	arxiv.org Internet	10 words — < 1%
24	"Computational Intelligence in Data Mining", Springer Science and Business Media LLC, 2020 Crossref	9 words — < 1%
25	"Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems", Springer Science and Business Media LLC, 2022 Crossref	9 words — < 1%
26	Edi Faisal, Farza Nurifan, Riyanarto Sarno. "Word Sense Disambiguation in Bahasa Indonesia Using SVM", 2018 International Seminar on Application for Technology of Information and Communication, 2018 Crossref	9 words — < 1%
27	www.mdpi.com Internet	9 words — < 1%
28	Gede Rizky Gustisa Wisnu, Ahmadi, Ahmad Rizaqu Muttaqi, Aris Budi Santoso, Prabu Kresna Putra, Indra Budi. "Sentiment Analysis and Topic Modelling of 2018 Central Java Gubernatorial Election using Twitter Data", 2020 International Workshop on Big Data and Information Security (IWBIS), 2020 Crossref	8 words — < 1%

29	K. Chitra, T. M. Saravanan, S.Naveen Prasath, G. Robin, N.K.Sriram Babu. "Sentiment Analysis on Smartphone Using Support Vector Machine", 2022 International Conference on Computer Communication and Informatics (ICCCI), 2022 Crossref	8 words — < 1%
30	Zul Indra, Azhari Setiawan, Yessi Jusman. "Implementation of Machine Learning for Sentiment Analysis of Social and Political Orientation in Pekanbaru City", Journal of Physics: Conference Series, 2021 Crossref	8 words — < 1%
31	annals-csis.org Internet	8 words — < 1%
32	cora.ucc.ie Internet	8 words — < 1%
33	diginole.lib.fsu.edu Internet	8 words — < 1%
34	dokumen.pub Internet	8 words — < 1%
35	downloads.hindawi.com Internet	8 words — < 1%
36	ir.kdu.ac.lk Internet	8 words — < 1%
37	jyx.jyu.fi Internet	8 words — < 1%
38	nti.khai.edu Internet	8 words — < 1%

39	pdfs.semanticscholar.org Internet	8 words — < 1%
40	A. Vinciarelli. "OCR Based Slide Retrieval", Eighth International Conference on Document Analysis and Recognition (ICDAR 05), 2005 Crossref	7 words — < 1%
41	Md. Jahed Hossain, Dabasish Das Joy, Sowmitra Das, Rashed Mustafa. "Sentiment Analysis on Reviews of E-commerce Sites Using Machine Learning Algorithms", 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET), 2022 Crossref	7 words — < 1%
42	Lecture Notes in Computer Science, 2013. Crossref	6 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF