


Sentiment Analysis of Bjorka Hacker Using the Naive Bayes and C.45 Algorithms

^{1*}Wowon Priatna , ²Eka Nur A'ini, ³Joni Warta, ⁴Agus Hidayat, ⁵Tyastuti Sri Lestari, ⁶Rasim
^{1,2,3,4,5}Faculty of Computer Science, Informatics, Bhayangkara University of Jakarta, Indonesia
E-mail: ^{1*}wowon.priatna@dsn.ubharajaya.ac.id, ²kanurainii240@gmail.com,
³joniwarta@dsn.ubharajaya.ac.id, ⁴agus.hidayat@dsn.ubharajaya.ac.id, ⁵tyas@ubharajaya.ac.id
***Corresponding author**

(Received September 30, 2023 Revised November 19, 2023 Accepted November 26, 2023, Available online December 19, 2023)

Abstract

In 2023, Indonesia was again devastated by a hacker known as Bjorka. Bjorka did not act just once or twice; every time, Bjorka made the entire Indonesian population proud. The 19 million BPJS Employment data belonging to the Indonesian people that Bjorka hacked is the BPJS Employment data belonging to the Indonesian people that Bjorka hacked. Since the release of the Bjorka story, there has been a surge in the number of people criticizing it on social media, particularly Facebook, so the criticism or opinions can be used to conduct sentiment analysis. Based on this, developing a method that can automatically classify beliefs into positive and negative categories through sentiment analysis is necessary. The sentiment analysis process begins with data preprocessing, followed by keyword analysis using the TF-IDF method, algorithm development, and analysis of classification results. The data classification methods used in this study are Naive Bayes and C4.5. The data will be analyzed using text mining and classified using the Naive Bayes and C4.5 algorithms. Based on the results of the tests, the best classification was achieved by Nave Bayes, with a score of 70 percent for the C4.5 algorithm and 68 percent for the C4.5 algorithm. The Nave Bayes algorithm can predict up to 70% data transmission rates for both positive and negative signals.

Keywords: BPJS, Hacker Bjorka, Classification, Sentiment Analysis, C.45, Naive Bayes

This is an open access article under the [CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/).
Copyright © 2023 IAIC - All rights reserved.



1. Introduction

Data security has become a critical component in the evolution of information technology. The use of information technology has impacted a variety of areas, including data collection and storage on a cloud-based basis [1]. Data breaches occur frequently due to advancements in technology or an individual's ability to detect a violation, commonly referred to as a hacker [2].

One of the hackers who is talked about quite a lot by people in Indonesia is hacker Bjorka [2]. Hacker Bjorka is reported to have leaked 11 GB of Tokopedia marketplace data, 26 million Indihome customer data, 1.3 billion SIM card registration data, 105 million KPU data [3][4], and recently Bjorka has leaked BPJS Employment data [4]. This has given rise to various opinions on social media, especially Facebook [5].

Opinions from the public are very diverse; some are amazed by Bjorka's figure, and some do not like it or comment that this is just a diversion of issues [6]. The large and ever-increasing number of opinions is a challenge in describing public sentiment, so an analytical approach to public views is needed. Text mining is one approach that can be applied to overcome this problem [7].

Previously, research on Bjorka was conducted using a Support Vector Machine, yielding accuracy results of 62.33 percent [8]. Furthermore, similar analysis using the Naive Bayes algorithm produced an accuracy score of 83.06% [9]. Another study compared the accuracy of Naive Bayes, C4.5, and Random Forest for classifying online motorcycle taxi services [10]. The comparison results showed that Naive Bayes accuracy was better than C4.5 and Random Forest accuracy with an average Naive Bayes accuracy of 69.18%, while Random Forest was 69.18%: 66.34% and C4.5 of 65% [11][12].

This sentiment analysis research will compare accuracy using the Naive Bayes and C4.5 algorithms [13][14]. This research adds TF-IDF feature weighting; with this additional method, it is hoped that the system will be able to classify sentiment better and produce better accuracy [15][16].

2. Research Method

2.1. Research Design

The process in this research involves the application of the Knowledge Discovery in Database method [17]. KDD is a step in extracting potential, implied, and previously unknown information from a data set [18]. The workflow of the research can be seen in Figure 1.



Figure 1. Research Design

2.2. Dataset

This study relied on 1,187 data points obtained from Facebook social media. The collected data is then manually labeled with the assistance of a linguist to determine positive and negative sentiments [19].

2.3. Term Frequency Inverse Document Frequency (TF-IDF) Weighting

TF-IDF is a feature extraction method that gives a value to each word in the training dataset. The TF-IDF approach gives a score based on how often words appear in the document [20][21]. TF-IDF calculations can be done using equation (1).

$$f - idf = tf \times \log\left(\frac{n}{df}\right) \quad (1)$$

2.4. Naive Bayes

Naive Bayes is a technique that can be applied to classify data. Naive Bayes Classification Approach is a statistical method used to estimate a class's membership probability [22]. Naive Bayes calculations can be done using the equation (2).

$$MAP = \underset{v \in V}{\operatorname{argmax}} \prod_{i=1}^n (P(x_i | V_j) P(v_j)) \quad (2)$$

$(Px_i|V_j)$ and $P(v_j)$ Calculated during training, where the equation can be seen in equations (3) and (4).

$$P(v_j) = \frac{|docs_j|}{|Example|} \tag{3}$$

$$(Px_i|V_j) = \frac{|n_k+1|}{n+|vocabulary|} \tag{4}$$

2.6. C.45

The C4.5 algorithm is the result of the development of the ID3 algorithm with various improvements and enhancements [23]. Some improvements include handling numeric attributes, managing missing values, and reducing noise in data sets [24]. The C4.5 calculation can be done using equation (5).

$$Entropy(S) = \sum_{i=1}^n - P_i \text{Log}_2 P_i \tag{5}$$

After the entropy is calculated, attribute selection is done using Information Gain [25]. Information Gain calculations can be carried out using equation (6).

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n * Entropy(S_i) \tag{6}$$

2.7. Evaluation

The final stage is the evaluation of Classification performance [26]. Based on the accuracy value, which trains how often the model is produced correctly, it is described using a confusion matrix [27]. Classification evaluation also measures its performance using recall and precision [28]. To calculate the accuracy value (7), it is found in equation, precision equation (8), recall equation (9). Besides using the confusion matrix, whether the prediction results are good or bad, a classification model can also use the Receiver Operating Characteristic (ROC) [29] and dan Area Under the Curve (AUC) [30].

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \tag{7}$$

$$Precision = (TP)/(TP+FP) \tag{8}$$

$$Recall = (TP) / (TP+FN) \tag{9}$$

Where TP= True Positive, TN=True Negative, FP=False Positive and FN = False Negative.

3. Results and Analysis

3.1. Text Preprocessing

In this stage, the data that has been labeled is then carried out by a data cleaning process. The stages of text preprocessing carried out can be seen in Table 1.

Table 1. Preprocessing Stages

Stage	Comment
Original Comment	In my opinion, Bjorka is a hacker who is kept by certain individuals with certain goals and objectives
Case Folding	In my opinion, Bjorka is a hacker who is kept by certain individuals with certain goals and objectives
Tokenizing	'in my opinion', 'bjorka', 'is', 'hacker', 'the', 'in' 'maintained', 'by', 'person', 'certain', 'with', 'purpose', 'and', 'intended', 'certain'
Filtering/Stopword Removing	According to Bjorka, hackers maintain individual goals
Stemming	Also, hackers protect individuals who aim at their goals

3.2. Transformation

At this stage, the data that has been labeled will then be calculated by weighting each word. The Term Frequency Inverse Document Frequency weight calculation uses equation (1). The results of the Term Frequency Inverse Document Frequency calculation can be seen in Table 2.

Table 2. TF-IDF calculation

Term	TF			DF	D/DF	IDF	IDF+1	W=TF*(IDF+1)	
	D1	D2	D3					D1	D2
Hurry	1	0	0	1	3	0,4771213	1,4771213	1,4771213	0
Sell	1	0	0	1	3	0,4771213	1,4771213	1,4771213	0
People	1	0	0	1	3	0,4771213	1,4771213	1,4771213	0
Privacy	1	0	0	1	3	0,4771213	1,4771213	1,4771213	0
Switch	0	1	1	2	1,5	0,1760913	1,1760913	0	1,1760913
Issue	0	1	1	2	1,5	0,1760913	1,1760913	0	1,1760913
Ministry of Finance	0	1	0	1	3	0,4771213	1,4771213	0	1,4771213
Order	0	1	0	1	3	0,4771213	1,4771213	0	1,4771213
Complete	0	0	1	1	3	0,4771213	1,4771213	0	0
Tax	0	0	1	1	3	0,4771213	1,4771213	0	0

3.3. Modeling

After the TF-IDF weight results are obtained, the next stage is implementing the Naive Bayes and C.45 algorithms. Examples of training and test data samples can be seen in Table 3 and Table 4.

Table 3. Sample Testing Data

D	Comment	Categorization
D1	Hurry to Sell People's Privacy	negative
D2	Transferring the Issue of the Ministry of Finance Case Order	negative
D3	Transferring Tax Issues	positive

Table 4. Sample and Testing

D	Comment	Category
D1	Transferring Tax Issues	?

3.1.1. Naive Bayes

The first stage in Naive Bayes calculations is to calculate the prior probability. Prior probability calculations can be done using equation (3), and can be seen in Table 5.

Table 5. Prior Probability

	$P_{(positive)}$	$P_{(negative)}$
Prior Probability	$\frac{4}{12} = 0.333$	$\frac{8}{12} = 0.66667$

After getting the prior probability results, the next stage calculates the likelihood. The likelihood calculation can be done using equation (4) and can be seen as follows.

Word probability 'switch':

$$P(a_{switch} | V_{positive}) = \frac{1+1}{4+12} = \frac{2}{16} = 0.125$$

$$P(a_{switch} | V_{negative}) = \frac{1+1}{8+12} = \frac{2}{20} = 0.1$$

Word Probability 'issue'

$$P(a_{issue} | V_{positive}) = \frac{1+1}{4+12} = \frac{2}{16} = 0.125$$

$$P(a_{issue} | V_{negative}) = \frac{1+1}{8+12} = \frac{2}{20} = 0.1$$

Word Probability 'tax'

$$P(a_{tax} | V_{positive}) = \frac{1+1}{4+12} = \frac{2}{16} = 0.125$$

$$P(a_{tax} | V_{negative}) = \frac{0+1}{8+12} = \frac{1}{20} = 0.05$$

After obtaining the likelihood probability results, the next stage is to classify the test data. Manual calculations can be calculated using equation (2) to classify test data.

$$P(test | positive)$$

$$= P(positive) \times P(switch | positive) \times P(issue | positive)$$

$$\times P(tax | positive)$$

$$= 0.333 \times 0.1 \times 0.1 \times 0.5$$

$$= 0.000650390625$$

$$P(test | negative)$$

$$= P(negative) \times P(switch | negative) \times P(issue | negative)$$

$$\times P(tax | negative)$$

$$= 0.66667 \times 0.1 \times 0.1 \times 0.5$$

$$= 0.00333335$$

From the results of these calculations, it can be concluded that the test data for 'tax issue experts' falls into the category of **negative**.

3.1.2. C.45

The next stage is calculating the C4.5 algorithm. It is known that the training data can be seen in Table 6.

Table 6. Sample Training Data

D	Data	Class
D1	Hurry to Sell People's Privacy	negative
D2	Responsibility Order	positive
D3	Transferring the Issue of the Ministry of Finance Case Order	negative
D4	Transferring Tax Issues	negative
D5	Transfer of Responsibility Order Issues	negative

After that, the value of each term of the training data is calculated. The calculations can be seen in Table 7

Table 7. Training Data Values

D	Hurry	Sell	People	Privacy	Order	Guaranteed	Switch	Issue	Case	Ministry of Finance	Tax	Answer	Amount
d1	1	1	1	1	0	0	0	0	0	0	0	0	4
d2	0	0	0	0	1	1	0	0	0	0	0	0	2
d3	0	0	0	0	1	0	1	1	1	1	0	0	5
d4	0	0	0	0	0	0	1	1	0	0	1	0	3
d5	0	0	0	0	1	1	1	1	0	0	0	1	5

After calculating all terms, the results are obtained in Table 8. Furthermore, it can be concluded that the root node obtained is as in Figure 2.

Table 8. Information Gain 1

Attribute	N(0)	P(0)	Entropy(0)	N(1)	P(1)	Entropy(1)	Gain
Hurry	3	1	0,811278	1	0	0	0,579471
Sell	3	1	0,811278	1	0	0	0,579471
People	3	1	0,811278	1	0	0	0,579471
Privacy	3	1	0,811278	1	0	0	0,579471
Order	2	0	0	2	1	0,889975	1,762.478
Not Quite Enough	3	0	0	1	1	1	1,628493
Switch	1	1	1	3	0	0	0,828493
Issue	1	1	1	3	0	0	0,828493
Case	3	1	0,811278	1	0	0	0,579471
Hurry	3	1	0,811278	1	0	0	0,579471
Sell	3	1	0,811278	1	0	0	0,579471
People	3	1	0,811278	1	0	0	0,579471
Highest Gain : 1,577841							
Node: Command							

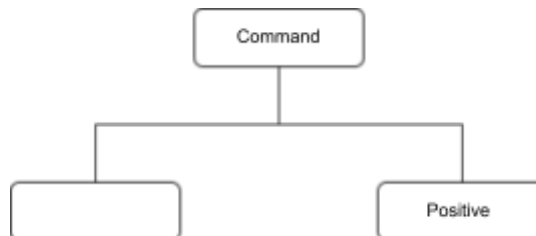


Figure 2. Node 1 Decision Tree

Because the entropy value for the "command" node branch is not zero, a recalculation is performed to determine the next node. The second information gain calculation can be seen in Table 9. From Table 9, it can be concluded that the root node obtained is as in Figure 3.

Table 9. Information Gain 2

Attribute	N (0)	P (0)	Entropy (0)	N (1)	P(1)	Entropy (1)	Gain
Hurry	2	1	0,918296	1	0	0	0,23464
Sell	2	1	0,918296	1	0	0	0,23464
People	2	1	0,918296	1	0	0	0,23464
Privacy	2	1	0,918296	1	0	0	0,23464
Not Quite Enough	2	0	0	1	1	1	0,9
Switch	1	1	1	2	0	0	0,1
Issue	1	1	1	2	0	0	0,1
Case	3	1	0,811278	0	0	0	0,025037

Attribute	N (0)	P (0)	Entropy (0)	N (1)	P(1)	Entropy (1)	Gain
Ministry of Finance	3	1	0,811278	0	0	0	0,025037
Tax	2	1	0,918296	1	0	0	0,637064
Answer	2	1	0,918296	1	0	0	0,637064
Hight Gain: 0.9							
Node: bear							

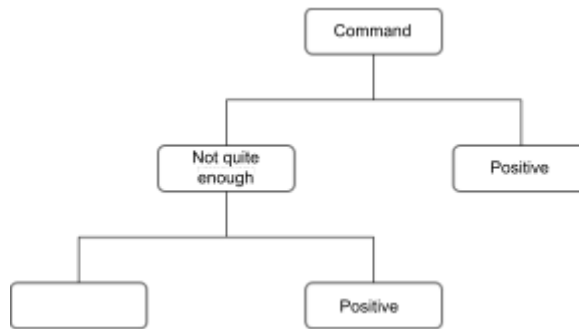


Figure 3. Node 1 Decision Tree

Because the entropy value for the "command" node branch is still not equal to 0, it is necessary to recalculate to determine the next node. The third information gain calculation can be seen in Table 10. From the results of Table 10, it can be concluded that the root node obtained is as in Figure 4.

Table 10. Information Gain Last Node

Attribute	N(0)	P(0)	Entropy(0)	N(1)	P(1)	Entropy(1)	Gain
Hurry	1	0	0	1	0	0	2,877443751
Sell	1	0	0	1	0	0	2,877443751
People	1	0	0	1	0	0	2,877443751
Privacy	1	0	0	1	0	0	2,877443751
Switch	1	0	0	1	0	0	2,877443751
Issue	1	0	0	1	0	0	2,877443751
Case	0	0	0	0	0	0	2,877443751
Ministry of Finance	0	0	0	0	0	0	2,877443751
Tax	1	0	0	1	0	0	2,877443751
Answer	0	0	0	0	0	0	2,877443751

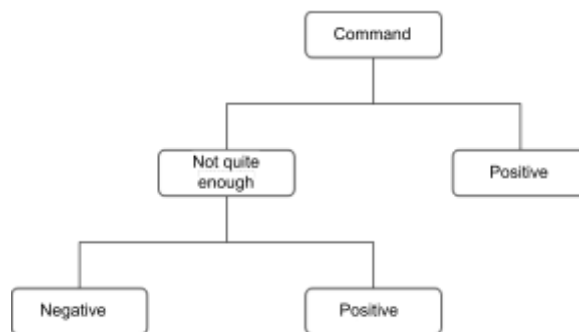


Figure 4. Node 3 Decision Tree

Because the entropy value on the 'tanggung' node branch has reached a value equal to 0, the 'lose' attribute branch is a leaf node, and no gain calculation is carried out to determine the next node.

3.3. Evaluation

At the evaluation stage, researchers carried out model tests and model evaluations to determine the performance of the Naive Bayes and C4.5 algorithms. The classification results will be displayed in the form of a confusion matrix. The results of the evaluation of each algorithm can be seen in Table 11.

Table 11. The Results of the Evaluation of Each Algorithm

Algorithm	Accuracy	Precision		Recall		F1	
		positive	negative	positive	negative	positive	negative
Naive Bayes	70%	72%	68%	73%	66%	73%	67%
C.45	68%	69%	67%	74%	62%	72%	64%

4. Conclusion

Based on the comparison accuracy results of the Naive Bayes and C4.5 algorithms, the highest accuracy value was obtained at 70% for the Naive Bayes algorithm. In comparison, the C4.5 algorithm obtained an accuracy value of 68%. This proves that the Naive Bayes algorithm can predict data accuracy of 70% for positive and negative sentiment.

References

- [1] M. Ahmad, S. Aftab, and I. Ali, "Sentiment Analysis of Tweets using SVM," *Int. J. Comput. Appl.*, vol. 177, no. 5, pp. 25–29, 2017, doi: 10.5120/ijca2017915758.
- [2] A. Eiji and S. Mehta, "Simulation-Based 5G Femtocell Network System Performance Analysis," *Int. J. Cyber IT Serv. Manag.*, vol. 3, no. 1, pp. 74–78, 2023.
- [3] Zulfikar Hardiansyah, "Rentetan Aksi Hacker Bjorka dalam Kasus Kebocoran Data di Indonesia Sebulan Terakhir," <https://tekno.kompas.com/>, 2022.
- [4] M. C. Saputra and E. Andajani, "Analysis of Factors Influencing Intention to Adopt Battery Electric Vehicle in Indonesia," *ADI J. Recent Innov.*, vol. 5, no. 2, pp. 100–109, 2024.
- [5] N. M. A. J. Astari, "Analisis Sentimen pada Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier," <https://repo.undiksha.ac.id/>, 2021.
- [6] O. A. D. Wulandari and D. Apriani, "Sustainable Institutional Entrepreneurial Culture and Innovation For Economic Growth," *APTISI Trans. Manag.*, vol. 7, no. 3, pp. 221–230, 2023.
- [7] R. Sholehurrohmah and I. Sabda Ilman, "Analisis Sentimen Tweet Kasus Kebocoran Data Penggunaan Facebook Oleh Cambridge Analytica," *J. Pepadun*, vol. 3, no. 1, pp. 140–147, 2022, doi: 10.23960/pepadun.v3i1.108.
- [8] A. W. Kusuma, Y. Jumaryadi, and A. Fitriani, "Examining the Joint Effects of Air Quality, Socioeconomic Factors on Indonesian Health," *Aptisi Trans. Technopreneursh.*, vol. 5, no. 2sp, pp. 186–195, 2023.
- [9] S. Andayani, "Formation of clusters in Knowledge Discovery in Databases by Algorithm K-Means," *SEMNAS Mat. dan Pendidik. Mat. 2007*, 2007.
- [10] Z. Lubis, M. Zarlis, and M. R. Aulia, "Performance Analysis of Oil Palm Companies Based on Barcode System through Fit Viability Approach: Long Work as A Moderator Variable," *Aptisi Trans. Technopreneursh.*, vol. 5, no. 1, pp. 40–52, 2023.
- [11] J. A. Septian, T. M. Fachrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," *J. Intell. Syst. Comput.*, vol. 1, no. 1, pp. 43–49, 2019, doi: 10.52985/insyst.v1i1.36.
- [12] L. K. Choi, P. A. Sunarya, and M. Fakhrezzy, "Blockchain technology as authenticated system for smart universities," *IAIC Trans. Sustain. Digit. Innov.*, vol. 4, no. 1, pp. 57–61, 2022.
- [13] A. Reza Satria and S. Adinugroho, "Analisis Sentimen Ulasan Aplikasi Mobile menggunakan Algoritma Gabungan Naive Bayes dan C4.5 berbasis Normalisasi Kata Levenshtein Distance," vol. 4, no. 11, pp. 4154–4163, 2020.
- [14] I. D. Girinzio, A. Ramadan, D. B. Saputra, and G. Mustika, "Improve Critical Thinking Students In Indonesia For New Learning Management System," *Int. Trans. Educ. Technol.*, vol. 1, no. 2, pp. 111–121, 2023.
- [15] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00327-4.

- [16] D. S. S. Wuisan, R. A. Sunardjo, Q. Aini, N. A. Yusuf, and U. Rahardja, "Integrating Artificial Intelligence in Human Resource Management: A SmartPLS Approach for Entrepreneurial Success," *Aptisi Trans. Technopreneursh.*, vol. 5, no. 3, pp. 334–345, 2023.
- [17] P. Sedgwick, "How to read a receiver operating characteristic curve," *BMJ*, vol. 350, no. May, 2015, doi: 10.1136/bmj.h2464.
- [18] C. Indonesia, "Baca artikel CNN Indonesia '10 Kasus Kebocoran Data 2022: Bjorka Dominan, Ramai-ramai Bantah,'" <https://www.cnnindonesia.com/>, 2022.
- [19] C. Indonesia, "BPJS Ketenagakerjaan Bantah Data Dibobol Bjorka," <https://www.cnnindonesia.com/>, 2023.
- [20] U. Rahardja, C. T. Sigalingging, P. O. H. Putra, A. Nizar Hidayanto, and K. Phusavat, "The impact of mobile payment application design and performance attributes on consumer emotions and continuance intention," *SAGE Open*, vol. 13, no. 1, p. 21582440231151920, 2023.
- [21] S. J. Kuryanti, S. N. Khasanah, S. Nusa, and M. Jakarta, "Comparison of Naive Bayes Algorithm, C4.5 and Random Forest for Service Classification Ojek Online," *J. Publ. Informatics Eng. Res.*, vol. 3, no. 2, 2019.
- [22] S. Pranata, K. Hadi, M. H. R. Chakim, Y. Shino, and I. N. Hikam, "Business Relationship in Business Process Management and Management with the Literature Review Method," *ADI J. Recent Innov.*, vol. 5, no. 1Sp, pp. 45–53, 2023.
- [23] B. Rawat, A. S. Bist, U. Rahardja, E. P. Harahap, and R. A. D. Septian, "Novel Framework to Define Extended & Mixed Reality for Online Learning," in 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS), IEEE, 2022, pp. 1–4.
- [24] D. A. H. D. Larasati and T. Sutrisno, "Tourism Site Recommendation in Jakarta Using Decision Tree Method Based on Web Review," *SSRN Electron. J.*, 2018, doi: 10.2139/ssrn.3268964.
- [25] A. Lubis, I. Iskandar, and M. M. L. W. Panjaitan, "Implementation of KNN Methods And GLCM Extraction For Classification Of Road Damage Level," *IAIC Trans. Sustain. Digit. Innov.*, vol. 4, no. 1, pp. 1–7, 2022.
- [26] T. R. Shultz and S. E. Fahlman, *Encyclopedia of Machine Learning and Data Mining*. 2017.
- [27] A. Williams and C. S. Bangun, "Artificial Intelligence System Framework in Improving The Competence of Indonesian Human Resources," *Int. J. Cyber IT Serv. Manag.*, vol. 2, no. 1, pp. 82–87, 2022.
- [28] A. M. Pravina, I. Cholissodin, and P. P. Adikara, "Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2789–2797, 2019.
- [29] Adhitya Karel Maulaya and Junadhi, "Analisis Sentimen Menggunakan Support Vector Machine Masyarakat Indonesia Di Twitter Terkait Bjorka," *J. CoSciTech (Computer Sci. Inf. Technol.)*, vol. 3, no. 3, pp. 495–500, 2022, doi: 10.37859/coscitech.v3i3.4358.
- [30] M. Annas and S. N. Wahab, "Data Mining Methods: K-Means Clustering Algorithms," *Int. J. Cyber IT Serv. Manag.*, vol. 3, no. 1, pp. 40–47, 2023.