JURNAL INFOTEL
Informatics - Telecommunication - Electronics
Website: https://ejournal.ittelkom-pwt.ac.id/index.php/infotel
ISSN: 2085-3688; e-ISSN: 2460-0997

Data preprocessing approach for machine learning-based sentiment classification

Sunneng Sandino Berutu[1,*], Haeni Budiati[2], Jatmika[3], Fornieli Gulo[4]
[1,2,3,4]Department of Informatic, Immanuel Christian University Yogyakarta
[1,2,3,4]Jl. Solo, Km. 11,1, Yogyakarta 55571, Indonesia
[*]Corresponding email: sandinoberutu@ukrimuniversity.ac.id

Received 27 August 2023, Revised 17 October 2023, Accepted 1 November 2023

Abstract — Public sentiment regarding a particular issue, product, activity, or organization can be measured and monitored with an application based on artificial intelligence. The data come from comments circulating on social media. However, the rules for writing comments on social media have yet to be standardized, so non-standard words often appear in these comments. Non-standard words affect the determination of sentiment into positive, neutral, and negative categories. Therefore, the author proposes a data preprocessing approach by inserting the Rabin-Karp algorithm to improve non-standard words. This research consists of several steps, namely crawling of data, preprocessing of data, extraction of features, development of the model (based on Naïve Bayes, Decision Tree, and Support Vector Machine methods), and the analysis of the results. The implementation results indicated that the approach influences the determination of the sentiment category composition. Then, model testing results showed that almost all models obtained the highest value in the Neutral category for the precision, recall, and f1-score parameters. In addition, the results of the accuracy parameter of the classification model showed that the Support Vector Machine-based model performs better than the Naïve Bayes and Decision Tree methods.

Keywords – data preprocessing, decision tree, Naïve Bayes, Rabin-Karp, sentiment, support vector machine,

Copyright ©2023 JURNAL INFOTEL
All rights reserved.

I. Introduction

Social media has become the most popular platform for user interaction and information sharing in the modern digital age. Social media is also a significant data source for understanding public opinion and sentiment about a particular product, brand, or topic in business and marketing. Therefore, an artificial intelligence-based sentiment identification application was developed to support this. The data sources used to develop this system are Instagram, Twitter, and Facebook [1]. The data obtained can be in the form of text, images, and videos. In text data, there are non-standard words and spelling mistakes that can affect the performance of the sentiment classification system. Data preprocessing is an essential part of developing a classification model. In general, the steps of data preprocessing that are completed are text cleaning [2], deleting stopwords [3], deleting non-letter characters, and converting uppercase to lowercase [4].

Research related to machine learning-based sentiment classification has been implemented in various problems and fields. Some examples of implementation in the health sector related to the prevention, treatment, and impact of COVID-19 include market sentiment [5] and the use of multidimensional data in determining sentiment [6] with the BERT framework. BERT is also adopted to measure sentiment on product review data [7], and Twitter comments data about finance [8]. Another research presents the development of a classification of opinion tweet data using a combination method based on Convolutional Neural Networks [9], a performance comparison of models using Decision Tree (DT), Support Vector Machine (SVM), and Random Forest (RF) for sentiment classification in the tourism sector [10] and the health sector in the COVID-19 dataset [11]. Other researchers [12] used five methods, such as Naïve Bayes (NB), DT, K-Nearest Neighbor, SVM, and RF, to evaluate sentiment data related to COVID-19. Other methods, such as Neural Networks (NN), are applied to data science [13], product recommendations [14], and sentiment analysis in the financial sector [15].

The Rabin-Karp algorithm adopts a hashing tech-

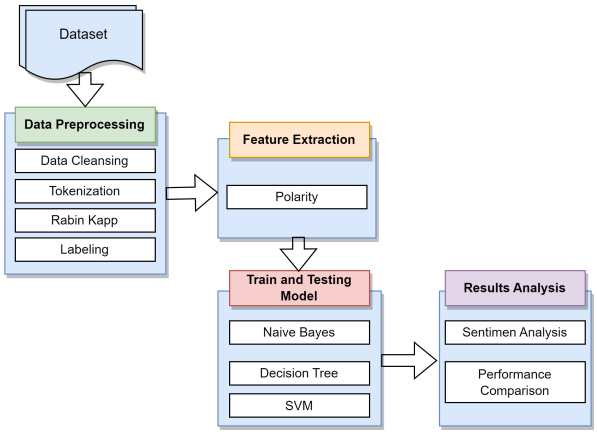

Fig. 1. Research method.

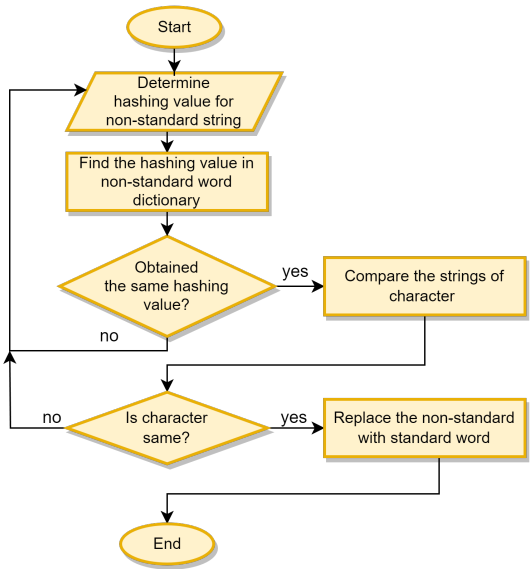

Fig. 2. Rabin-Karp algorithm for repairing the non-standard words.

nique to find similar words. This algorithm is implemented effectively for multiword identification [16] in text. Previous researchers have adopted this approach to solve various problems such as potential plagiarism [16], detection of the level of document similarity [17], and detection of fake news [18]. This study proposes a data preprocessing approach to detect and improve non-standard words. In this approach, the workings of the Rabin-Karp algorithm are adopted to refine non-standard words into standard words. Then, the dataset obtained by the abovementioned approach is used to develop machine learning-based classification models such as NB, SVM, and DT.

II. Research Method

This section explains the steps in optimizing data preprocessing to improve sentiment classification performance. The stages proposed in this study consist of five steps, namely data acquisition, preprocessing, extracting features, model training and testing, and results analysis. These steps are illustrated in Fig. 1, and explained in further detail in the following subsection.

A. Dataset

The initial stage of this study involved collecting data on comments in Indonesian from social media platform Twitter. The sentiment comments that will be processed are the comments regarding one of Indonesia's online transportation services. The snscrape library is used to collect Twitter data. The keyword used in the data search is "gojek". The data collection period was from July 01, 2022, to October 30, 2023. The total number of tweets obtained was 3178. The data is then saved in CSV format.

B. Data Preprocessing

Generally, the steps of data preprocessing that are completed are text cleaning [2], deleting stopwords [3], deleting non-letter characters [4], and converting uppercase to lowercase. Furthermore, tokenization is carried out to separate words from text so that the data obtained consists of several words. The next stage is identifying non-standard words from the text. The Rabin-Karp algorithm is implemented to fix the non-standard word, as illustrated in Fig. 2. In order to facilitate the implementation of this algorithm, an initial non-conventional word dictionary is created, consisting of a total of 500 data words.

In general, the workings of the Rabin-Karp algorithm in this study are described as follows:

1) Calculating hash value of non-standard words.
2) Looking up the first stage of the hash value in the dictionary.
3) Comparing word characters.
4) Replacing non-standard words with standard words.

C. Feature Extraction

Tweet sentiments were grouped into three categories, namely Positive, Neutral, and Negative. These categories are obtained by calculating the text polarity value [18] with the following rules:

1) If the polarity value > 1, the text was indentified as Positive opinion category.
2) If the polarity value $= 0$, the text was indentified as the Neutral opinion category.
3) If the polarity value < 1, the text was classified into the category of Negative opinion.

D. Implementation Machine Learning Model

The sentiment classification model was developed using NB, DT, and SVM methods. The author used the scikit-learn library for model training and testing. The dataset processed at the feature extraction stage was split into data from training and testing. The percentage of dataset composition was determined at 80 % as data from training, and the rest was data from testing.

The results of the testing were measured by the confusion matrix. Model performance was assessed

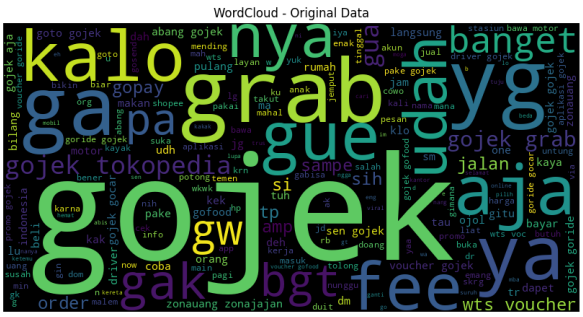

Fig. 3. The original dataset word cloud.

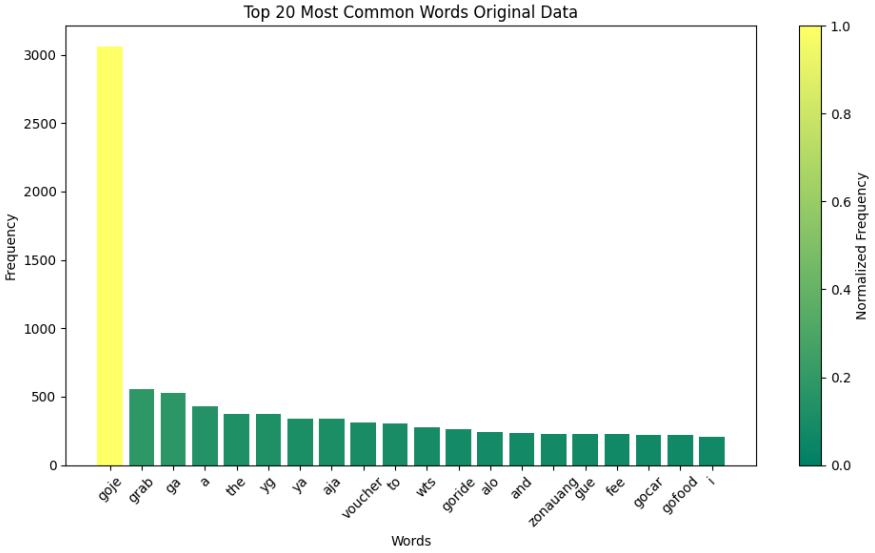

Fig. 4. Word frequency of original dataset.

by four parameter indicators: accuracy, f1-score, precision, and sensitivity. Then, the authors conducted analyses to measure the RK algorithm implementation as follows:

1) The results of sentiment measurement on the original dataset.
2) The results of sentiment measurement on datasets processed with the Rabin-Karp algorithm.
3) The classification model performance using a confusion matrix.

III. Result

In this section, we delve into the results of the experiment. First, we explore the outcomes of implementing the proposed approach, examining its impact on sentiment category composition. Subsequently, we discuss the results related to sentiment measurement, shedding light on the effectiveness of the applied methods. Finally, we analyze the performance of the classification models, with a focus on the comparative evaluation of the SVM-based model against the NB and DT methods.

A. Proposed Approach Implementation

The results of crawling on Twitter obtained data of 3178 tweets. The data preprocessing stage is done by data cleaning (removing users, removing emoticons, and removing stop words), case folding, and text stemming. The result of this stage is displayed in the word cloud shown in Fig. 3, while the frequency of word occurrences is shown in the following graph of Fig. 4. Fig. 3 shows that the word ”gojek” has the largest font size. It signifies that the word ”gojek” has the most significant number in the dataset. Fig. 4 shows a graph of the list of common words in the dataset, in which the word ”gojek” is in the first position.

The next stage is the identification of non-standard words. The identification result is shown as a word cloud in Fig. 5, while the frequency of word occurrences is visualized in the bar chart in Fig. 6. Based on Fig. 5, the words “yg” and “ga” have almost the same font size. It signifies that these three words dominate the dataset. Fig. 6 shows the frequency of the words ”yg” and ”ga,” which are in the top three orders of the dataset.

Meanwhile, the word cloud and the bar chart of the frequency of fixed words conducted by the Rabin-Karp algorithm are shown in Fig. 7 and Fig. 8. Based on Fig. 7 and 8, the word ”gojek” is the word that appears most frequently.

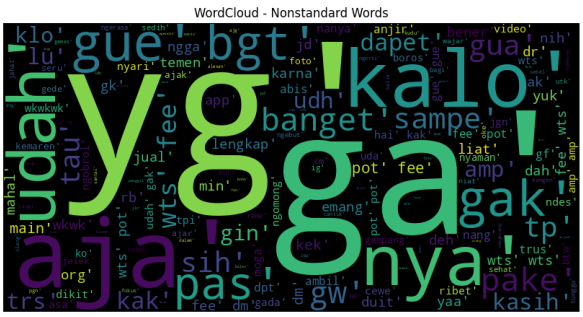

Fig. 5. The word cloud of non-standard words.

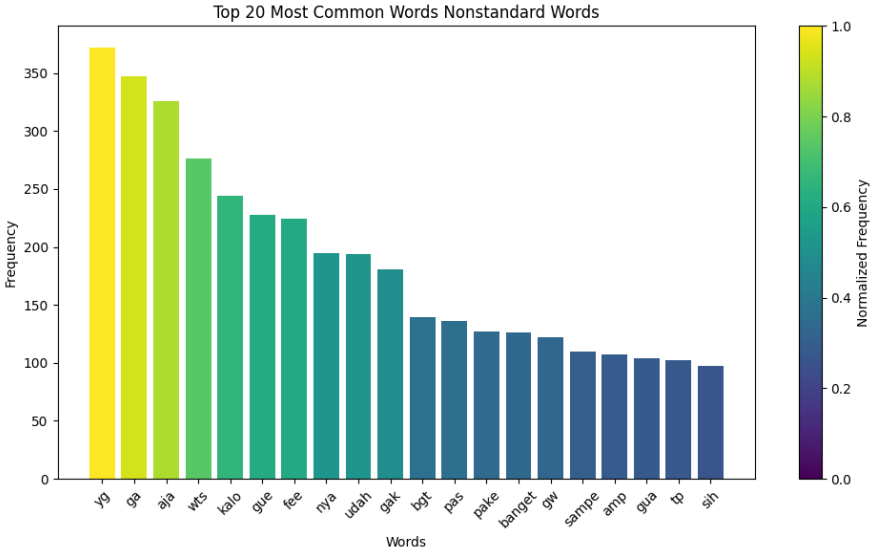

Fig. 6. The frequency of identified non-standard words.

B. Sentiment Measurement

The implementation of the polarity technique in determining sentiment categories with the original dataset is described in Fig. 9 and Table 1. The dataset is named ds_original, which is used in model development. Sentiment in the Neutral category reached the highest percentage of 81.7 %. Meanwhile, the Negative category received the lowest percentage.

Table 1. Sentiment Results on the ds_original Dataset

Sentiment categories	Quantity	Percentage
Positive	398	12,5
Negative	184	5,8
Neutral	2596	81,7

Furthermore, the results of sentiment measurement on the word repairing results are shown in Fig. 10 and Table 2. This dataset is named ds_RabinKarp, which is used in training and model testing. Sentiment in the Neutral category reached the highest percentage, namely 80.1 %. Meanwhile, the Negative category received the smallest percentage.

Table 2. Sentiment Results on the ds_RabinKarp Dataset

Text sentiments	Quantity	Percentage
Positive	426	13,4
Negative	208	6,5
Neutral	2544	80,1

C. Classification Models

The datasets in Table 1 and Table 2 are used as training and test datasets in developing a classification model based on the NB, DT, and SVM methods. The model is adjusted to the name of the dataset. Therefore, there are six classification models. Model names are explained in Table 3.

Table 3. List of Developed Models

Metode	Dataset	Model
Naïve Bayes	*ds_original*	*NBO_model*
Naïve Bayes	*ds_RabinKarp*	*NBRK_model*
Decision Tree	*ds_original*	*DTO_model*
Decision Tree	*ds_RabinKarp*	*DTRK_model*
Support Vector Machine	*ds_original*	*SVMO_model*
Support Vector Machine	*ds_RabinKarp*	*SVMRK_model*

After completing the training phase, the authors tested the model, visualizing the results in a confusion matrix [19] and presenting performance parameter values in a table. Two models were generated using the NB method: NBO_model and NBRK_model. The confusion matrix for NBO_model test results is depicted in Fig. 11. The obtained parameter values, including accuracy, f1-score, recall, and precision [20], are detailed in Table 4.

The highest value of the precision parameter is achieved by the Neutral category of 0.83. Then, the highest f1-score and recall values are also in the

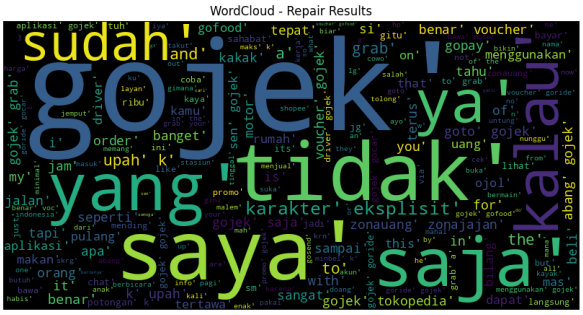

Fig. 7. The word cloud of word repair.

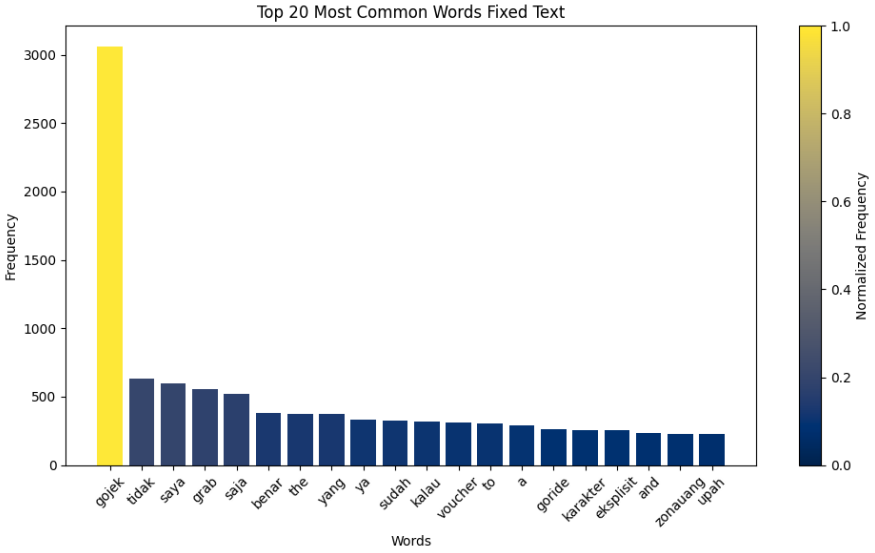

Fig. 8. The frequency in the fixed dataset.

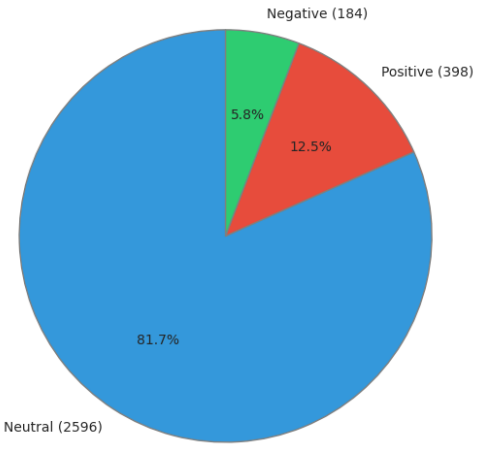

Fig. 9. The percentage of sentiment in the original dataset.

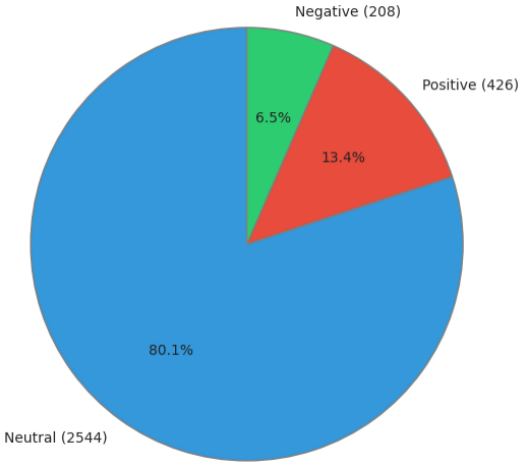

Fig. 10. The percentage of sentiment in the repaired dataset.

Table 4. Parameter of the NBO_model Test Results

Text sentiment	Parameters			Support
	Precision	Recall	f1-score	
Positive	0.61	0.20	0.30	35
Negative	0.00	0.00	0.00	506
Neutral	0.83	0.99	0.91	95

Table 5. Parameters of the NBRK_model Test Result

Text sentiment	Parameters			Support
	Precision	Recall	f1-score	
Positive	0.64	0.14	0.23	100
Negative	0.00	0.00	0.00	45
Neutral	0.80	1	0.89	491

Neutral category of 0.91 and 0.99. The prediction of accuracy is 0.82. The confusion matrix from the NBRK_model test results is shown in Fig. 12. Table 5 shows the resulting parameters from the confusion matrix of Fig. 12.

The Neutral category has the highest score on the precision, f1-score, and recall parameters of 0.80, 0.89, and 1. The result of the DTO_model test is displayed as a confusion matrix in Fig. 13. Parameters resulting from the confusion matrix Fig. 13 are described in

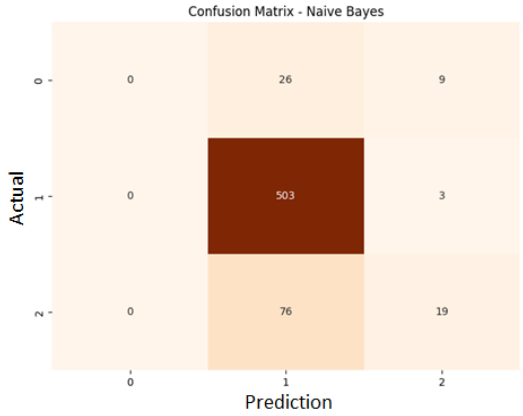

Fig. 11. The confusion matrix from NBO_model.

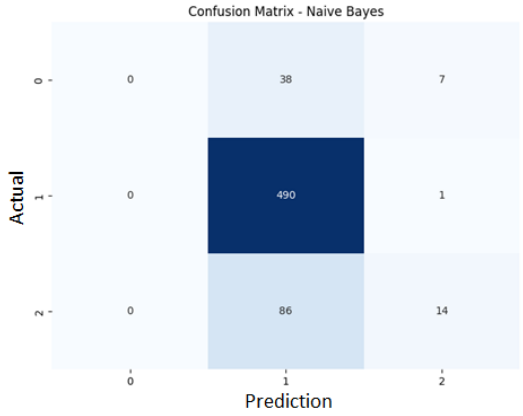

Fig. 12. The confusion matrix from the NBRK_model test.

Table 6.

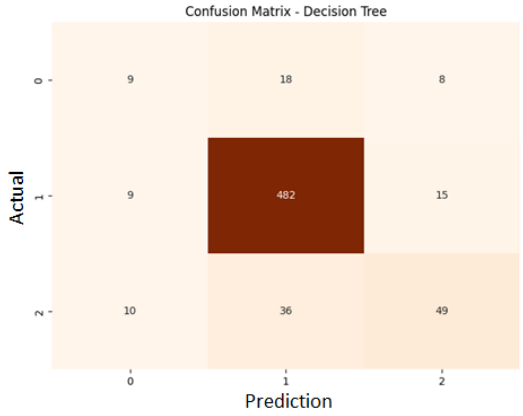

Fig. 13. Confusion matrix from the DTO_model test result.

Table 6. Parameters from the DTO_model Test Result

Text sentiment	Parameters			Support
	Precision	Recall	f1-score	
Positive	0,68	0.52	0.59	95
Negative	0.32	0.26	0.29	35
Neutral	0.90	0.95	0.93	506

The Neutral category has the highest score on the precision, f1-score, and recall parameters of 0.90, 0.93, and 0.95. At the same time, the accuracy value is 0.85. On the other hand, the DTRK_model test result is shown in the confusion matrix in Fig. 14. The parameter values resulting from the confusion matrix in Fig. 14 is shown in Table 7.

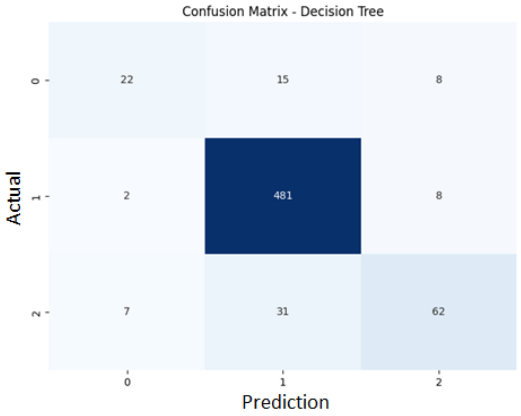

Fig. 14. Confusion matrix from DTRK_model test result.

Table 7. Parameters from DTRK_model Test Result

Text sentiment	Parameters			Support
	Precision	Recall	f1-score	
Positive	0.79	0.62	0.70	100
Negative	0.71	0.49	0.58	45
Neutral	0.91	0.98	0.94	491

The highest value of the precision, f1-score and recall parameters achieved by the Neutral category of 0.91, 0.94 and 0.98. While the accuracy value is 0.86.

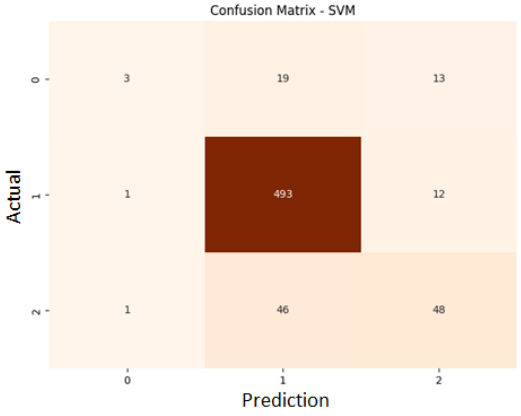

Fig. 15. Confusion matrix from the SVMO_model test result.

Furthermore, the SVMO_model test result is illustrated in the following Fig. 15. The parameter values resulting from the confusion matrix is described in Table 8.

Table 8. Parameters from SVMO_model Test Result

Text sentiment	Parameters			Support
	Precision	Recall	f1-score	
Positive	0.66	0.51	0.57	95
Negative	0.60	0.09	0.15	35
Neutral	0.88	0.97	0.93	506

The precision, f1-score, and recall parameters achieved the highest results for the Neutral category, attaining 0.88, 0.93, and 0.97, respectively. Meanwhile, the prediction accuracy for this model is 0.86. The classification performance of the SVMRK_model is visualized in Fig. 16 through a confusion matrix. The parameters, including f1-score, recall, and precision, are detailed in Table 9.

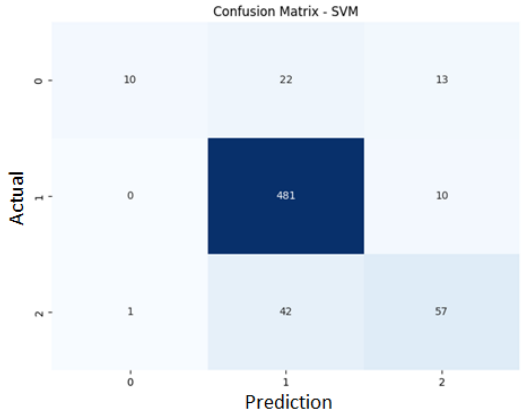

Fig. 16. Confusion matrix from the SVMRK_model test result.

Table 9. Parameters of SVMRK_model Performance

Text sentiment	Parameters			Support
	Precision	Recall	f1-score	
Positive	0.71	0.57	0.63	45
Negative	0.91	0.22	0.36	491
Neutral	0.88	0.98	0.93	100

The highest value of the precision parameter is achieved by the Negative category of 0.91. Then, the Neutral sentiment category obtained the highest results on the f1-score and Recall parameters with values of 0.93 and 0.98. Meanwhile, the accuracy parameter is 0.79.

IV. Discussion

To assess the performance of the proposed data preprocessing during the labeling stage, we compared the labeling results obtained with the previous method and those achieved with the proposed data preprocessing. The results of this comparison are detailed in Table 10.

Table 10. The Data Preprocessing Performance

Sentiment category	Data preprocessing	
	Previous	Proposed
Positive	398 (12,5%%)	426(13,4%)
Negative	184 (5.8%)	208(6,5%)
Neutral	2596 (81.7%)	2544 (80.1%)

Table 10 shows a change in the percentage composition of the Positive, Neutral, and Negative opinion categories after the proposed data preprocessing was applied. In the Positive category, there was an increase from 12.5 % to 13.4 %. Then, in the Negative category, there was also an increase from 5.8 % to 6.5 %. Meanwhile, in the Neutral category, there was a decrease in the percentage from 81.7 % to 80.1 %. In general, implementing the Rabin-Karp algorithm in data preprocessing causes changes in the labeling stage, even though the percentage changes are insignificant.

In this study, a performance comparison analysis between models with the ds_original and ds_RabinKarp datasets could not be carried out because the data composition per category differed at the training and testing stages.

For models with ds_original dataset, the performance measurement between models by measuring the value of the precision parameter was analyzed. Comparison results per category are shown in Fig. 17. The NB model achieved the highest precision parameter value in the Neutral category of 0.90.

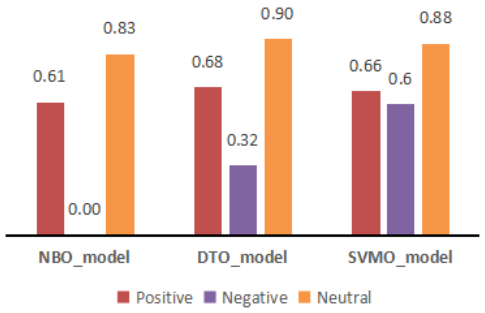

Fig. 17. The performance comparison of parameter precision on models.

Secondly, an analysis of performance between models was conducted through the results of measuring the recall parameter. A comparison of the values of these parameters is shown in Fig. 18.

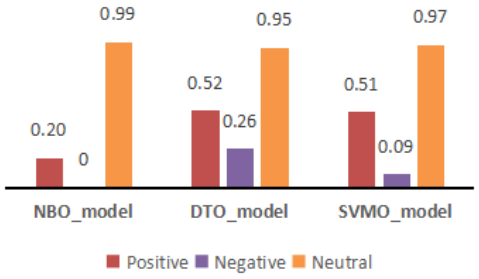

Fig. 18. Performance comparison of the recall parameter on models.

The NB model in the Neutral category of 0.99 achieved the recall parameter's highest value. Thirdly, the performance analysis between models was conducted by measuring the value of the f1-score. A comparison of f1-score results for each model is shown in Fig. 19.

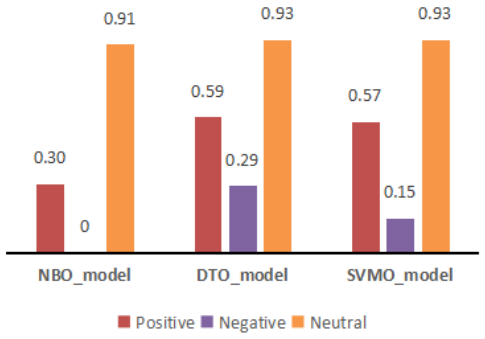

Fig. 19. Performance comparison of the f1-score parameter on models.

The SVM model achieved the highest value of the f1-score parameter in the Neutral sentiment category of 0.93. For models with ds_RabinKarp database, the

performance analysis between models was measured based on the precision parameter. A comparison of the precision results for each model is shown in Fig. 20.

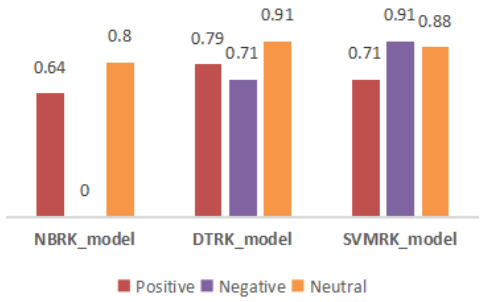

Fig. 20. Performance comparison of the precision parameter.

The DT and SVM models obtained the highest score in the Neutral and Negative categories of 0.91. Furthermore, the performance between models was analyzed based on the recall parameter. The obtained value of this parameter in each model is shown in Fig. 21.

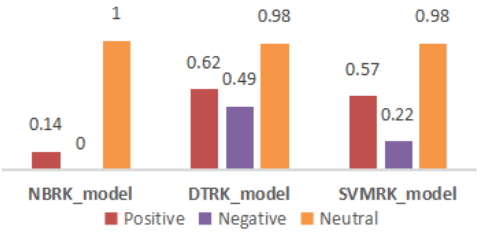

Fig.21. Performance comparison of recall parameter

Fig. 21. Performance comparison of recall parameter.

The performance of the NB models on the recall parameter obtained the highest value of 1. Thirdly, performance analysis between models was carried out based on the f1-score parameter. A comparison between parameter values in each model is shown in Fig. 22. The DT model obtained the highest score in the Neutral category of 0.94.

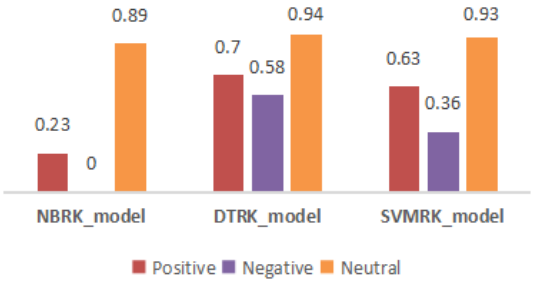

Fig. 22. Performance comparison of the fl-score parameter.

We found that Neutral sentiment received the highest scores when comparing the three parameters across all models. This outcome is due to the use of an unbalanced dataset, where the Neutral sentiment category forms a significant portion. We further analyzed accuracy differences between models using ds_original and ds_RabinKarp datasets, as shown in Fig. 23 and Fig. 24.

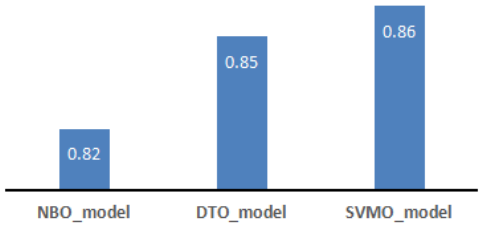

Fig. 23. Performance comparison of the accuracy parameter of models with *ds_original* dataset.

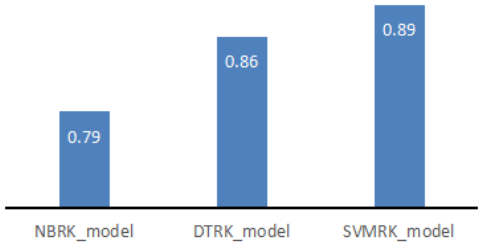

Fig. 24. Performance comparison of the accuracy parameter of models with *ds_RabinKarp* dataset.

The SVM model achieved the highest accuracy parameter value with the ds_original and ds_RabinKarp datasets of 0.86 and 0.89. Based on the comparison results on this parameter, the performance of the SVM model is better than the DT and NB models.

V. Conclusion

In this study, an approach at the data processing stage was proposed. The Rabin-Karp algorithm was inserted to correct non-standard words in the text. The implementation results showed that this approach can improve non-standard words. Based on the results of sentiment analysis, there were changes in the percentage composition of the total data where the Neutral sentiment category changed from 81.7 % to 80.1 %, the Positive category changed from 12.5 % to 13.4 %, and the Negative category changed from 5.8 % to 6.5 %. Furthermore, this study produced six classification models based on the NB, DT, and SVM methods where each method was developed with the ds_original and ds_RabinKarp datasets. The classification models perform better on the Neutral sentiment based on the measurement results of precision, fl-score, and recall parameters. Meanwhile, based on the accuracy parameter, the SVM model performance is superior to the NB and DT models. For further research, feature extraction with the frequency-inverse document approach can be applied to measure this data processing performance.

Acknowledgment

This research was funded by The Indonesian Ministry of Research, Technology and Higher Education

(Ristekdikti), grant number 181/E5/PG.02.00.PL/2023 and 0423.8/LL5-/AL.04/2023.

REFERENCES

[1] M. Bibi, W. A. Abbasi, W. Aziz, S. Khalil, M. Uddin, C. Iwendi, and T. R. Gadekallu, "A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis," *Pattern Recognition Letters*, vol. 158, pp. 80–86, 2022, doi: 10.1016/j.patrec.2022.04.004.

[2] P. Atandoh, F. Zhang, D. Adu-gyamfi, P. H. Atandoh, and R. Elimeli, "Integrated deep learning paradigm for document-based sentiment analysis," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 7, p. 101578, 2023, doi: 10.1016/j.jksuci.2023.101578.

[3] M. Ojeda-Hernández, D. López-Rodríguez, and Á. Mora, "Lexicon-based sentiment analysis in texts using Formal Concept Analysis," *Int. J. Approx. Reason.*, vol. 155, pp. 104–112, 2023, doi: 10.1016/j.ijar.2023.02.001.

[4] D. Suhartono, K. Purwandari, N. H. Jeremy, S. Philip, P. Arisaputra, and I. H. Parmonangan, "Deep neural networks and weighted word embeddings for sentiment analysis of drug product reviews," *Procedia Comput. Sci.*, vol. 216, no. 2022, pp. 664–671, 2023, doi: 10.1016/j.procs.2022.12.182.

[5] M. Costola, O. Hinz, M. Nofer, and L. Pelizzon, "Machine learning sentiment analysis, COVID-19 news and stock market reactions," *Res. Int. Bus. Financ.*, vol. 64, no. January, p. 101881, 2023, doi: 10.1016/j.ribaf.2023.101881.

[6] H. Luo, X. Meng, Y. Zhao, and M. Cai, "Exploring the impact of sentiment on multi-dimensional information dissemination using COVID-19 data in China," *Comput. Human Behav.*, vol. 144, no. November 2022, p. 107733, 2023, doi: 10.1016/j.chb.2023.107733.

[7] M. Elahi, D. Khosh Kholgh, M. S. Kiarostami, M. Oussalah, and S. Saghari, "Hybrid recommendation by incorporating the sentiment of product reviews," *Inf. Sci. (Ny).*, vol. 625, pp. 738–756, 2023, doi: 10.1016/j.ins.2023.01.051.

[8] M. Wilksch and O. Abramova, "PyFin-sentiment: Towards a machine-learning-based model for deriving sentiment from financial tweets," *Int. J. Inf. Manag. Data Insights*, vol. 3, no. 1, p. 100171, 2023, doi: 10.1016/j.jjimei.2023.100171.

[9] M. Umer, S. Sadiq, M. Nappi, M. U. Sana, and I. Ashraf, "ETCNN: Extra tree and convolutional neural network-based ensemble model for COVID-19 tweets sentiment classification: ETCNN: COVID-19 Tweets Sentiment Classification," *Pattern Recognit. Lett.*, vol. 164, pp. 224–231, 2022, doi: 10.1016/j.patrec.2022.11.012.

[10] N. Leelawat, S. Jariyapongpaiboon, A. Promjun, S. Boonyarak, K. Saengtabtim, A. Laosunthara, A. K. Yudha, and J. Tang, "Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning," *Heliyon*, vol. 8, no. 10, p. e10894, 2022, doi: 10.1016/j.heliyon.2022.e10894.

[11] M. Qorib, T. Oladunni, M. Denis, E. Ososanya, and P. Cotae, "Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination twitter dataset," *Expert Syst. Appl.*, vol. 212, no. January 2022, p. 118715, 2022, doi: 10.1016/j.eswa.2022.118715.

[12] D. Purwitasari, C. B. P. Putra, and A. B. Raharjo, "A stance dataset with aspect-based sentiment information from Indonesian COVID-19 vaccination-related tweets," *Data Br.*, vol. 47, p. 108951, 2023, doi: 10.1016/j.dib.2023.108951.

[13] G. Revathy, S. A. Alghamdi, S. M. Alahmari, S. R. Yonbawi, A. Kumar, and M. Anul Haq, "Sentiment analysis using machine learning: Progress in the machine intelligence for data science," *Sustain. Energy Technol. Assessments*, vol. 53, no. PC, p. 102557, 2022, doi: 10.1016/j.seta.2022.102557.

[14] A. Solairaj, G. Sugitha, and G. Kavitha, "Enhanced Elman spike neural network based sentiment analysis of online product recommendation," *Appl. Soft Comput.*, vol. 132, p. 109789, 2023, doi: 10.1016/j.asoc.2022.109789.

[15] C. Qian, N. Mathur, N. H. Zakaria, R. Arora, V. Gupta, and M. Ali, "Understanding public opinions on social media for financial sentiment analysis using AI-based techniques," *Inf. Process. Manag.*, vol. 59, no. 6, p. 103098, 2022, doi: 10.1016/j.ipm.2022.103098.

[16] A. K. Saputra, K. Muludi, and T. Thamrin, "Comparative analysis between Rabin Karp algorithm, Winnowing, and Turnitin applications for detecting plagiarized words," *Proceeding 6th ICITB 2020*, pp. 40–49, 2020, [Online]. Available: https://jurnal.darmajaya.ac.id/index.php/icitb/article/view/2505

[17] W. Hidayat, E. Utami, and A. Sunyoto, "Selection of the best k-gram value on modified Rabin-Karp algorithm," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 16, no. 1, p. 11, 2022, doi: 10.22146/ijccs.63686.

[18] M. Harahap, R. Marcel, D. Y. Milatrisna, and S. Kuswulandari, "Fake article detection application with Rabin Karp algorithm," *Sinkron*, vol. 8, no. 1, pp. 166–170, 2023, doi: 10.33395/sinkron.v8i1.11949.

[19] C. Chen, S. S. Berutu, Y. Chen, H. Yang, and C. Chen, "Regulated two-dimensional deep convolutional neural network-based power quality classifier for microgrid," *Energies.*, vol. 15, no.7, p. 2532 2022.

[20] S. S. Berutu, "Text mining dan klasifikasi sentimen berbasis Naïve Bayes pada opini masyarakat terhadap makanan tradisional," *J. Sist. Komput. dan Inform.*, vol. 4, no. 2, p. 254, 2022, doi: 10.30865/json.v4i2.5138.