Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach

Mariana Purba[1]
Doctor of Engineering
Universitas Sriwijaya, Indonesia

Ermatita Ermatita[2], Abdiansah Abdiansah[3]
Faculty of Computer Science
Universitas Sriwijaya, Indonesia

Handrie Noprisson[4], Vina Ayumi[5], Umniy Salamah[7]
Faculty of Computer Science
Universitas Mercu Buana, Indonesia

Hadiguna Setiawan[6]
Department of Computer Science
Stikhafi Indonesia

Yadi Yadi[8]
Informatics Department
Institut Teknologi Pagar Alam, Indonesia

***Abstract*—Opinion mining has been a prominent topic of research in Indonesia, however there are still many unanswered questions. The majority of past research has been on machine learning methods and models. A comparison of the effects of random splitting and cross-validation on processing performance is required. Text data is in Indonesian. The goal of this project is to use a machine learning model to conduct opinion mining on Indonesian text data using a random splitting and cross validation approach. This research consists of five stages: data collection, pre-processing, feature extraction, training & testing, and evaluation. Based on the experimental results, the TF-IDF feature is better than the Count-Vectorizer (CV) for Indonesian text. The best accuracy results are obtained by using TF-IDF as a feature and Support Vector Machine (SVM) as a classifier with cross validation implementation. The best accuracy reaches 81%. From the experimental results, it can also be seen that the implementation of cross validation can improve accuracy compared to the implementation of random splitting.**

Keywords*—Random splitting; cross validation; machine learning; Indonesian text

I. Introduction

Opinion mining technology examines and interprets enormous amounts of text data automatically. Opinion mining is the technique of extracting useful information and knowledge from unstructured natural language texts automatically [1]–[5]. Opinion mining is a specific sub-field of text mining that seeks to automatically discover the polarity of opinions (positive, negative, and others) associated with natural language texts [6]–[11].

The language structure of the dataset determines the main challenge in opinion mining. Sentences can be ironic or have several meanings depending on the context. For example, someone can support school policies in the education sector while also breaking school rules—another challenge in obtaining opinions in determining the difference between subjective and objective texts [12]–[14]. A subjective texts is one person's point of view, bias, or opinion. News stories and neutral texts are examples of objective writing that deal with facts and are supposed to be fully unbiased [6], [15], [16].

A machine learning approach can be used for opinion mining. Machine learning refers to methods and systems that can learn from data automatically. The most common machine learning method is supervised learning. It entails creating a prediction model that can inductively learn from a training data collection [17]. The training data is a set of labelled instances, with each example consisting of a pair of input objects (specified in a feature set) and the desired output value, in the case of a classification model, a class label. After the model has been trained, it is ready to be applied to new data [6].

Opinion mining in Indonesia has been a popular topic of study, yet there are still many open challenges. Indonesia is morphologically diverse and ambiguous, with complicated morpho-syntactic rules and many irregular forms and a wide range of dialects with no written standards. Learning a robust general model from Indonesian text might be challenging without suitable processing and handling [18]. Furthermore, compared to English, Indonesian opinion mining has fewer freely available resources in terms of huge net sentiment lexicons and annotated opinion sets. These difficulties have piqued researchers' interest in Indonesian opinion mining [19].

Apart from increasing research on opinion mining on Indonesian text data, there are still some gaps. Most of the previous research has focused on machine learning models and algorithms. There is a need to compare the effect of random splitting and cross-validation on improving performance for processing Indonesian text data [19], [20].

Based on the above background, the purpose of this study is to conduct opinion mining on Indonesian text data using a machine learning model by implementing a random splitting and cross validation approach.

II. Related Work

The previous research of opinion mining on Indonesian text data using a machine learning model has been done by several scholars. Research by Fachrina & Widyantoro (2017) compares Support Vector Machine and Naïve Bayes Classifier to process 2960 Indonesian text data [21]. Research by Suciati & Budi (2019) compared the performance of Random Forest (RF), Multinomial Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Extra Tree classifier (ET) using ten folds cross-validation to process Indonesian text data.

The algorithm that achieved the highest score was obtained by Logistic Regression (LR) and Decision Tree (DT) [22]. Research by Miranda et al. (2019) used Bayes classification to process Indonesian text data. This study obtained an accuracy of 74.94% [23]. Research by Wisnu et al. (2020) uses the Naïve Bayes classifier and K-Nearest Neighbor or KNN to process Indonesian text data. This study obtained an accuracy of KNN (91.00%) and Naïve Bayes (83.50%) [24]. Research by Buntoro et al. (2021) uses the Naïve Bayes Classifier (NBC) and a Support Vector Machine (SVM) to process Indonesian text data. This study obtained an accuracy of SVM (79.02%) and NBC (44.94%) [25].

Most of the previous research has focused on machine learning models and algorithms. There is a need to compare the effect of random splitting and cross-validation on improving performance for processing Indonesian text data. This research is proposed to fill the gap by implementing random splitting and cross-validation for improving performance for processing Indonesian text data.

III. Research Methodology

This study uses a public dataset to determine negative and positive comments on the Indonesian feedback dataset. The stages of the research can be seen in the following Fig. 1.

The first stage is the data collection stage. The dataset used for training and testing the model is sentiment data on Twitter obtained publicly provided by Sulistya in 2021 at Kaggle [26]. The dataset is a collection of feedback data in Indonesian by Twitter users on Covid-19. The dataset consists of 1000 records with 500 records each for the positive class and 500 tweets for the negative class. The following is an example of a dataset. The second stage is the preprocessing stage. This stage consists of six stages: data cleansing, case folding, tokenizing, stopword, normalization, and stemming. Details of these stages can be seen in the Fig. 2:

Based on Fig. 2 above, at the data cleansing stage, a cleaning process is carried out for words that are not needed in order to reduce the computational burden, such as text containing HTML, links, and scripts. In addition, this stage also removes punctuation marks such as periods (.), commas (,) and other punctuation marks. In this pre-processing process, the case folding method is also applied, namely the process of converting words into lower-case letters.The third stage is tokenization. This method is implemented to transform the text's words into several sequences truncated by spaces or specific characters [23], [27], [28].

The stop word removal method is a method of deleting a word with a unique word from text data such as conjunctions and possessive words. In addition, types of words that are less meaningful will be removed, such as words: I, and, or by using this method. Stop word removal is meant to reduce the burden on system performance because the words taken are considered essential [29]–[31]. The last stage in the pre-processing process is the stemming stage. The method at this stage is done by transforming the words in the text to become essential words.

The third research stage is to perform feature extraction. We compare two text features at this stage, namely Count Vectorizer (CV) and TF-IDF. The Term Frequency-Inverse Document Frequency method, abbreviated as TF-IDF, is the most commonly used word weight calculation method in opinion mining. The method is known for its efficient computation time, ease of implementation, and good results or accuracy values. The method works by calculating the value of Term Frequency (TF) and Inverse Document Frequency (IDF) for each token (word) in the document in the corpus or dataset [32]–[34].

The fourth stage is the Training and Testing Model. Training and testing are done by comparing four classifiers, namely Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machine (SVM)

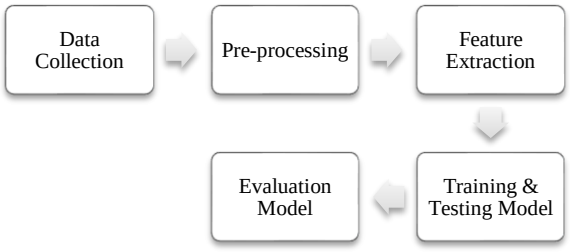

Fig. 1. Main Research Methods.

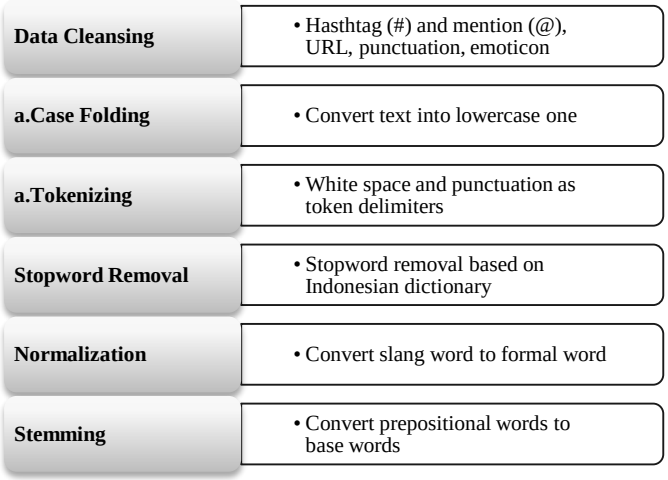

Fig. 2. Pre-Processing Methods.

The Random Forest (RF) method is the development of the Classification and Regression Tree (CART) algorithm. This method applies bootstrap aggregating and random feature selection. This method can be used for classification that works by building a classification tree. Increasing the accuracy of the RF method is done by generating a child node on each node (the node above it) with random selection. Then, the classification results from each tree are accumulated and selected based on the classification results that appear the most [35]–[37]. The RF method has three main parts: the root node, internal node, and leaf node. The root node is the node at the very top, commonly referred to as the root of the decision tree. The internal node is the branch's node with one to two inputs. Finally, the leaf node or terminal node is the end node that has one input and no output. The calculation on the decision tree begins by calculating the entropy value as a determinant of the level of attribute impurity and the value of information gain [35]–[37]. The Logistic Regression (LR) method is used to express the relationship between categorical response variables (in the form of polychotomous or dichotomous) with either continuous or categorical predictor variables. Logistic regression aims to classify each event or observation into positive and negative classes [38] [39].

The Naïve Bayes (NB) method is a method that can be used in opinion mining. This method applies Bayes' theorem theory in classification based on attribute values that are conditionally independent if given an output value. In short, Bayes' theorem is a fundamental statistical approach to pattern recognition [40], [41]. The advantage of using the NB method is that the value or amount of training data required in data processing can be on a small scale and can still be used to determine parameter estimates in the data classification process [42].

The Support Vector Machine (SVM) method is a method with the concept of statistical learning theory, which has given better performance results than other classification methods in several research cases. This method does not study all the training data in the training process, but only a selected number of data is used to build a model in the classification process. The SVM method does not store all training data but only stores a small portion of training data for further processing. This has become an advantage in choosing the SVM method because not all training data is involved in each training iteration [43]. The SVM method works by maximizing the decision limit (hyperplane) or finding the best decision limit (hyperplane) as a separator of the two data classes depicted in the Fig. 3:

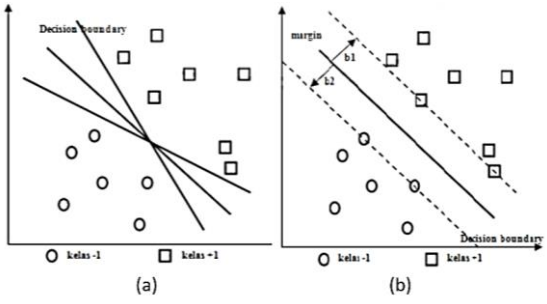

Fig. 3. Hyperplane [44].

In the picture above (a), there is a choice of possible decision limits (hyperplane) for the data set; then, in the picture above (b), there is a decision limit (hyperplane) with the maximum margin. The margin is the distance between the decision boundary (hyperplane) and the closest data from each class. This closest data is known as the support vector. The hyperplane component with the maximum margin will better generalize the classification process.

For example, in Fig 3 (b), the solid line component shows the best decision boundary (hyperplane) with a location in the middle of the two classes, while the dotted line component that passes through the circle and square data is a support vector. The central concept of training on the SVM method is finding the hyperplane's location [44], [45]. Experiments were carried out using the results of random splitting and cross-validation. The third and fourth stages are illustrated in the Fig. 4:

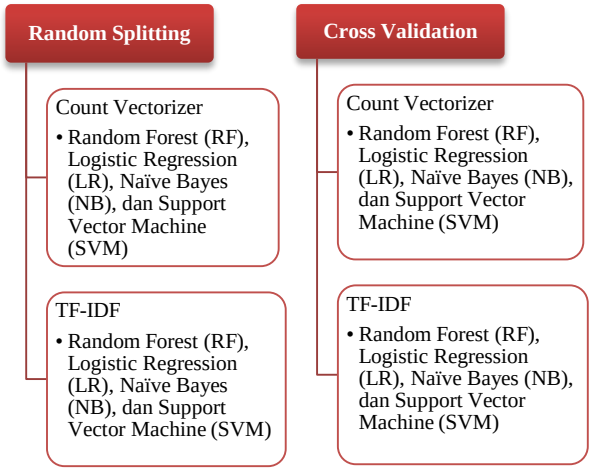

Fig. 4. Experiment Scenario.

The next stage is an evaluation to compare the best accuracy, precision, recall, f1-measure obtained. The performance evaluation measurement model is an approach that aims to measure the performance or performance of a system. This approach is widely used in the case of training or data training. Several formulas or equations in the performance evaluation measure are usually applied separately or in combination to get a better performance analysis perspective. Some of the calculations contained in the performance evaluation are as follows [46]. The precision method calculates the level of accuracy or accuracy of the results between user testing and system answers.

$$pre = \frac{TP}{TP+FP} \quad (1)$$

The recall is a measurement of the accuracy or accuracy of the same information with information that has existed before.

$$rec = \frac{TP}{TP+FN} \quad (2)$$

Accuration is a comparative calculation between the information the system answers correctly with the comprehensive information.

$$acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

IV. Result and Discussion

This study aims to determine negative and positive comments on the Indonesian feedback dataset publicly provided by Sulistya (2021) on the Kaggle web using a machine learning approach. This study also compares the performance of several machine learning methods to find out which method has the best performance.

The second stage carried out in the experiment is the pre-processing stage. The first stage of pre-processing is data cleansing. The cleansing of the dataset is: removing hashtags (#) and mentions (@), deleting URLs, deleting punctuation and deleting emoticons. The example result of data cleansing can be seen in the Table I:

TABLE I. Result of Data Cleansing

Process	Data #1	Data #2
Data source	"Indonesia: APBN Sekarat, Covid-19 Meningkat,	#BREAKING:Pemerintah mengonfirmasi kasus posi
Text_remove _hastag_ and mentions	"Indonesia:APBN Sekarat, Covid-19 Meningkat,	: Pemerintah mengonfirmasi kasus positif Covid
Text_remove _url	"Indonesia:APBN Sekarat, Covid-19 Meningkat,	: Pemerintah mengonfirmasi kasus positif Covid
Text_remove _punc	Indonesia:APBN Sekarat, Covid-19 Meningkat Rakya	Pemerintah mengonfirmasi kasus positif Covid1
Text_remove _emojis	Indonesia:APBN Sekarat, Covid-19 Meningkat Rakya	Pemerintah mengonfirmasi kasus positif Covid1
Text_remove_emoticons	Indonesia:APBN Sekarat, Covid-19 Meningkat Rakya	Pemerintah mengonfirmasi kasus positif Covid1

The second preprocessing stage is case folding. This stage is done by converting text into lowercase ones. For example, "Pemerintah mengkonfirmasi kasus positif COVID19...", will be converted as "pemerintah mengkonfirmasi kasus covid19". The example results of case folding for our dataset can be seen in the Fig. 5:

```
0    bang gimna pemerintah mau peduli rrc urus abk ...
1    erinx  tidak percaya data covid19 dari pemerin...
2    indonesia apbn sekarat covid19 meningkat rakya...
3    \n\nuntuk mengurangi penyebaran virus covid19\...
4      hingga kamis 752020 pk 1200 wibdata pemerint...
5     pemerintah mengonfirmasi kasus positif covid1...
```

Fig. 5. Example Result of Case Folding.

The next stage is tokenizing. This stage separates the text with white space and punctuation as token delimiters. For example, "pemerintah mengkonfirmasi kasus covid19", will be converted as "[pemerintah, mengkonformasi, kasus, positif, covid19]". The example result of tokenizing for our dataset can be seen in the Fig. 6:

```
0    [bang, gimna, pemerintah, mau, peduli, rrc, ur...
1    [erinx, tidak, percaya, data, covid, dari, pem...
2    [indonesia, apbn, sekarat, covid, meningkat, r...
3    [untuk, mengurangi, penyebaran, virus, covid, ...
4    [hingga, kamis, pk, wibdata, pemerintah, mempe...
5    [pemerintah, mengonfirmasi, kasus, positif, co...
```

Fig. 6. Example Results of Tokenizing.

The next step is stop-word removal. At this stage, the words included in the stop-word will be deleted. Stop-word deletion is done by matching the dataset with the Indonesian stop-word dictionary. For example, "[untuk, mengurangi, penyebaran virus, covid]", will be converted "[mengurangi, penyebaran virus, covid]". The word "untuk" stop-word so as it will be deleted. The example result of stop-word removal can be seen in the Fig. 7:

case_folding_tweets	tweet_tokens	tweet_stopword_removal
bang gimna pemerintah mau peduli rrc urus abk ...	[bang, gimna, pemerintah, mau, peduli, rrc, ur...	[bang, gimna, pemerintah, peduli, rrc, urus, a...
erinx tidak percaya data covid dari pemerintah...	[erinx, tidak, percaya, data, covid, dari, pem...	[erinx, percaya, data, covid, pemerintah, perc...
indonesia apbn sekarat covid meningkat rakyat ...	[indonesia, apbn, sekarat, covid, meningkat, r...	[indonesia, apbn, sekarat, covid, meningkat, r...
untuk mengurangi penyebaran virus covid menduk...	[untuk, mengurangi, penyebaran, virus, covid, ...	[mengurangi, penyebaran, virus, covid, menduku...

Fig. 7. Example Results of Stop-Removal.

The fifth preprocessing stage is normalization. Normalization changes non-standard words (slang words) and acronyms into familiar words by matching the dataset with the Indonesian normalization dictionary. The results of some normalization of words in Indonesian are shown in the Table II and Fig. 8:

TABLE II. Result of Normalization

No.	Real Data	Normalization
1	&	Dan
2	1 pun	Satupun
3	7an	Tujuan
4	@	Di
5	Jkt	Jakarta
6	Nasihat	Nasehat
7	SE	Surat edaran
8	Ababil	Abglabil
9	Abis	Habis
10	Ad	Ada

tweet_tokens	tweet_stopword_removal	tweet_normalized
[bang, gimna, pemerintah, mau, peduli, rrc, ur...	[bang, gimna, pemerintah, peduli, rrc, urus, a...	[bang, gimana, pemerintah, peduli, rrc, urus, ...
[erinx, tidak, percaya, data, covid, dari, pem...	[erinx, percaya, data, covid, pemerintah, perc...	[erinx, percaya, data, covid, pemerintah, perc...
[indonesia, apbn, sekarat, covid, meningkat, r...	[indonesia, apbn, sekarat, covid, meningkat, r...	[indonesia, apbn, sekarat, covid, meningkat, r...

Fig. 8. Example Results of Normalization.

The sixth preprocessing stage is stemming. At this stage, the affixed words are transformed into basic words using the literary method. The method at this stage is done by transforming the words in the text to become essential words. At this stage, the transformation of affixed words into basic words is carried out using the Sastrawi library. For example, "[mengurangi, penyebaran virus, covid]", will be converted "[kurang, sebar virus, covid]. The basic words of "mengurangi" and "menyebarkan" are "kurang" and "sebar. The example result of stemming can be seen in the Fig. 9:

After completing six stages: data cleansing, case folding, tokenization, stopword, normalization, and stemming. The example of every step has been elaborated above. Attribute of Data#1 is text before method is applied and attribute of Data#2 is text after applied method. Moreover, an overview result of preprocessing stages can be seen in the Table III.

The third research stage is to perform feature extraction. At this stage, feature extraction is carried out on the dataset that has been processed in the previous stage. The feature extraction stage aims to obtain features used in model training and testing. We compare two text features at this stage, namely Count Vectorizer (CV) and TF-IDF.

The fourth stage is the training and testing model. Training and testing are done by comparing four classifiers, namely Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machine (SVM). Experiments were carried out using the approach of random splitting and cross-validation.

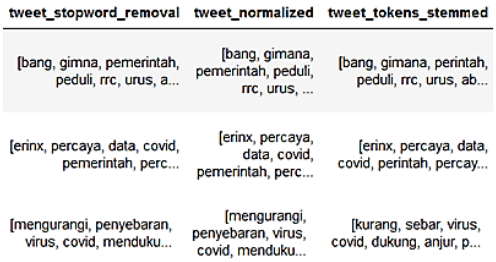

Fig. 9. Example Results of Stemming.

TABLE III. RESULT OF DATA PREPROCESSING

Method	Data #1	Data #2
Cleansing _tweets	Indonesia APBN Sekarat Covid Meningkat Rakya…	\n\Untuk Mengurangi Penyebaran Virus Covid19\ …
case_folding _tweets	indonesia apbn sekarat covid meningkat rakyat…	untuk mengurangi penyebaran virus covid menduk…
tweet_tokens	[Indonesia, apbn, sekarat, covid, meningkat r…	[untuk,mengurangi, penyebaran, virus, covid, …
tweet _stopword _removal	[Indonesia, apbn, sekarat, covid, meningkat, r…	[mengurangi, penyebaran, virus, covid, menduku…
tweet _normalized	[Indonesia, apbn, sekarat, covid, meningkat, r…	[mengurangi, penyebaran, virus, covid, menduku…
tweet_tokens _stemmed	[Indonesia, apbn, sekarat, covid, tingkat, rak…	[kurang, sebar, virus, covid, dukung, anjur, p…

The first experiment is training and evaluating machine learning models using random splitting. This experiment was carried out by randomly dividing the dataset into training and testing data with 80% of the training data and 20% of the testing data, respectively. Furthermore, an evaluation was carried out to compare the best accuracy, precision, recall, f1-measure obtained. The results of this experiment can be seen in the Table IV:

TABLE IV. PERFORMANCE EVALUATION

Process	Accuracy	Precision	Recall	F1
CV&RF	0.76	0.77	0.76	0.76
TF-IDF&RF	0.77	0.77	0.77	0.77
CV&LR	0.75	0.75	0.75	0.75
TF-IDF&LR	0.76	0.76	0.76	0.76
CV&NB	0.75	0.75	0.74	0.75
TF-IDF&NB	0.74	0.75	0.74	0.74
CV & SVM	0.76	0.77	0.76	0.76
TF-IDF&SVM	0.78	0.78	0.78	0.78

The second experiment is training and evaluation of machine learning models using cross-validation. This stage is carried out using 10-fold cross-validation. At this stage, cross-validation is implemented to find the maximum accuracy of the model. After cross-validation, training and model evaluation were carried out to measure the resulting accuracy, precision, recall, f1-measure results. The results of this experiment can be seen in the Table 5:

TABLE V. EXPERIMENT WITH CROSS VALIDATION

Process	Accuracy	Precision	Recall	F1
CV&RF	0.79	0.80	0.79	0.79
TF-IDF&RF	0.78	0.79	0.79	0.79
CV&LR	0.78	0.79	0.79	0.79
TF-IDF&LR	**0.81**	0.81	0.81	0.81
CV&NB	0.79	0.80	0.79	0.79
TF-IDF&NB	0.79	0.80	0.79	0.79
CV & SVM	0.79	0.79	0.79	0.79
TF-IDF&SVM	**0.81**	0.81	0.81	0.81

According to the Table above, utilizing TF-IDF as a feature and Support Vector Machine (SVM) as a classifier with cross validation implementation yields the best accuracy results. The highest level of accuracy is 81%. It can also be observed from the experimental results that cross validation can improve accuracy when compared to random splitting. Furthermore, for Indonesian language, the TF-IDF feature outperforms the Count-Vectorizer (CV).

V. CONCLUSION

Based on the experimental results, the TF-IDF feature is better than the Count-Vectorizer (CV) for Indonesian text. The best accuracy results are obtained by using TF-IDF as a feature and Support Vector Machine (SVM) as a classifier with cross validation implementation. The best accuracy reaches 81%. From the experimental results, it can also be seen that the implementation of cross validation can improve accuracy compared to the implementation of random splitting.

This research has not discussed the problem of negation. In future research, issues that will be investigated further include the implementation of negation handling with the modified

syntactic rule method in the pre-processing process to increase the accuracy of opinion mining.

ACKNOWLEDGMENT

The authors would like to thank Universitas Sriwijaya that have supported this research.

REFERENCES

[1] M. Misuraca, G. Scepi, and M. Spano, "Using Opinion Mining as an educational analytic: An integrated strategy for the analysis of students' feedback," Stud. Educ. Eval., vol. 68, p. 100979, 2021.

[2] R. Annisa and I. Surjandari, "Opinion mining on Mandalika hotel reviews using latent dirichlet allocation," Procedia Comput. Sci., vol. 161, pp. 739–746, 2019.

[3] E. Sonalitha et al., "Combined Text Mining: Fuzzy Clustering for Opinion Mining on the Traditional Culture Arts Work," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 8, pp. 294–299, 2020.

[4] D. Ramayanti et al., "Tuberculosis Ontology Generation and Enrichment Based Text Mining," in 2020 International Conference on Information Technology Systems and Innovation (ICITSI), 2020, pp. 429–434.

[5] H. Noprisson et al., "Influencing factors of knowledge sharing among students in Indonesia higher educational institutions," in 2016 International Conference on Information Technology Systems and Innovation (ICITSI), 2016, pp. 1–6.

[6] L. Tavoschi et al., "Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy," Hum. Vaccin. Immunother., vol. 16, no. 5, pp. 1062–1069, 2020.

[7] P. Rajkumar, "Opinion mining for user experience evaluation model using kernel-naive bayes classification algorithm," J. Eng. Res., 2021.

[8] C. J. Rameshbhai and J. Paulose, "Opinion mining on newspaper headlines using SVM and NLP," Int. J. Electr. Comput. Eng., vol. 9, no. 3, p. 2152, Jun. 2019.

[9] H. Noprisson, E. Ermatita, A. Abdiansah, V. Ayumi, M. Purba, and M. Utami, "Hand-Woven Fabric Motif Recognition Methods: A Systematic Literature Review," in 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS, 2021, pp. 90–95.

[10] V. Ayumi, E. Ermatita, A. Abdiansah, H. Noprisson, M. Purba, and M. Utami, "A Study on Medicinal Plant Leaf Recognition Using Artificial Intelligence," in 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS, 2021, pp. 40–45.

[11] M. Purba, E. Ermatita, A. Abdiansah, V. Ayumi, H. Noprisson, and A. Ratnasari, "A Systematic Literature Review of Knowledge Sharing Practices in Academic Institutions," in 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS, 2021, pp. 337–342.

[12] Y. Huang, "Opinion Mining Algorithm Based on the Evaluation of Online Mathematics Course with Python," in 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2021, pp. 1395–1398.

[13] J. Serrano-Guerrero, F. P. Romero, and J. A. Olivas, "Fuzzy logic applied to opinion mining: a review," Knowledge-Based Syst., vol. 222, p. 107018, 2021.

[14] S. Zad, M. Heidari, J. H. Jones, and O. Uzuner, "A survey on concept-level sentiment analysis techniques of textual data," in 2021 IEEE World AI IoT Congress (AIIoT), 2021, pp. 285–291.

[15] A. Easwaran, "Opinion Mining and Emotion Detection in Social Network Data and Student Survey Data in Cloud Environment." The University of North Carolina at Charlotte, 2021.

[16] S. Sagnika, B. S. P. Mishra, and S. K. Meher, "An attention-based CNN-LSTM model for subjectivity detection in opinion-mining," Neural Comput. Appl., vol. 33, no. 24, pp. 17425–17438, 2021.

[17] [17] H.-C. Soong, N. B. A. Jalil, R. Kumar Ayyasamy, and R. Akbar, "The Essential of Sentiment Analysis and Opinion Mining in Social Media : Introduction and Survey of the Recent Approaches and Techniques," in 2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE), 2019, pp. 272–277.

[18] W. P. Sari and H. Fahmi, "Opinion Mining Analysis on Online Product Reviews Using Naïve Bayes and Feature Selection," in 2021 International Conference on Information Management and Technology (ICIMTech), 2021, vol. 1, pp. 256–260.

[19] G. Badaro et al., "A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations," ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 18, no. 3, pp. 1–52, 2019.

[20] S. H. Sahir, R. S. Ayu Ramadhana, M. F. Romadhon Marpaung, S. R. Munthe, and R. Watrianthos, "Online learning sentiment analysis during the covid-19 Indonesia pandemic using twitter data," IOP Conf. Ser. Mater. Sci. Eng., vol. 1156, no. 1, p. 012011, Jun. 2021.

[21] Z. Fachrina and D. H. Widyantoro, "Aspect-sentiment classification in opinion mining using the combination of rule-based and machine learning," in 2017 International Conference on Data and Software Engineering (ICoDSE), 2017, pp. 1–6.

[22] A. Suciati and I. Budi, "Aspect-based Opinion Mining for Code-Mixed Restaurant Reviews in Indonesia," in 2019 International Conference on Asian Language Processing (IALP), 2019, pp. 59–64.

[23] E. Miranda, M. Aryuni, R. Hariyanto, and E. S. Surya, "Sentiment Analysis using Sentiwordnet and Machine Learning Approach (Indonesia general election opinion from the twitter content)," in 2019 International Conference on Information Management and Technology (ICIMTech), 2019, vol. 1, pp. 62–67.

[24] H. Wisnu, M. Afif, and Y. Ruldevyani, "Sentiment analysis on customer satisfaction of digital payment in Indonesia: A comparative study using KNN and Naïve Bayes," J. Phys. Conf. Ser., vol. 1444, no. 1, p. 012034, Jan. 2020.

[25] G. A. Buntoro, R. Arifin, G. N. Syaifuddiin, A. Selamat, O. Krejcar, and F. Hamido, "The Implementation of the machine learning algorithm for the sentiment analysis of Indonesia's 2019 Presidential election," IIUM Eng. J., vol. 22, no. 1, pp. 78–92, 2021.

[26] Y. I. Sulistya, "Covid-19 Indonesian Tweet," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/yudhaislamisulistya/covid19-tweet-indonesia-positif-dan-negatif/versions/5. [Accessed: 01-Feb-2022].

[27] B. Haryanto, Y. Ruldeviyani, F. Rohman, J. D. TN, R. Magdalena, and Y. F. Muhamad, "Facebook analysis of community sentiment on 2019 Indonesian presidential candidates from Facebook opinion data," Procedia Comput. Sci., vol. 161, pp. 715–722, 2019.

[28] Y. D. Kirana and S. Al Faraby, "Sentiment Analysis of Beauty Product Reviews Using the K-Nearest Neighbor (KNN) and TF-IDF Methods with Chi-Square Feature Selection," J. Data Sci. Its Appl., vol. 4, no. 1, pp. 31–42, 2021.

[29] P. Desai, J. R. Saini, and P. B. Bafna, "POS-based Classification and Derivation of Kannada Stop-words using English Parallel Corpus," in 2022 3rd International Conference for Emerging Technology (INCET), 2022, pp. 1–5.

[30] R. Hendrawan and S. Al Faraby, "Multilabel classification of hate speech and abusive words on Indonesian Twitter social media," in 2020 International Conference on Data Science and Its Applications (ICoDSA), 2020, pp. 1–7.

[31] R. Rosnelly, "The Similarity of Essay Examination Results using Preprocessing Text Mining with Cosine Similarity and Nazief-Adriani Algorithms," Turkish J. Comput. Math. Educ., vol. 12, no. 3, pp. 1415–1422, 2021.

[32] A. Fauzi, E. B. Setiawan, and Z. K. A. Baizal, "Hoax News Detection on Twitter using Term Frequency Inverse Document Frequency and Support Vector Machine Method," J. Phys. Conf. Ser., vol. 1192, no. 1, p. 012025, Mar. 2019.

[33] M. Z. Naeem, F. Rustam, A. Mehmood, I. Ashraf, and G. S. Choi, "Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms," PeerJ Comput. Sci., vol. 8, p. e914, 2022.

[34] Z. Balani and C. Varol, "Combining Approximate String Matching Algorithms and Term Frequency In The Detection of Plagiarism.," Int. J. Comput. Sci. Secur., vol. 15, no. 4, pp. 97–106, 2021.

[35] G. A. Sandag, "Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest," CogITo Smart J., vol. 6, no. 2, pp. 167–178, 2020.

[36] P. Chakraborty, F. Nawar, and H. A. Chowdhury, "A Ternary Sentiment Classification of Bangla Text Data using Support Vector Machine and Random Forest Classifier," in International Conference on Computational Techniques and Applications, 2022, pp. 69–77.

[37] H. M. Pandey, P. Tiwari, A. Khamparia, and S. Kumar, "Twitter-based opinion mining for flight service utilizing machine learning," Inform., vol. 43, no. 3, pp. 381–386, 2019.

[38] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, "Supervised data mining approach for predicting student performance," Indones. J. Electr. Eng. Comput. Sci., vol. 16, no. 3, p. 1584, Dec. 2019.

[39] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," J. Clin. Epidemiol., vol. 110, pp. 12–22, 2019.

[40] K. S. Rawat and I. V Malhan, "A hybrid classification method based on machine learning classifiers to predict performance in educational data mining," in Proceedings of 2nd International Conference on Communication, Computing and Networking, 2019, pp. 677–684.

[41] S.-S. M. Ajibade, N. B. Ahmad, and S. M. Shamsuddin, "A data mining approach to predict academic performance of students using ensemble techniques," in International Conference on Intelligent Systems Design and Applications, 2018, pp. 749–760.

[42] B. Khotimah, M. Miswanto, and H. Suprajitno, "Optimization of Feature Selection Using Genetic Algorithm in Naïve Bayes Classification for Incomplete Data," Int. J. Intell. Eng. Syst., vol. 13, no. 1, pp. 334–343, Feb. 2020.

[43] Y. Liu, H. Chen, L. Zhang, X. Wu, and X. Wang, "Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: A case study in China," J. Clean. Prod., vol. 272, p. 122542, 2020.

[44] A. A. Putra, R. Magdalena, and R. Y. N. Fu'adah, "Klasifikasi Kanker Usus Besar Menggunakan Metode Ekstraksi Ciri Principal Component Analysis Dan Klasifikasi Support Vector Machine," eProceedings Eng., vol. 6, no. 2, 2019.

[45] P. Birzhandi, K. T. Kim, B. Lee, and H. Y. Youn, "Reduction of training data using parallel hyperplane for support vector machine," Appl. Artif. Intell., vol. 33, no. 6, pp. 497–516, 2019.

[46] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "Multi-label classifier performance evaluation with confusion matrix," Comput Sci Inf Technol, vol. 10, pp. 1–14, 2020.