Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 11, No. 2, June 2025, pp. 263-275
ISSN: 2338-3070, DOI: 10.26555/jiteki.v11i2.30605

Impact of Cosine Similarity Function on SVM Algorithm for Public Opinion Mining About National Sports Week 2024 on X

Abil Mansyur, Ichwanul Muslim Karo Karo, Muliawan Firdaus, Elmanani Simamora, Muhammad Badzlan Darari, Rizki Habibi, Suvriadi Panggabean
Faculty of Mathematics and Natural Science, Universitas Negeri Medan, Jl. William Iskandar Ps. V, Medan 20371, Indonesia

ARTICLE INFO

Article history:

Received January 20, 2025
Revised March 26, 2025
Accepted May 16, 2025

Keywords:

PON;
Opinion mining;
SVM;
Cosine similarity;
Kernel function;
Performance

ABSTRACT

Public opinion on PON 2024 (National Sports Week in Indonesia) became a trending topic on X (formerly Twitter), reflecting both positive and negative sentiments. Understanding these sentiments is important for evaluating the event and preparing for the upcoming. However, baseline SVM algorithms using standard kernel functions are not optimized for text similarity and limit performance in sentiment analysis. This research proposes cosine similarity as a substitution for the kernel function on SVM, enhancing the sentiment analyzer's performance on public opinions about PON 2024. The approach leverages cosine similarity's strength in handling text-based data. The key contribution of this research is the integration of cosine similarity into the SVM algorithm as a replacement for kernel functions, improving performance in sentiment analysis. Additionally, this study offers a comprehensive comparison with baseline SVM and provides actionable insights for upcoming PON. The study collected 1,011 tweets related to PON 2024 using web scraping and the Twitter API, followed by labeling sentiments as positive, neutral, or negative. Several preprocessing techniques also were applied to prepare the data. Two models were developed: baseline SVM and another using SVM integrated with cosine similarity, both evaluated through accuracy, precision, recall, and F1-score. The baseline SVM achieved 85.1% accuracy, 85% precision, 83% recall, and 83.3% F1-score, struggling particularly with negative sentiment. Opposite, by integrating cosine similarity on SVM, the performance improved to 88.73% accuracy, 88.3% precision, 89.3% recall, and 88.3% F1-score—a boost of 3.3-6.3%. Additionally, the public opinion revealed that positive sentiments mostly focused on athlete achievements and medal awards, while negative sentiments highlighted issues like referee performance and specific sports (e.g., football). This approach can serve as a valuable tool for event organizers to identify public concerns and maintain positive aspects for the upcoming PON 2028.

Corresponding Author:

Ichwanul Muslim Karo Karo, Faculty of Mathematics and Natural Science, Universitas Negeri Medan, Jl. William Iskandar Ps. V, Medan 20371, Indonesia
Email: ichwanul@unimed.ac.id

1. INTRODUCTION

The National Sports Weeks (Indonesian: *Pekan Olahraga Nasional*, abbreviated as *PON*) is biggest multi-sport competition event held every four years in Indonesia [1]. The participants are athlete delegates from various provinces in Indonesia. The host is Aceh and North Sumatra province in 2024. Discourse about PON 2024 is trending on X social media with various issues and topics through the hashtags *#PON2024*, *#PONXXIAcehSumut*, *#BersatuKitaJuara*, and *#TwibbonPON2024*. Active user expressed their opinions on

the pros and cons of issues during PON 2024. Tweets related to dissatisfaction, criticism or diatribes on X are negative sentiments, and vice versa.

Sentiment analysis, or also known as opinion mining [2], [3], is the process of analyzing digital text to determine whether it has positive, negative, or neutral content [4]. This study notes that several studies have been realized and evaluated public opinion mining for sporting events. Public opinion mining on X about the FIFA World Cup event in 2014 [5]. The tweets were crawled and labeled manually. The sentiment analyzer used the CNN algorithm. The performance was evaluated by the F1 metric of 0.6634. Tweets about the 2018 FIFA World Cup have also been analyzed for sentiment by [6]. The study evaluated of several machine learning techniques. SVM outperforms other models such as Naive Bayes, ANN, k-NN, and Logistic Regression. It also analyzed sentiment for the Qatar 2022 World Cup from Twitter users in Arab countries [7]. The study explored the opinions using machine learning algorithms (logistic regression, random forest classifier, Naive Bayes classifier, and SVM). One of the best algorithms was SVM (accuracy = 93%). Other researchers have also analyzed public sentiment about FIFA World Cup 2022 [8]. The study used deep learning models such as RoBERTa, DistilBERT, and XLNet to analyze the sentiment of opinion on X social media for the first day of the tournament. The performance of the models was assessed using measures such as accuracy, F1-score, precision, and recall. It is shown that roBERTa is a best model for classifying sentiment than others. In addition to the FIFA World Cup, other sporting events have also become topics of conversation on social media and analyzed the opinion. A football match in Turkey also attracted talk on X and had its sentiment analyzed [9]. There are 30,000 Turkish tweets about it. In the labeling process, four sentiment classes were used as positive, negative, neutral, and irrelevant. There are four machine learning algorithms (Naïve Bayes, K-Nearest Neighborhood, C4.5, and SVM) that were evaluated. The accuracy of the SVM algorithm was 92.30% and outperformed other algorithms. Part of this study [10] analyzed sentiment for four Olympic games: from 2008 to 2022. There were two sources of data sets: YouTube and Twitter. The sentiment analyzer used a Naive Bayes machine learning approach. Unfortunately, this study did not evaluate the model. This study focused on word-by-word associations in the corpus.

Analyzing public opinion about *PON* 2024 can be an evaluation resource for the upcoming *PON* 2028. Positive sentiments about the organization of *PON* 2024 indicate that the aspects discussed and appreciated by the public should be maintained. Meanwhile, negative sentiment is an unfavorable record in the eyes of the public. It could be an input for improvement. In other words, this analysis will not only provide an overview of public perceptions of *PON* 2024 but can also help organizers to identify potential problems and aspects that need to be improved. For example, if there are many negative sentiments related to infrastructure or security, organizers can make immediate improvements. On the other hand, extracting public opinion is also another evaluator for stakeholders outside the *PON* 2024 committee report.

Support Vector Machine (SVM) algorithm has been widely used and popular machine learning algorithm for opinion mining cases [11], [12]. The SVM algorithm performs by finding the optimal hyperplane that separates classes of data by maximizing the margin [13], thereby classifying data with high precision [14], [15]. Typically, the margin function optimization in SVM algorithms is based on a modified distance [16], [17] or Kernel function [18], [19]. However, the kernel function and not all modified distances are not designed for text data and natural language cases [20], [21], [22]. This research aims to analyze public sentiment about PON 2024 on social media X with the SVM algorithm and cosine similarity function. Cosine similarity is more effective in handling text [23], as it measures similarity based on the orientation of word vectors in a multidimensional space [22], whereas the kernel function is designed more for numerical and non-linear data [24]. In addition, cosine similarity is not affected by document length [25], whereas kernel function can be affected by the scale of the data, requiring additional normalization of the text. The cosine similarity function is a substitute for the distance function or kernel in SVM. Model evaluation using accuracy, precision, recall, and F1. The key contribution of this research is the integration of cosine similarity into the SVM algorithm as a replacement for kernel functions, improving accuracy in sentiment classification. Additionally, this study offers a comprehensive comparison with baseline SVM and provides actionable insights into public sentiment regarding PON 2024. This analysis aims to identify the most effective model for accurately analyzing public opinion. In addition, this research also provides insights regarding aspects that should be maintained for the upcoming *PON* 2028 and aspects that need to be improved. The insights are based on word cloud analysis of each sentiment class. In addition, this research is also expected to contribute to the development of science in the field of sentiment analysis in Indonesia and become material for evaluating and planning the hosting of the upcoming *PON*.

2. METHODS

There are five main stages in this research (shown in Fig. 1). This research begins by collecting datasets from social media X, and the resulting output is tweets. The next process will go through text pre-processing with various techniques. The output of this stage is in the form of structured raw data. In other words, the tweets have been successfully extracted into several features. The raw data is split into training and testing datasets. The training dataset aims to build and train the text classification model. Whereas the testing dataset is used to evaluate the model of text classification.

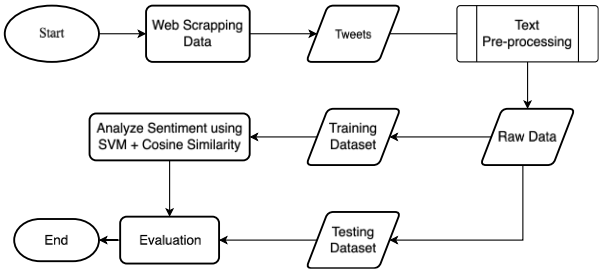

Fig. 1. Research flowchart

2.1. Data Collection

The data collection process uses web scraping techniques. The technique is a technology that allows us to extract structured data from text [26]; in this case, texts come from social media X. The process is supported by the Twitter API feature. The feature allows developers to use Tweet-Harvest, which is available on the website https://developer.twitter.com/. Tweet-Harvest is a Node.js-based tool used to collect tweet data from X [27]. However, the web scraping process will attract spam and biased tweets; therefore, to filter relevant tweets, there are several criteria limitations in a text that will be scraped from social media X (shown in Table 1). Tweets must contain keywords/hashtag elements (such as: *PON 2024, PON Aceh-Sumut,* and *PON XXI*); otherwise, they are removed. Although there are many tweets before the start of *PON* 2024, this study focuses on tweets within the time frame of *PON* 2024. As a result, 1,011 tweets were successfully obtained from the process.

Table 1. Tweet criteria

Criteria	Limitation
Keywords/hashtag	*PON 2024, PON Aceh-Sumut, PON XXI/,* etc.
Time spends	9 – 20 September 2024
Language	Indonesia
Tweet	Tweets are not repeated

2.2. Text Pre-processing

Text preprocessing is a natural language processing technique that transforms raw data into a format that is understandable, predictable, and analyzable [28]. In many cases, it could increase the performance model [29], [30], [31]. The objective of this process is that tweets are transformed and shaped into features. There are six text preprocessing techniques used in this research (Table 2). The cleansing technique removes predefined symbols, such as #, $, and @ on usernames. This study used case folding to homogenize the letter. A brief explanation of text preprocessing techniques is shown in Table 2.

2.3. Sentiment Analyzer

The identifying sentiment of the public about PON 2024 used the Support Vector Machine (SVM) algorithm. The SVM algorithm is a machine learning algorithm that can handle numerous variables and classes [32], [33], [34]. The idea of the SVM algorithm is to create a hyperplane function that can separate the data [3], [20], [35], the algorithm is shown below. the algorithm is shown below. The SVM algorithm input includes the number of input vectors (N_{sv}), number of features (N_{ft}), and bias (b). The input vector and support vector are in array format. Bias (b) is initial state by 1. The algorithm output is an equation model in hyperplane format. However, the SVM algorithm has limitations and poor performance for finding patterns on a non-linear dataset [36], [37], [38]. Many researchers use the kernel trick to enhance the performance of the model [39], [40], [41], [42]. Unfortunately, Kernel tricks are not designed as text similarity functions [22].

Table 2. Text pre-processing techniques

Techniques	Description
Cleansing	Aims to reduce noise in the collected tweet, such as removing symbols, URL links: hashtag "#", and at "@" when mentioning the username.
Case Folding	Converting words that have uppercase into lowercase. The goal is to eliminate data redundancy.
Tokenization	Separation of sentences into single words and checking the word of the *i*-th character is not a separator such as period (.), comma (,), space and other separators, it will be combined with the next character.
Stopwords Removal	Each word will be checked, if there are conjunctions, prepositions or pronouns, then the word will be removed.
Stemming	The process of converting words in tweet data that have affixed words into basic words [21]. This study uses the Sastrawi library in python.
Feature Extraction	The process of identifying and extracting important features from raw data. This study used Bag of Word, like previous studies [13].

SVM Algorithm

Input:
N_{in}; N_{sv}; N_{ft}; b; $SV[N_{sv}]$ (Support vector array); $IN[N_{in}]$ (Input vector array)

Output:
F (decision function output)

```
for i ← 1 to N_in by 1 do
    F = 0
    for j ← 1 to N_sv by 1 do
        dist = 0
        for k ← 1 to N_ft by 1 do
            dist += (SV[j].feature[k] − IN[i].feature[k])^2
        end
        k = exp (−γ × dist)
        F+= SV[j].a ∗ × k
    end
    F = F + b ∗)
end
```

This study proposed an idea that substitutes the kernel trick with a text similarity function. Cosine similarity is one of the most popular distance measures in text classification problems [43], [44]. Cosine similarity is used frequently as an alternative to the kernel function in SVM to handle text-based data [45]. Cosine similarity is more effective in handling text [25], [46]. It measures similarity based on the orientation of word vectors in a multidimensional space. Whereas the kernel function is designed for numerical and non-linear data [24]. Cosine similarity is not affected by document length [47]. The kernel function is affected by the scale of the data, so it requires data normalization. Furthermore, this study modified the SVM algorithm by replacing the kernel function with the cosine similarity function. The cosine similarity function is used to measure the closeness of tweets. Technically, the cosine similarity function is used to replace the process in line 8 in the SVM algorithm.

2.4. Model Evaluation

Model of sentiment analyzer will be evaluated and tested against the testing data. This study used confusion matrix (Table 3). Confusion matrix is a table used to evaluate the performance of sentiment analysis models [48]. The confusion matrix shows the number of correctly and incorrectly classified data, and helps identify which class of data is most frequently misplaced [35], [49]. If a model predicts a tweet as positive sentiment, and it is actually labeled as positive, it's a true positive. A false positive occurs when the model incorrectly classifies a neutral or negative sentiment as positive. A false negative occurs when the model incorrectly classifies a positive sentiment as neutral or negative. A true negative occurs when the model correctly classifies a negative sentiment as negative.

A confusion matrix can be used to calculate metrics performance [35], [49]. This study used accuracy (*Acc*), precision (*P*), recall (*R*), and F1 score metric (shown in Table 4). The formula of metric is calculated from the confusion matrix. These metrics are important for evaluating the performance of the models, especially when the data is unbalanced [50].

Table 3. Confusion matrix

		Predict	
		Positive	**Negative**
Actual	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True Negative (TN)

Table 4. Explanation performance metric

Metric	Description	Formula
Precision (*P*)	Indicates when the model predicts a positive, how often the prediction is correct [49]	$P = \frac{TP}{TP+FP} .100\%$
Recall (*R*)	Indicated sensitivity of model to predict the data [51]	$R = \frac{TP}{TP+FN} .100\%$
F1 score	Comparison of weighted average precision and recall [49].	$F1\ score = \frac{2PR}{P+R}$
Accuracy (*Acc*)	A measure that determines the degree of similarity between the result of a measurement and the value actually [27].	$Acc = \frac{TP+TN}{TP+TN+FP+FN} .100\%$

3. RESULTS AND DISCUSSION

This section examines the results of the experiment while simultaneously providing an in-depth discussion. This research ran two experiments. The first experiment analyzed sentiment with the baseline SVM algorithm. This means that the distance function in the SVM algorithm has not been substituted by cosine similarity. The second experiment analyzed sentiment with the SVM algorithm with the cosine similarity function. Each experiment was run with a *K*-fold cross-validation scenario ($K = 5$). An illustration of the data split for running both experiments is presented in Fig. 2. In other words, each experiment is run five times with 80% training data and 20% test data. The aim of *K*-Fold cross-validation is to avoid overfitting the model. The models derived from the two experiments were evaluated with accuracy, precision, recall, and F1 score metrics. In addition, this section also presents an analysis of the performance of both. Lastly, the section presents an analysis of topics in negative and positive content as an insight for the upcoming *PON* 2028.

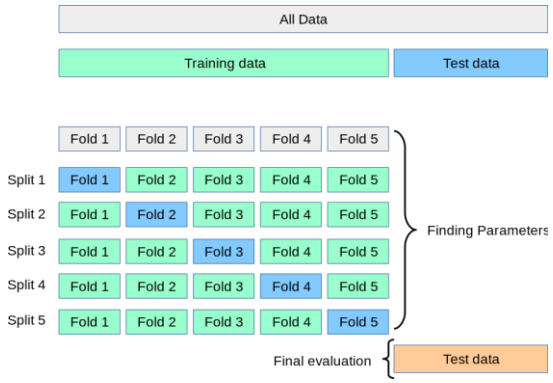

Fig. 2. 5-Fold cross validation

3.1. Dataset

This section presents the results of web scraping and tweet melting. The number of tweets that were successfully scraped was 1011 and then stored in a CSV file. The sample tweets can be seen in Table 5. There are nine pieces of information for a successfully collected tweet, the identity of the crawled tweets (Id), the identity of the conversation (*conversation_id_str*), the time of user created tweets (*creat_at*), the number of people who liked the tweet (*favorite_count*), the full text of tweets (*full_text*), unique identifier for the tweet (*id_str*), if this tweet contains an image, the field will contain the URL of the image (*Image_url),* If the represented tweet is a reply, this field will contain the screen name of the original post's author (*In_reply_to_screen_name*), and the corresponding language identifier (*lang*).

Tabel 5. Example of tweets

Id	*Conversation_id_str*	*Created_at*	*Favourite_count*	*Full_text*	*Id_str*	*Image_url*	*In_reply_to_screen_name*	*Lang*
1	1837356819104288874	Sat Sep 21 05:02:56 +0000 2024	0	Presiden Jokowi diwakili Menteri Koordinasi (M...	1837356819104288874	NaN	NaN	in
.	.	.		.	.	.	.	.
.	.	.		.	.	.	.	.
.	.	.		.	.	.	.	.
1011	1837355846088720822	Sat Sep 21 04:59:04 +0000 2024	0	Sobat Polri mari kita dukung semangat persatua...	1837355846088720822	https://pbs.twimg.com/ext_tw_video_thumb/18373...	NaN	in

This study only uses the full_text field. In other words, only tweets will be used for the next stage. Furthermore, each tweet is labeled positive, negative, or neutral. The proportion shown in Fig. 3. There are 591 or 58.5% of tweets showed positive opinions in the form of pleasure, agreement, or opinions that accept, benefit, or support the implementation of *PON* 2024 in Indonesia. There are 32.4% or 328 tweets giving opinions of complaints, anger, or negative sentiment about *PON* 2024, while the rest are neutral.

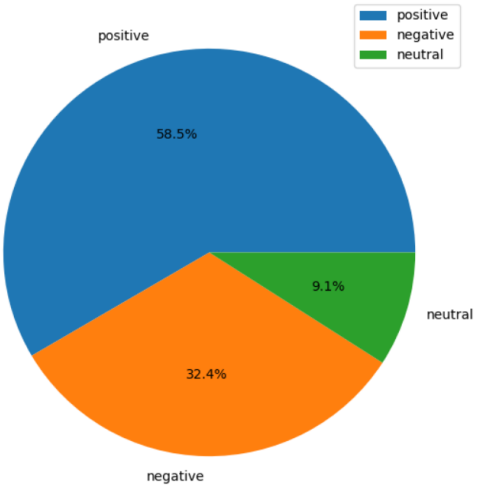

Fig. 3. Proportion of tweets

3.2. Text Pre-processing Result

This section presents the text pre-processing results. There are five text pre-processing techniques used in this study: clean review, case folding, tokenization, Stopword removal, and stemming. The results of the five techniques can be seen in Table 6. In addition, this study used the bag-of-words method as feature extraction. There are 1908 unique terms. It means the unique terms are the variables of the dataset. Thus, unstructured text can be more easily processed by machine learning algorithms.

3.3. Performance of Baseline SVM

This section presents the results of the first experiment. Baseline SVM performance is the preliminary result as a benchmark before applying cosine similarity on the SVM algorithm. This study used a kernel linear function. Twenty percent, or 809 tweets of the dataset, were used for training the model. The remaining tweets were used to test the model. The training accuracy of the model is 90.741%. While the model performance, accuracy 85.1%, precision 85%, recall 83%, and F1 score 83.3%. The details of the model performance in analyzing each class are presented in Table 7. The resulting model is quite reliable in identifying neutral and

positive classes, but not for negative classes. This can be seen from the values of precision, recall, and F1 score for the negative class, which are low compared to other classes.

Table 6. Example of text pre-processing results

id	Full Text	Cleaned Reviews	Casefolding	Tokenizing	Stopword removal	Stemming Text
0	*Masuk tim Tiger Cup 2004 era Peter Withe pdhl...*	*Masuk tim Tiger Cup era Peter Withe pdhl CMII...*	*masuk tim tiger cup era peter withe pdhl cmii...*	*[masuk, tim, tiger, cup, era, peter, withe, pd...*	*[masuk, tim, tiger, cup, era, peter, white, pd...*	*[masuk, tim, tiger, cup, era, peter, white, pd...*
1	*Isyana/Rinjani perlebar lagi gapnya yang sebel...*	*IsyanaRinjani perlebar lagi gapnya yang sebelu...*	*isyanarinjani perlebar lagi gapnya yang sebelu...*	*[isyanarinjani, perlebar, lagi, gapnya, yang, ...*	*[isyanarinjani, perlebar, gapnya, menipis, mat...*	*[isyanarinjani, lebar, gap, tip, match, dexter...*
2	*HEBAT BANGEETT !!! Selamat kepada Mokesano Shir...*	*HEBAT BANGEETT Selamat kepada Mokesano Shirraj...*	*hebat bangeett selamat kepada mokesano shirraj...*	*[hebat, bangeett, selamat, kepada, mokesano, s...*	*[hebat, bangeett, selamat, mokesano, shirraja,...*	*[hebat, bangeett, selamat, mokesano, shirraja,...*

Tabel 7. Performance of baseline SVM model for every label

Class	Precision	Recall	F1-score
Negative	0.82	0.80	0.81
Neutral	0.88	0.83	0.85
Positive	0.85	0.86	0.85

3.4. Performance of SVM with Cosine Similarity

This section presents the results of the second experiment. Cosine similarity substitutes the distance function on the SVM algorithm. The cosine formula is shown in Equation 1, where Tw_i is *i*-th tweet. The composition of the dataset used for model training and testing is still the same as the previous experiment. The training accuracy of the model is 90.39%. While the model performance, accuracy 88.73%, precision 88.3%, recall 89.3%, and F1 score 88.3%. The neutral class has the highest precision (0.98), meaning the model is highly confident in neutral classifications. The positive class has the highest recall (0.95), meaning most positive samples were correctly classified, but with lower precision (0.75), suggesting some false positives. The negative class maintains a balanced performance across all metrics (shown in Table 8).

$$Cosine\left(Tw_i, Tw_j\right) = \frac{Tw_i Tw_j}{\|Tw_i\| \|Tw_j\|} \quad (1)$$

Tabel 8. Performance of SVM + cosine similarity model for every label

Class	Precision	Recall	F1-score
Negative	0.92	0.86	0.89
Neutral	0.98	0.87	0.92
Positive	0.75	0.95	0.84

These results indicate that SVM with cosine similarity achieves high classification accuracy, with minimal performance drop between training and testing, suggesting a well-generalized model. The SVM with the cosine similarity model demonstrates strong classification performance, especially in neutral and negative sentiment detection. However, improvements in precision for the positive class could be made, possibly through feature selection, data balancing, or hyperparameter tuning.

3.5. Comparison of Experimental Results

This section presents a comparative analysis of the scenarios. The comparison chart is presented in Fig. 4. The performance of the baseline SVM for analyzing public opinion about PON 2024 on social media X is in the range of 83 - 85.1%. Based on the data in Fig. 4, the sensitivity of the baseline SVM model is the lowest. While the sensitivity of the model of SVM with cosine similarity is the highest. In addition, all performance metrics show that the SVM model with cosine similarity is able to analyze the sentiment of public opinion

about PON 2024 on X better by 3.3 - 6.3%. The sensitivity of the SVM model with cosine similarity has increased the most compared to other metrics.

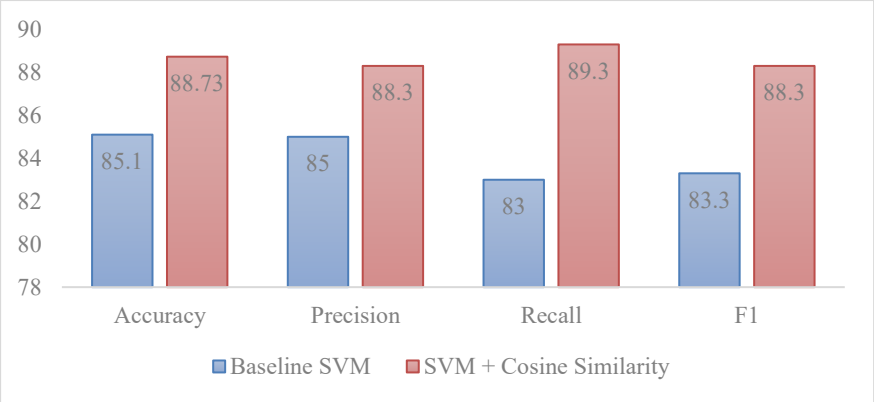

Fig. 4. Comparison of the performance of the two experimental scenarios

This study also captures some differences in sentiment analysis results by the two experiment scenarios. The baseline SVM model mostly analyzes a tweet as a negative opinion, even though the actual class is not. Surely, this condition makes the performance of the baseline SVM model have low performance for analyzing negative labels. This condition is also supported by the data in Table 9, where the F1 score of the model for analyzing the negative class is the lowest compared to the analysis results for other classes. Cosine similarity in SVM provides improved performance over the baseline SVM that uses a linear kernel. It is more suitable for text data, as it considers word patterns without being affected by document length. This combination improves accuracy, precision, recall, and F1-score compared to the linear kernel. It reduces overfitting, as it does not depend on the vector length but only on the similarity angle between documents. It is more stable in handling differences in document size, resulting in more accurate and consistent classification.

Tabel 9. Performance of SVM + cosine similarity model for every label

Tweets	Prediction using Baseline SVM	Prediction using SVM + Cosine Similarity	Actual class
presiden jokowi diwakili Menteri Koordinasi (m...	Negative	Positive	Positive
pukul wasit di pon xxi aceh-sumut muhammad riz...	Negative	Negative	Negative
Sobat Polri mari kita dukung semangat persatua...	Negative	Neutral	Positive

3.6. Insight of Sentiment Analysis

This section presents the topic of each positive tweet and negative tweet with Word cloud. Word cloud makes it easy to describe words by providing interesting and informative text visualizations. The goal is to capture what things should be improved and maintained in the next *PON* activities. The more often a word appears, the larger its size, and vice versa. Words with low frequency will be displayed in a smaller size. Green Word cloud in Fig. 5 is a visualization of topics of positive tweets, while red Word cloud is a visualization of negative tweets. Based on Fig. 5, there are several topics of positive tweets, including the West Java contingent becoming the overall champion and bonuses for medal acquisition. The recommendation is that it should be maintained or improved in upcoming *PON*. Meanwhile, the topics of negative tweets include poor referee performance, and aspects of the football branch. Therefore, the recommendation is that these two aspects must be improved for the upcoming *PON*.

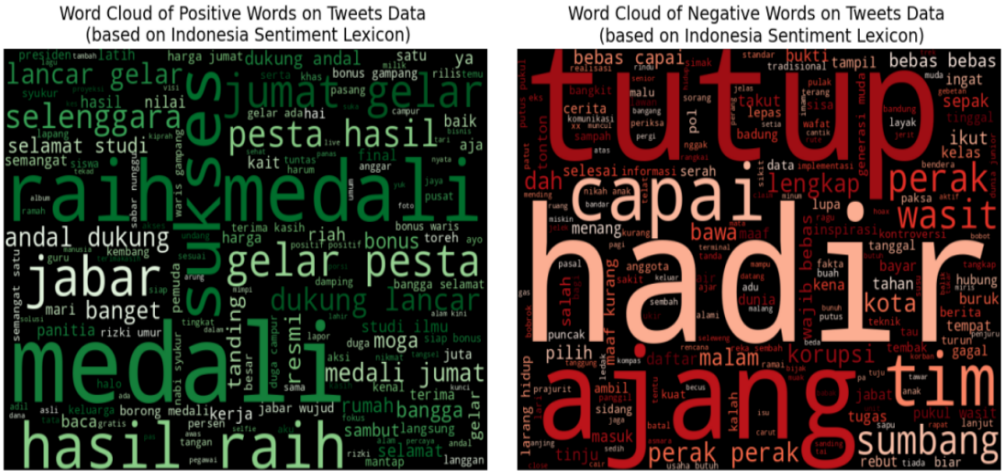

Fig. 5. Topic of positive and negative tweet

4. CONCLUSION

This research conducted a sentiment analysis of public opinion regarding PON 2024 using the Support Vector Machine (SVM) algorithm, integrating the cosine similarity function as an alternative to traditional kernel functions. The study utilized a dataset of 1,011 tweets collected from social media platform X, classified into positive, neutral, and negative sentiments. The data underwent a rigorous preprocessing phase, including text cleaning, tokenization, and stopword removal, to enhance the performance of sentiment classification models. Two experimental scenarios were implemented and evaluated using accuracy, precision, recall, and F1-score metrics. The first scenario, which used the baseline SVM model, achieved an accuracy of 85.1%, a precision of 85%, a recall of 83%, and an F1-score of 83.3%. The analysis revealed that the baseline model exhibited weaknesses in accurately classifying negative sentiment. The second scenario, integrating cosine similarity into the SVM model, significantly improved classification performance, achieving an accuracy of 88.73%, a precision of 88.3%, a recall of 89.3%, and an F1-score of 88.3%. These findings underscore the effectiveness of cosine similarity in enhancing the model's ability to distinguish sentiment in textual data. The sentiment analysis provided insightful evaluations of public perceptions regarding *PON* 2024. Positive sentiments were primarily associated with aspects such as athlete achievements and medal awards, suggesting strong public support for these elements. Conversely, negative sentiments highlighted concerns over referee performance and specific sports categories, notably football. These insights provide valuable feedback for event organizers and stakeholders, allowing them to maintain successful elements while addressing public concerns in preparation for *PON* 2028. Overall, this study demonstrates the applicability of sentiment analysis as a strategic evaluation tool for large-scale events. By improving the accuracy of sentiment classification through cosine similarity, this research contributes to the field of opinion mining and text-based machine learning applications. The model struggles with negative sentiment classification, future research could explore additional modifications to the model, including hyperparameter tuning and feature selection techniques, to further enhance performance and applicability in diverse domains.

REFERENCES

[1] K. Kogoya, T. S. Guntoro, and M. F. P. Putra, "Sports Event Image, Satisfaction, Motivation, Stadium Atmosphere, Environment, and Perception: A Study on the Biggest Multi-Sport Event in Indonesia during the Pandemic," *Soc Sci*, vol. 11, no. 6, 2022, https://doi.org/10.3390/socsci11060241.

[2] N. T. Mhaske and A. S. Patil, "Resource creation for opinion mining: a case study with Marathi movie reviews," *International Journal of Information Technology (Singapore)*, vol. 13, no. 4, 2021, https://doi.org/10.1007/s41870-021-00698-8.

[3] P. Kiran Kumar, N. Jahna Tejaswi, M. L. Vasanthi, L. L. Srihitha, and B. Phanindra Kumar, "Sentimental Analysis on Multi-domain Sentiment Dataset Using SVM and Naive Bayes Algorithm," in *Communications in Computer and Information Science*, pp. 201-213, 2022. https://doi.org/10.1007/978-3-030-95502-1_16.

[4] W. Chansanam and K. Tuamsuk, "Thai twitter sentiment analysis: Performance monitoring of politics in Thailand using text mining techniques," *International Journal of Innovation, Creativity and Change*, vol. 11, no. 12, pp. 436-452, 2020, https://www.ijicc.net/images/vol11iss12/111227_Chansanam_2020_E_R.pdf.

[5] H. Hettiarachchi, D. Al-Turkey, M. Adedoyin-Olowe, J. Bhogal, and M. M. Gaber, "TED-S: Twitter Event Data in Sports and Politics with Aggregated Sentiments," *Data (Basel)*, vol. 7, no. 7, 2022, https://doi.org/10.3390/data7070090.
[6] N. Pombo, M. Rodrigues, Z. Babic, M. Punceva, and N. Garcia, "Computerised Sentiment Analysis on Social Networks. Two Case Studies: FIFA World Cup 2018 and Cristiano Ronaldo Joining Juventus," in *Advances in Intelligent Systems and Computing*, vol. 29, pp. 126-140, 2021. https://doi.org/10.1007/978-3-030-72651-5_13.
[7] M. Faisal, Z. Abouelhassan, F. Alotaibi, R. Alsaeedi, F. Alazmi, and S. Alkanadari, "Sentiment Analysis Using Machine Learning Model for Qatar World Cup 2022 among Different Arabic Countries Using Twitter API," in *IEEE World AI IoT Congress*, pp. 0222-0228, 2023. https://doi.org/10.1109/AIIoT58121.2023.10188463.
[8] S. S. Arnob, M. A. A. Shikder, T. A. Ovey, E. R. Rhythm, and A. A. Rasel, "Analyzing Public Sentiment on Social Media during FIFA World Cup 2022 using Deep Learning and Explainable AI," in *26th International Conference on Computer and Information Technology, ICCIT*, pp. 1-6, 2023. https://doi.org/10.1109/ICCIT60459.2023.10441156.
[9] R. KORKUSUZ and A. CARUS, "Futbol Müsabakaları ile İlgili Tweetlerin Anlık Duygu Analizi," *European Journal of Science and Technology*, pp. 386-396, 2020, https://doi.org/10.31590/ejosat.821200.
[10] K. Jia, Y. Zhu, Y. Zhang, F. Liu, and J. Qi, "International public opinion analysis of four olympic games: From 2008 to 2022," *Journal of Safety Science and Resilience*, vol. 3, no. 3, 2022, https://doi.org/10.1016/j.jnlssr.2022.03.002.
[11] A. P. Gopi, R. N. S. Jyothi, V. L. Narayana, and K. S. Sandeep, "Classification of tweets data based on polarity using improved RBF kernel of SVM," *International Journal of Information Technology (Singapore)*, vol. 15, no. 2, 2023, https://doi.org/10.1007/s41870-019-00409-4.
[12] I. M. Karo Karo, M. F. M. Fudzee, S. Kasim, and A. A. Ramli, "Sentiment Analysis in Karonese Tweet using Machine Learning," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 219–231, Mar. 2022, https://doi.org/10.52549/ijeei.v10i1.3565.
[13] I. M. K. Karo, M. F. M. Fudzee, S. Kasim, and A. A. Ramli, "Karonese Sentiment Analysis: A New Dataset and Preliminary Result," *International Journal on Informatics Visualization*, vol. 6, no. 2–2, 2022, https://doi.org/10.30630/joiv.6.2-2.1119.
[14] A. H. Ali and M. Z. Abdullah, "An efficient model for data classification based on SVM grid parameter optimization and PSO feature weight selection," *International Journal of Integrated Engineering*, vol. 12, no. 1, 2020, https://doi.org/10.30880/ijie.2020.12.01.001.
[15] A. P. Gopi, R. N. S. Jyothi, V. L. Narayana, and K. S. Sandeep, "Classification of tweets data based on polarity using improved RBF kernel of SVM," *International Journal of Information Technology (Singapore)*, vol. 15, no. 2, 2023, https://doi.org/10.1007/s41870-019-00409-4.
[16] G. S. Lumacad and R. A. Namoco, "Multilayer Perceptron Neural Network Approach to Classifying Learning Modalities Under the New Normal," *IEEE Trans Comput Soc Syst*, vol. 11, no. 1, 2024, https://doi.org/10.1109/TCSS.2023.3251566.
[17] R. Sundar and M. Punniyamoorthy, "Performance enhanced Boosted SVM for Imbalanced datasets," *Applied Soft Computing Journal*, vol. 83, 2019, https://doi.org/10.1016/j.asoc.2019.105601.
[18] L. Muflikhah, W. Widodo, W. F. Mahmudy, and S. Solimun, "A support vector machine based on kernel k-means for detecting the liver cancer disease," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 3, 2020, https://doi.org/10.22266/IJIES2020.0630.27.
[19] T. Y. Kim, H. Ko, S. H. Kim, and H. Da Kim, "Modeling of recommendation system based on emotional information and collaborative filtering," *Sensors*, vol. 21, no. 6, 2021, https://doi.org/10.3390/s21061997.
[20] H. Hasanli and S. Rustamov, "Sentiment Analysis of Azerbaijani twits Using Logistic Regression, Naive Bayes and SVM," in *13th IEEE International Conference on Application of Information and Communication Technologies, AICT 2019 - Proceedings*, 2019. https://doi.org/10.1109/AICT47866.2019.8981793.
[21] S. Fahmi, L. Purnamawati, G. F. Shidik, M. Muljono, and A. Z. Fanani, "Sentiment analysis of student review in learning management system based on sastrawi stemmer and SVM-PSO," in *Proceedings International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, iSemantic*, pp. 643-648, 2020. https://doi.org/10.1109/iSemantic50169.2020.9234291.
[22] K. Meena, R. Lawrance, S. Suresh, and M. Ahmad Khder, "Evaluation of descriptive type answer using transformed weight and Cosine-SVM," *J Stat Appl Probab*, vol. 11, no. 2, 2022, https://doi.org/10.18576/jsap/110206.
[23] Z. Rajabi, M. R. Valavi, and M. Hourali, "A Context-Based Disambiguation Model for Sentiment Concepts Using a Bag-of-Concepts Approach," *Cognit Comput*, vol. 12, no. 6, 2020, https://doi.org/10.1007/s12559-020-09729-1.
[24] Y. Li, J. Lou, X. Tan, Y. Xu, J. Zhang, and Z. Jing, "Adaptive Kernel Learning Kalman Filtering With Application to Model-Free Maneuvering Target Tracking," *IEEE Access*, vol. 10, 2022, https://doi.org/10.1109/ACCESS.2022.3193101.
[25] M. Y. Saeed *et al.*, "An abstractive summarization technique with variable length keywords as per document diversity," *Computers, Materials and Continua*, vol. 66, no. 3, 2021, https://doi.org/10.32604/cmc.2021.014330.
[26] M. A. Khder, "Web scraping or web crawling: State of art, techniques, approaches and application," *International Journal of Advances in Soft Computing and its Applications*, vol. 13, no. 3, 2021, https://doi.org/10.15849/ijasca.211128.11.

[27] N. S. Mullah and W. M. N. W. Zainon, "Improving detection accuracy of politically motivated cyber-hate using heterogeneous stacked ensemble (HSE) approach," *J Ambient Intell Humaniz Comput*, vol. 14, no. 9, 2023, https://doi.org/10.1007/s12652-022-03763-7.
[28] M. Işik and H. Dağ, "The impact of text preprocessing on the prediction of review ratings," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 3, pp. 1405-1421. 2020, https://doi.org/10.3906/elk-1907-46.
[29] J. H. Lee, M. Lee, and K. Min, "Natural Language Processing Techniques for Advancing Materials Discovery: A Short Review," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 10, no. 5, pp. 1337-1349. 2023. https://doi.org/10.1007/s40684-023-00523-6.
[30] A. G. L. Babu and S. Badugu, "Extractive Summarization of Telugu Text Using Modified Text Rank and Maximum Marginal Relevance," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 9, 2023, https://doi.org/10.1145/3600224.
[31] M. Mujahid, K. Kanwal, F. Rustam, W. Aljadani, and I. Ashraf, "Arabic ChatGPT Tweets Classification Using RoBERTa and BERT Ensemble Model," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 8, 2023, https://doi.org/10.1145/3605889.
[32] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *Journal of Data Analysis and Information Processing*, vol. 08, no. 4, 2020, https://doi.org/10.4236/jdaip.2020.84020.
[33] M. U. Nasir, S. Khan, S. Mehmood, M. A. Khan, M. Zubair, and S. O. Hwang, "Network Meddling Detection Using Machine Learning Empowered with Blockchain Technology," *Sensors*, vol. 22, no. 18, 2022, https://doi.org/10.3390/s22186755.
[34] A. Kammoun and M. S. Alouinifellow, "On the Precise Error Analysis of Support Vector Machines," *IEEE Open Journal of Signal Processing*, vol. 2, 2021, https://doi.org/10.1109/OJSP.2021.3051849.
[35] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Comput Oper Res*, vol. 152, 2023, https://doi.org/10.1016/j.cor.2022.106131.
[36] R. Kashef, "A boosted SVM classifier trained by incremental learning and decremental unlearning approach," *Expert Syst Appl*, vol. 167, 2021, https://doi.org/10.1016/j.eswa.2020.114154.
[37] F. Budiman and E. Sugiarto, "Non-linear Multiclass SVM Classification Optimization using Large Datasets of Geometric Motif Image," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 9, 2021, https://doi.org/10.14569/IJACSA.2021.0120932.
[38] M. S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, "Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset," *Computer Methods and Programs in Biomedicine Update*, vol. 4, 2023, https://doi.org/10.1016/j.cmpbup.2023.100118.
[39] N. Mohd Hatta, Z. Ali Shah, and S. Kasim, "Evaluate the Performance of SVM Kernel Functions for Multiclass Cancer Classification," *International Journal of Data Science*, vol. 1, no. 1, 2020, https://doi.org/10.18517/ijods.1.1.37-41.2020.
[40] C. B. Pande *et al.*, "Comparative Assessment of Improved SVM Method under Different Kernel Functions for Predicting Multi-scale Drought Index," *Water Resources Management*, vol. 37, no. 3, 2023, https://doi.org/10.1007/s11269-023-03440-0.
[41] P. H. Prastyo, I. Ardiyanto, and R. Hidayat, "Indonesian Sentiment Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF," in *International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy, ICDABI*, pp. 1-6, 2020. https://doi.org/10.1109/ICDABI51230.2020.9325685.
[42] Z. Gao, R. Xia, and P. Zhang, "Prediction of Anti-proliferation Effect of [1,2,3]Triazolo[4,5-d]pyrimidine Derivatives by Random Forest and Mix-Kernel Function SVM with PSO," *Chem Pharm Bull (Tokyo)*, vol. 70, no. 10, 2022, https://doi.org/10.1248/cpb.c22-00376.
[43] S. Abuhaimed, M. Al-Jasir, H. Al-Juaid, and A. Alhameidi, "Supervised Learning-Based Indoor Positioning System Using WiFi Fingerprints," in *Lecture Notes in Networks and Systems*, pp. 56-71 2023. https://doi.org/10.1007/978-3-031-33743-7_52.
[44] Y. Yuan, "Enhanced EDAS technique for colleges business English teaching quality evaluation based on Euclid distance and cosine similarity measure," *Journal of Intelligent and Fuzzy Systems*, vol. 46, no. 1, 2024, https://doi.org/10.3233/JIFS-233786.
[45] H. Öztürk, E. Ozkirimli, and A. Özgür, "A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction," *BMC Bioinformatics*, vol. 17, no. 1, 2016, https://doi.org/10.1186/s12859-016-0977-x.
[46] H. Froud, A. Lachkar, and S. A. Ouatik, "Arabic Text Summarization Based on Latent Semantic Analysis to Enhance Arabic Documents Clustering," *International Journal of Data Mining & Knowledge Management Process*, vol. 3, no. 1, 2013, https://doi.org/10.5121/ijdkp.2013.3107.
[47] M. Wibowo, C. Quix, N. S. Hussien, H. Yuliansyah, and F. D. Adhinata, "Similarity Identification of Large-scale Biomedical Documents using Cosine Similarity and Parallel Computing," *Knowledge Engineering and Data Science*, vol. 4, no. 2, 2022, https://doi.org/10.17977/um018v4i22021p105-116.

[48] I. M. K. Karo, R. Ramdhani, A. W. Ramadhelza, and B. Z. Aufa, "A Hybrid Classification Based on Machine Learning Classifiers to Predict Smart Indonesia Program," in *Proceeding - 3rd International Conference on Vocational Education and Electrical Engineering: Strengthening the framework of Society 5.0 through Innovations in Education, Electrical, Engineering and Informatics Engineering, ICVEE*, pp. 1-5, 2020. https://doi.org/10.1109/ICVEE50212.2020.9243195.
[49] N. E. Ramli, Z. R. Yahya, and N. A. Said, "Confusion Matrix as Performance Measure for Corner Detectors," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 29, no. 1, 2022, https://doi.org/10.37934/araset.29.1.256265.
[50] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit*, vol. 91, 2019, https://doi.org/10.1016/j.patcog.2019.02.023.
[51] O. V. Putra, F. M. Wasmanson, T. Harmini, and S. N. Utama, "Sundanese Twitter Dataset for Emotion Classification," in *CENIM Proceeding: International Conference on Computer Engineering, Network, and Intelligent Multimedia*, pp. 391-395, 2020. https://doi.org/10.1109/CENIM51130.2020.9297929.

BIOGRAPHY OF AUTHORS

Abil Mansyur, was born in Gunting Saga on September 6, 1972. He is Associate Professor at Undergraduate Mathematics Study Program, the Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan. He is also Vice Rector for Academic Affairs at Universitas Negeri Medan for the term 2023–2027. He earned a Bachelor of Science (S.Si) degree in Mathematics from Universitas Sumatera Utara in 1997, a Master of Science (M.Si) degree in Mathematics from Universitas Gadjah Mada in 2005, and a Doctorate (Dr.) in Mathematical Sciences from Universitas Sumatera Utara in 2014.
Email: abilmansyur@unimed.ac.id.

Ichwanul Muslim Karo Karo, was an Assistant Professor at computer science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan. He obtained a Ph.D from Universiti Tun Hussein Onn Malaysia in 2024. He is active in various research fields, such as Data mining, spatial mining, sentiment analysis and business intelligence. Besides, he is also actively involved as a data scientist and consultant for various projects in various institutions/organizations.
Email: ichwanul@unimed.ac.id.
Orcid ID: 0000-0002-2824-5654

Muliawan Firdaus, was born in Batangkuis on May 22, 1979. He is an Assistant Professor and lecturer at Teacher Professional Education Program (*Program Profesi Guru/PPG*), Universitas Negeri Medan. He earned a Bachelor of Education (S.Pd) degree in Mathematics Education from Universitas Muhammadiah Sumatera Utara, a Master of Science (M.Si) degree in Mathematics from Universitas Sumatera Utara in 2004, and completed his Doctorate (Dr.) in Mathematics at Universitas Sumatera Utara in 2024. He is actively involved in teaching and research in the field of mathematics education.
Email: muliawanfirdaus@unimed.ac.id.
Orcid ID: 0000-0003-3561-4486.

Elmanani Simamora, was born on November 16, 1972. He holds a Master of Science (M.Si) degree in Mathematics and a Doctorate (Dr.) in Mathematics, both from Universitas Gadjah Mada, with his doctoral degree completed in 2016. He is a lecturer at the Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan, assigned to the Undergraduate Mathematics Study Program. Holding the academic title of Lektor and also serves as the Secretary of the Doctoral Program in Mathematics Education at the Postgraduate Program of Universitas Negeri Medan.
Email: elmanani_simamora@unimed.ac.id.
Orcid ID: 0000-0002-0409-349X.

Muhammad Badzlan Darari, was born in Medan on March 25, 1987. He earned a Bachelor of Education (S.Pd) degree in Mathematics Education and a Master of Education (M.Pd) degree in Mathematics Education from Universitas Negeri Medan, completing his master's degree in 2012. He is a lecturer at the Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan, assigned to the Undergraduate Mathematics Education Study Program. He holds the academic rank of Lektor and is actively engaged in teaching and educational development in the field of mathematics education.
Email: badzlan@unimed.ac.id.
Orcid ID: 0009-0008-2628-1193.

Rizki Habibi, was born in Alur Dua on February 25, 1985. He earned a Bachelor of Education (S.Pd) degree in Mathematics Education from Universitas Negeri Medan in 2007 and later obtained a Master of Science (M.Si) degree in Mathematics from Universitas Sumatera Utara in 2016. Currently, He is a lecturer at the Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan, assigned to the Undergraduate Mathematics Study Program since 2023. In addition to his teaching responsibilities, he is pursuing a Doctorate in Mathematics at Universitas Sumatera Utara. His academic title is Lektor, and his research interests lie in the field of Applied Mathematics.
Email: rizki@unimed.ac.id.
Orcid ID: 0009-0007-2026-7692.

Suvriadi Panggabean, was born on April 18, 1988. He earned a Bachelor of Education (S.Pd) degree in Mathematics Education from Universitas Negeri Medan and a Master of Education (M.Pd) degree in Mathematics Education from Universitas Sumatera Utara, graduating in 2014. He is currently a lecturer at the Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan, assigned to the Undergraduate Mathematics Study Program, and holding the academic title of Lektor.
He can be contacted via email at suvriadi@unimed.ac.id.
Orcid ID: 0009-0009-9556-3984