Matrix: Jurnal Manajemen Teknologi dan Informatika
Volume 12 Issue 3 Year 2022 Pages 109-116
DOI: http://dx.doi.org/10.31940/matrix.v12i3.109-116
URL: http://ojs2.pnb.ac.id/index.php/matrix

# Topic modeling and sentiment analysis about Mandalika on social media using the latent Dirichlet allocation method

***Veny Cahya Hardita [1], Rifqi Hammad [2]*, Ahmad Zuli Amrullah [3]***

[1] *STMIK Palangkaraya, Indonesia*
[2,3] *Universitas Bumigora, Indonesia*

*****Corresponding Author: *rifqi.hammad@universitasbumigora.ac.id*

**Abstract:** The rapid and widespread dissemination of information currently affects the tourism sector. One tourist area that is quite widely discussed is the Mandalika Circuit. Twitter is one platform that provides comments related to the Mandalika Circuit. The amount of information related to the Mandalika Circuit is currently not being utilized properly by managers (government or private). It causes many topics related to the Mandalika Circuit that are currently trending, and public sentiment regarding the Mandalika Circuit is unknown to the government or private sector. Ignorance can result in delays in decision making which can harm the manager. To overcome this problem, research on sentiment analysis and topic modeling related to the Mandalika Circuit was carried out. The sentiment analysis method used is SVM and for modeling using LDA. Based on the results of the sentiment analysis, 1500 tweets were obtained before doing the pre-processing process, thus getting a dataset of 500 tweets divided into 398 positive and 102 negative tweets. So it can be concluded that more Twitter users give positive than negative responses to the Mandalika Circuit. The test results show that the SVM algorithm can classify sentiment toward the Mandalika Circuit well, as indicated by the measurement of the performance of the SVM algorithm, namely 87% accuracy, 77% precision, 84.81% recall, and 98.52% specificity. These results also show that the F1 Score compares the average precision and recall, which is weighted at 80.72%.

**Keywords:** LDA, Mandalika Circuit, sentiment analysis, SVM, topic modeling

**History Article:** Submitted 23 October 2022 | Revised 26 October 2022 | Accepted 17 November 2022

**How to Cite:** V. H. Hardita, R. Hammad, and A. Z. Amrullah, "Topic modeling and sentiment analysis," *Matrix: Jurnal Manajemen Teknologi dan Informatika*, vol. 12, no. 3, pp. 109-116, 2022.

## Introduction

Tourism is one of the leading sectors in the development of the island of Lombok, especially the development of tourist areas, culture, facilities, and infrastructure. Mandalika as a super-priority tourism destination is one of the main focuses of national tourism [1]. The number of tourist destinations owned by the island of Lombok, from beach tourism to mountains, and now the construction of the MotoGP Circuit in Kuta Mandalika provides a positive trend for NTB tourism. The public showed enthusiasm by providing information and comments on social media about MotoGP [2].

In the era of rapid technological development as it is today, it provides easier access for the public to provide information [3]. It certainly affects various sectors, one of which is the tourism sector. Visitors to tourist attractions can provide comments in the form of praise, criticism, or suggestions for the attractions visited. One of the social media that is widely used to provide comments is Twitter [4]. Twitter is a social media platform that allows users to send and read 140 characters, commonly called tweets [5]. With the existence of Twitter, it is wide open for information and comments from visitors and sentiment [6] on the Mandalika Circuit.

The amount of social information on Twitter related to the Mandalika Circuit has not been utilized properly by the government or private parties who manage the Mandalika Circuit area. It causes many topics related to the Mandalika Circuit, which are currently trending unknown to the government and the private sector. Ignorance can result in decision-making delays, harming the management (government and private). From the MotoGP event to date, more than 1500 tweets related to the Mandalika Circuit have been collected. The tweets contained positive

and negative comments about the experiences and observations of visitors and the public about the Mandalika Circuit. However, the number of tweets with various topics discussed makes it difficult to identify community sentiments related to the Mandalika Circuit.

To overcome these problems, research on sentiment analysis and topic modeling related to the Mandalika Circuit was carried out. Sentiment analysis is a study used to help identify comments related to something [7]. One method used in sentiment analysis is the Support Vector Machine (SVM) [8]. SVM is a learning model which it has defined inputs and outputs. SVM proved to be the best algorithm for text categorization [9].

Topic modeling is one approach to text mining used to find data in text and find the relationship between one text and another from a corpus [10] [11]. One method that can be used in optical modeling is Latent Dirichlet Allocation (LDA) [12]. LDA is a method that can be used to group topics based on the probability of words in a topic [13].

This study proposes using the Support Vector Machine (SVM) method for sentiment analysis and the Latent Dirichlet Allocation (LDA) method for topic modeling or extraction. The result of this research is a prototype or initial model from the results of data exploration in the form of sentiment analysis and Topic Modeling. The results of this study are expected to be used as supporting data for the government and the private sector in determining policies and making decisions related to the development of the tourism sector that is relevant to the needs of tourists, such as policies for promotion, improvement of facilities, and so on.

Several previous studies support this research, such as the research conducted by Chotijah with the title "Media Relations of the Mandalika Superpriority Destination in Reporting on the Readiness of the 2021 MotoGP Event Implementation", utilizing various media through exposure and approaching through framing and agendas on the distribution of information. In addition, the organizers communicated positive messages to the public in a strategic and planned manner to create optimism for the Mandalika MotoGP event [1].

Research conducted by Siswanto, Wibawa, Gata, and Kusumawardhani in a paper published at the 2018 ICAITI conference discusses the analysis of the classification of comments related to MotoGP on Twitter using the Support Vector Machine and Naïve Bayes algorithms. The results show that the accuracy of the SVM algorithm is 95.50%, and Naïve Bayes has an accuracy of 93% [2].

Research conducted by Annisa, Surjandari, and Zulkarnain in 2019 showed opinions on Mandalika Hotel reviews using Latent Dirichlet Allocation. In their paper, they extracted eight topics from tourism keywords and reviews, where the review or public opinion relates to complaints, experiences, opinions, and hotel management responses are written on Traveloka [14]. And there are many other studies, such as research conducted by Merawati and Amrullah [15], Dwi and Adri [16], Zuriel et al. [17], and others.

## Methodology

The method used in this research is SVM for sentiment analysis and LDA for topic modeling. The stages in this research can be seen in Figure 1.


**Figure 1.** Research flow

Figure 1 shows the stages in this research, which consist of collecting data, pre-processing data, sentiment analysis, topic modeling, and analysis and evaluation.

1. Data Collecting
   The data used in this study comes from social media Twitter, so it is necessary to carry out a data crawling process. Therefore, the data taken only contains content related to the Mandalika Circuit.
2. Pre-processing Data

Because the data comes from social media, most of the data does not have a definite structure, so the information cannot be extracted directly, so pre-processing is needed. Pre-processing aims to change the form of unstructured data into structured data. The stages of pre-processing Twitter data are [18]:

a. Case Folding aims to convert all letters in the document to lowercase.
b. Tokenization is the stage of cutting the input string based on each word that composes it. This process removes numbers, punctuation, characters, HTML tags, links, scripts, etc.
c. Filtering the stage of taking important words from the tokenization results.
d. Phrase detection aims to detect the presence of 2 or more words that are the same.
e. Stemming aims to change suffixes into base words.

3. Sentiment Analysis
The next process is sentiment analysis. The pre-processed document will be labeled positive, negative, and neutral, depending on the opinion expressed in the document. The classification method used in this study is SVM.
4. Topic Modeling
Classified data with two categories, namely positive and negative sentiment, will be modeled into topic modeling using the LDA method to conclude the topics hidden in the document. Stages in LDA [19] :
   a. The first step is to initialize parameters such as the number of documents, the number of words in a document, the number of topics, the number of iterations, and the LDA coefficient.
   b. Marks a word with a predetermined topic in a semi-random distribution based on the Dirichlet distribution.
   c. The iteration stage will produce parameters that can determine the distribution of the number of topics in the document and the distribution of words from topics.
   d. Validate the model topic by analyzing the right number of iterations to form the model, the appropriate number of topics based on the perplexity distribution, and the probability distribution of each learning document on the formed topic model.
5. Analysis and Evaluation

Topic analysis can be done by observing the entire distribution of topics, including the distribution of words in the topic. Then the topic coherence test is carried out to test the ease of the model topic being interpreted by humans.

## Results and Discussions

The data used in this study is tweet data taken on Twitter of approximately 1500 tweets where this tweet was taken from the Mandalika MotoGP event until July 2022 with the keyword "Mandalika Circuit". The data that has been collected then enters the pre-processing stage. An example of this stage can be seen in Table 1.

**Table 1**. Pre-processing data

| Step | Before | After |
|---|---|---|
| Case Folding | The Mandalika Circuit is stunning and beautiful | the Mandalika Circuit is stunning and beautiful |
| Tokenization | the Mandalika Circuit is stunning and beautiful | 'the', 'Mandalika', 'Circuit', 'is', 'stunning', 'and', 'beautiful' |
| Filtering | the Mandalika Circuit is stunning and beautiful | 'Mandalika', 'Circuit', 'stunning, 'beautiful' |
| Phrase detection | the Mandalika Circuit is stunning and beautiful | 'Mandalika', 'Circuit', 'stunning, 'beautiful' |
| Stemming | the Mandalika Circuit is stunning and beautiful | 'Mandalika', 'Circuit', 'amazed', 'beautiful' |

Table 1 shows the pre-processing data stage on all data sets used in this study. After going through the pre-processing stage, it is known that not all tweet data collected can be used as a dataset. The amount of data that can qualify as a dataset is 500 tweets with a distribution of 398 positive sentiments and 102 negative sentiments with an inappropriate amount of data or data

imbalance for each category, both positive and negative sentiments. Therefore, under-sampling is applied to balance the data by dividing 100 positive and 100 negatives data.

The data set from the classification of positive and negative sentiments are then visualized as a word cloud. For example, the positive sentiment word cloud can be seen in Figure 2.

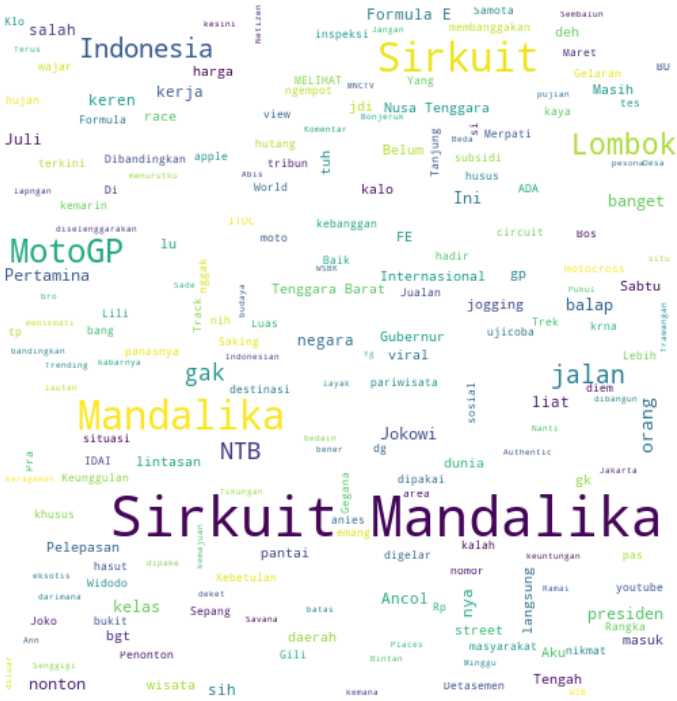


**Figure 2.** Positive tweet visualization

Figure 2 shows a positive sentiment word cloud where the words often appear for positive sentiment are Circuit Mandalika, Mandalika, Circuit, MotoGP, NTB, Lombok, Indonesia, and Pertamina. As for negative sentiments, words that often appear are Mandalika Circuit, Mandalika, Circuit, no, asphalt, preseason, and international. The negative sentiment word cloud can be seen in Figure 3.

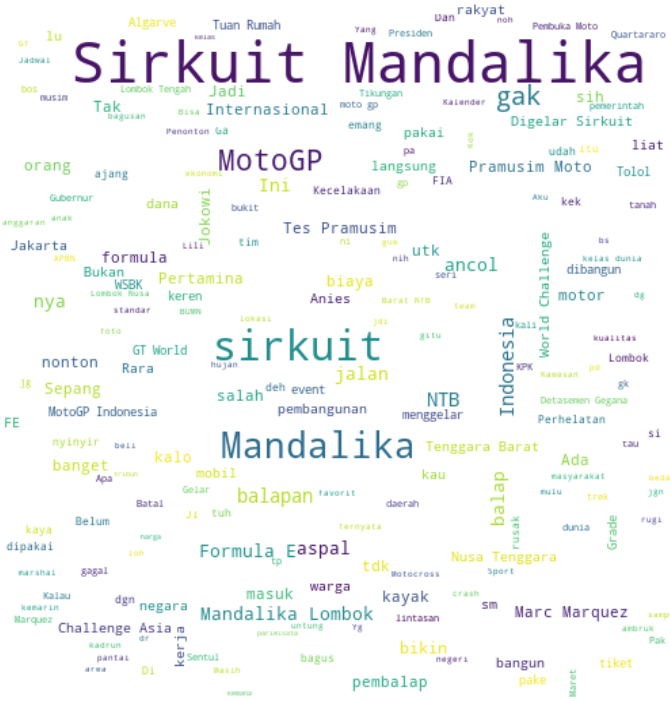


**Figure 3.** Negative tweet visualization

The classification method applied to the created dataset is SVM (Support Vector Machine). In doing the classification required training data and data testing. This training data consists of 200 tweets used to model the SVM algorithm. In contrast, data testing is needed to test the performance of the previously created model. The training process will mathematically determine the prior and term values of each data being trained and how likely it is to enter a certain class.

Tests in this study will use a confusion matrix where the test data can be known for accuracy, precision, recall [20], and specificity. The testing process is carried out with 100 tweets of test data. Based on the test results, the following results were obtained, 87 positive tweets and 13 negative tweets. In addition to model predictions, the results obtained were 67 True Positive, 20 True Negatives, 1 False Positive, and 12 False Negatives, as shown in Table 2 below.

**Table 2.** Confusion matrix

| Prediction | Actual | |
|---|---|---|
| | Positive | Negative |
| Positive | 67 | 1 |
| Negative | 20 | 12 |

Based on the calculations concerning the table above, the Accuracy Value: is 87.00, while the Precision Value is 77.01, Recall Value is 84.81, Specificity Value is 98.52, and the F1 Value is 80.72. With this, it can be said that the SVM algorithm has fairly good accuracy for classifying sentiment data.

Exploration of sentiment data using the Latent Dirichlet Allocation (LDA) method is carried out on positive and negative classes so that the topic modeling in this study is used to obtain information from the collection of opinions of tourists who have visited the Mandalika Circuit. The information will be interpreted as a collection of main topics from the tweets obtained.

This research will take ten topics to be tested to find the best group of topics. Topics resulting from topic modeling are not necessarily easy to interpret. Therefore, a topic coherence calculation is carried out to distinguish between good and bad topics. Overall, the coherence value obtained from modeling this topic is 0.473. The results of the ten topics generated are shown in the following Table 3.

**Table 3.** Topic

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 | Topic 9 | Topic 10 |
|---|---|---|---|---|---|---|---|---|---|
| Circuit | Circuit | Mandalika | Circuit | Mandalika | Circuit | Circuit | Circuit | Circuit | Circuit |
| Mandalika | Mandalika | Circuit | Mandalika | Circuit | Mandalika | Mandalika | Mandalika | Mandalika | Mandalika |
| moto | no | racer | kayak | marquez | built | lombok | no | racing | motogp |
| preseason | asphalt | rich | there is | marc | held | ntb | Jokowi | ticket | test |
| for | negara | Indonesia | Sepang | Honda | time | Street | racer | world | roll call |
| no | inhabitant | formula | motogp | accident | formula | kpk | cost | Asia | deposit |
| develop-ment | Jakarta | no | direct | race | president | interna-tional | budget | race | season |
| land | motogp | motor | damaged | motogp | formula | Indonesia | BUMN | not | wear |
| fail | Ancol | take | more | make | test | Pertamina | peeled off | challenge | no |
| asphalt | formula | bend | no | paved | WSBK | motogp | collapse | car | ntb |

Table 3 shows ten topics in this study. Each topic has its word list. The topics that have been obtained are then visualized using pyLDAvis. The visualization can be seen in Figure 4.

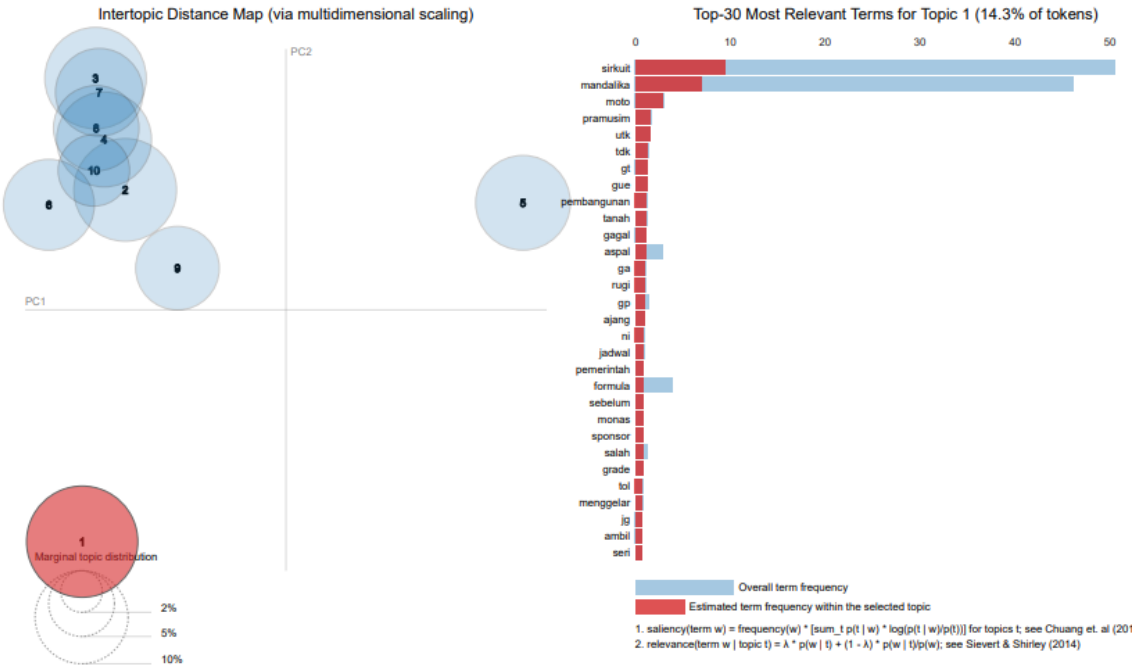


**Figure 4.** Topic visualization with pyLDAvis

Figure 4 shows the results of the visualization of all topics using pyLDAvis. The figure shows two different colors, namely blue and red. The blue color shows the frequency of occurrence of the word in all documents, while the red color shows the frequency of occurrence of the word in each topic. Figure 4 also shows the existence of circles (topics) that intersect and do not intersect. Intersecting topics indicate that several words are the same in each topic, while those that do not intersect indicate that there are no words in the topic that are the same as other topics. The bigger the circle, the higher the frequency of the topic. In this study, the words often appear are circuit, Mandalika, moto, and formula.

## Conclusion

This study succeeded in conducting sentiment analysis using the Support Vector Machine (SVM) method and Topic modeling with the Latent Dirichlet Allocation (LDA) method for Mandalika Circuits. Based on the sentiment analysis results, 1500 tweets were obtained before the pre-processing process, thus getting a dataset of 500 tweets divided into 398 positive tweets and 102 negative tweets. So it can be concluded that more Twitter users give positive than negative responses to the Mandalika Circuit. The test results show that the SVM algorithm can classify sentiment toward the Mandalika Circuit well, as indicated by the measurement of the performance of the SVM algorithm, namely the accuracy value of 87%, the precision of 77%, recall of 84.81%, and specificity of 98.52%. These results also show that the F1 Score is a comparison of the average precision and recall, which is weighted at 80.72%. In addition, based on the research conducted, it was found that the words that most often appear are circuit, Mandalika, motto, and formula.

## Acknowledgments

We thank the Ministry of Education, Culture, Research, and Technology for funding this research with a novice lecturer research grant (PDP). We also thank LLDIKTI XI and UP3M STMIK Palangka Raya for facilitating this research.

## References

[1] S. Chotijah, "Relasi media mandalika sebagai destinasi superprioritas dalam pemberitaan kesiapan pelaksanaan event MotoGP 2021," *JCommsci - J. Media Commun. Sci.*, vol. 4, no. 1, pp. 14–22, 2021.

[2] Siswanto, Y. P. Wibawa, W. Gata, G. Gata, and N. Kusumawardhani, "Classification analysis of MotoGP comments on media social twitter using algorithm support vector machine and naive Bayes," *Proc. ICAITI 2018 - 1st Int. Conf. Appl. Inf. Technol. Innov. Towar. A New Paradig. Des. Assist. Technol. Smart Home Care*, pp. 96–101, 2018.

[3] I. Heliany, "Wonderful digital tourism Indonesia dan peran revolusi industri dalam menghadapi era ekonomi digital 5.0," *Destin. (Journal Hosp. Pariwisata)*, vol. 1, no. 1, 2018.

[4] A. R. Rinaldi, D. Mutiarin, and J. Damanik, "Analisis netnografi sentimen pengguna Twitter terhadap pembukaan kembali pariwisata di tengah pandemi Covid," *Pariwisata Budaya J. Ilm. Pariwisata Agama dan Budaya*, vol. 6, no. 1, pp. 27–36, 2021.

[5] F. N. Hikmah, S. Basuki, and Y. Azhar, "Deteksi topic tentang tokoh publik politik menggunakan latent Dirichlet allocation (LDA)," *J. Repos.*, vol. 2, no. 4, p. 415, 2020, doi: 10.22219/repositor.v2i4.52.

[6] M. R. Firdaus, F. M. Rizki, F. M. Gaus, and I. K. Susanto, "Analisis sentimen dan topic modeling dalam aplikasi Ruangguru," *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 4, no. 1, p. 66, 2020, doi: 10.30645/j-sakti.v4i1.188.

[7] S. Fanissa, M. A. Fauzi, and S. Adinugroho, "Analisis sentimen pariwisata di Kota Malang menggunakan metode naive Bayes dan seleksi fitur query expansion ranking," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 8, pp. 2766–2770, 2018.

[8] M. Tripathi, "Sentiment analysis of Nepali COVID-19 tweets using NB, SVM, and LSTM," *J. Artif. Intell. Capsul. Networks*, vol. 3, no. 3, pp. 151–168, 2021.

[9] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of naive Bayes and SVM algorithm based on sentiment analysis using review dataset," in *8th International Conference on System Modeling & Advancement in Research Trends*, 2019, pp. 266–270.

[10] Y. H. Kee, C. Li, L. C. Kong, C. J. Tang, and K. L. Chuang, "Scoping review of mindfulness research: a topic modeling approach," *Mindfulness 2019 108*, vol. 10, no. 8, pp. 1474–1488, Apr. 2019.

[11] T. Heidenreich, F. Lind, J. M. Eberl, and H. G. Boomgaarden, "Media framing dynamics of the 'European refugee crisis': a comparative topic modeling approach," *J. Refug. Stud.*, vol. 32, no. Special_Issue_1, pp. i172–i182, Dec. 2019.

[12] U. Chauhan and A. Shah, "Topic modeling using latent Dirichlet allocation," *ACM Comput. Surv.*, vol. 54, no. 7, Sep. 2021.

[13] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimed. Tools Appl. 2018 7811*, vol. 78, no. 11, pp. 15169–15211, Nov. 2018.
[14] R. Annisa, I. Surjandari, and Zulkarnain, "Opinion mining on Mandalika hotel reviews using latent Dirichlet allocation," *Procedia Comput. Sci.*, vol. 161, pp. 739–746, 2019.
[15] N. L. P. M. Putu, A. Z. Amrullah, and Ismarmiaty, "Analisis sentimen dan pemodelan topic pariwisata lombok menggunakan algoritma naive bayes dan latent dirichlet allocation," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 123–131, 2021.
[16] D. D. A. Nugroho and A. Alamsyah, "Analisis konten pelanggan Airbnb pada network sosial media Twitter content analysis of airbnb customer based on twitter social media," in *E-Proceeding of Management*, 2018, pp. 1622–1628.
[17] H. P. P. Zuriel and A. Fahrurozi, "Implementasi algoritma klasifikasi support vector machine untuk analisa sentimen pengguna Twitter terhadap kebijakan PSBB," *J. Ilm. Inform. Komput.*, vol. 26, no. 2, pp. 149–162, 2021.
[18] N. L. P. Merawati and Ahmad Zuli Amrullah Ismarmiaty, "Analisis sentimen dan pemodelan topic pariwisata Lombok menggunakan algoritma naive Bayes dan latent dirichlet allocation," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 123–131, 2021.
[19] Y. Sahria and D. H. Fudholi, "Analysis of health research topics in Indonesia using the LDA (Latent Dirichlet Allocation) topic modeling method," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 2, pp. 336–344, 2020.
[20] R. Hammad, A. C. Nurcahyo, A. Z. Amrullah, P. Irfan, and K. A. Latif, "Optimization of data integration using schema matching of linguistic-based and constraint-based in the university database," *J. Manaj. Teknol. dan Inform.*, vol. 11, no. 3, pp. 119–129, 2021.