Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.8 No.1 January 2023, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Sentiment Analysis of the Indonesian National Team in the 2020 AFF Cup Using Naïve Bayes and K-Nearest Neighbor Algorithms

Muhammad Ilham Fadila[1], Hanafi[2*], Anggit Dwi Hartanto[3]

[1,2,3]*Information System, Faculty of Computer Science, University of AMIKOM Yogyakarta, Indonesia*

[1]muhammad.fadila@students.amikom.ac.id, [3]anggit@amikom.ac.id

[2]hanafi@amikom.ac.id(*)

Received: 2022-10-24; Accepted: 2023-01-05; Published: 2023-01-24

Abstract— The AFF Cup is a football competition organized by the ASEAN Football Federation, or AFF for short. The 2020 AFF Cup was held in 2021 due to the COVID-19 pandemic. The Indonesian National Team advanced to the final round and became runner-up in the championship. With the end of the championship and the Indonesian National Team having to accept defeat in the final, the public responded through tweets on Twitter. Through these tweets, it will be known how the public evaluates the performance of the Indonesian National Team in the 2020 AFF Cup. It is vital to carry out this research to obtain information regarding society's response. The research that will be conducted is sentiment analysis. Sentiment analysis will be carried out on Rapid Miner software, with the algorithms used being Naïve Bayes and K-Nearest Neighbor. The data used to perform sentiment analysis are tweets from Twitter taken using SNScrape. This research aims to analyze public responses to the Indonesian National Team in the 2020 AFF Cup. This research will determine the percentage of positive, neutral, and negative sentiments from public responses. So that later it can be concluded how the public responds to the Indonesian National Team, whether positive, neutral, or negative. It is also to find out which algorithm has the higher accuracy. The results obtained for Naive Bayes with an accuracy of 64.74% are 71.54% positive sentiment, 15.45% neutral sentiment, and 13.01% negative sentiment. For K-Nearest Neighbor, with an accuracy of 65.64% is 80.49% positive sentiment, 15.45% neutral sentiment, and 4.06% negative sentiment. Both algorithms have the highest accuracy compared to other algorithms in Rapid Miner when the sentiment analysis is performed, with K-Nearest Neighbor having slightly higher accuracy. Most tweets about the Indonesian National Team in the 2020 AFF Cup had positive sentiments. Based on these results, it can be concluded that even though the Indonesian National Team did not win the 2020 AFF Cup, the public still responded positively.

Keywords— 2020 AFF Cup, Indonesian National Team, Sentiment Analysis, Naïve Bayes, K-Nearest Neighbor

I. INTRODUCTION

The most widely played sport in the world is football. The ease with which it may be played is one among the factors contributing to its appeal. Football is played on expansive fields in developing nations as well as in large stadiums in developed cities [1].

Football is a match played by two teams of 11 players each, one player in each team plays as a goalkeeper. The goalkeeper has a special position, since they are the sole member of the team with the ability to handle the ball. Only inside of their own penalty box are they allowed to handle the ball [2].

Each team defends its goal and tries to score against the opponent's goal. Football is played in two halves where each half lasts 45 minutes with a 15 minutes break after the first half [3].

The composition of the football players who play is commonly referred to as a formation. A formation can differ from team to team [3]. A great team is well-organized and has a distinct playing philosophy. To provide the team the best possibilities for achieving a balance between defensive unity and attacking flexibility, the coach assigns player positions and duties. [4].

In football, none is fixed in stone. Depending upon the outcome of a match, an attacker may eventually drop back to defend a lead, and the goalkeeper of the losing team may sometimes be seen going up to contest a corner in the closing seconds of a match out of desperation [2].

Social media is a digital medium for communication where users may exchange information, connect quickly, have two-way conversations, and send simultaneous direct messages to several users [5]. Social media users are growing and getting more vocal as a result of technological advancements, giving voice to thoughts on hot-button topics [6].

Twitter is a microblogging platform that links users through tweets, or short communications that can be sent immediately [7]. It is a social network that enables users to find intriguing accounts that interest them, whether or not they are individually familiar with them. Additionally, it enables users to publish their activities to anybody in the world, including strangers and members of their family [8].

They can continue to tune in if the channel piques their interest. They can follow specific people whose timelines they are interested in, as well as businesses, organizations, and others [9].

In some aspects, the function of following users is similar to a television guide, where they can view a list of channels with a few details about what is currently being shown on each channel [9].

Twitter has evolved into an important social network. It's a popular way for many individuals to communicate and learn. For a website that only enables users to publish messages of up to 280 characters, this is really amazing. Twitter provides a

DOI : http://dx.doi.org/10.25139/inform.v8i1.5222

platform for users to discuss topics they want to address. No subjects are forbidden in this open forum [10].

The AFF Cup is a biennial football competition organized by the ASEAN Football Federation (AFF) [11]. In the 2020 AFF Cup, the Indonesian National Team is coached by a new coach, Shin Tae Yong. The appointment of this new coach certainly provides new hope for the Indonesian National Team. This championship was the first competition that the Indonesian National Team participated in with Shin Tae Yong.

The Indonesian National Team managed to qualify for the group phase. It then proceeded to the semi-finals against Singapore with an aggregate score of 5-3, which made the team eligible to move to the final round against Thailand. In the final round, the Indonesian National Team lost to Thailand with an aggregate score of 6-2. It made Thailand the 2020 AFF Cup champion and Indonesia the runner-up under Shin Tae Yong's leadership.

After the championship ended, the public gave their responses on Twitter. Based on the background previously presented, to find out how the public responds to the Indonesian National Team in the 2020 AFF Cup, it is necessary to conduct research. The research that will be carried out is sentiment analysis using Rapid Miner software, with the algorithms using Naive Bayes and K-Nearest Neighbor.

Previous research analysis of public sentiment for the Indonesian national team competing in the 2020 AFF Cup using the K-Nearest Neighbors Algorithm yielded the following results: an accuracy of 67.49 percent, a precision of 78.99 percent, and a recall of 47.69 percent, with a percentage of the positive sentiment equaling 50 percent and the percentage of the negative sentiment equaling 49 percent [12].

By carrying out this sentiment analysis, it will be possible to determine the degree of accuracy possessed by the two algorithms and the proportion of tweets containing positive, neutral, or negative sentiments concerning the participation of the Indonesian National Team in the 2020 AFF Cup. The results of this % allow for inferences to be made regarding the public's reaction to the Indonesian National Team, and these inferences can be favorable, neutral, or negative.

II. RESEARCH METHODOLOGY

Analysis of people's thoughts, sentiments, judgments, attitudes, and feelings regarding an item and its attributes as expressed in written texts is known as sentiment analysis, and it's an area of study in its own right. Something can be an entity if it's a product, service, organization, person, event, issue, or topic [13].

Naive Bayes is a straightforward probabilistic classifier that calculates a set of probabilities by adding up the frequency and value combinations in a given dataset. The algorithm relies on the Bayes' theorem and presupposes that all qualities are independent of one another or not depending on the value of the class variable [14].

K-Nearest Neighbor is a semi-supervised learning algorithm that uses distance calculations to locate the K Nearest data. It requires training data and a predetermined K value. The algorithm predicts that the class of the unknown data will be similar to the majority class if K data have different classes [15].

Rapid Miner is an all-encompassing tool for data science, featuring visual workflow design and full automation of processes. It denotes that coding is not necessary to carry out tasks relating to data mining. Rapid Miner makes it possible to do sentiment analysis straightforwardly and expediently. Rapid Miner is one of the most popular data science tools [16].

This study uses a qualitative method, with text data in the form of tweets taken from Twitter and software called Rapid Miner to perform the sentiment analysis. Figure 1 provides a visual representation of the activities or procedures that will be carried out as part of this investigation, and the explanation will continue below it.

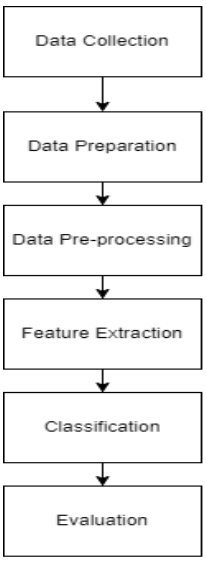

Figure 1. Research Stages

A. Data Collection

The gathering of data takes place at this phase of the process. The data will be analyzed to determine people's feelings about the topic. Using SNScrape, we scraped Twitter for Indonesian tweets containing "Indonesia AFF 2020." These tweets were collected as the data. SNScrape is a tool created in the Python computer language used to retrieve tweets from Twitter [17]. A total of 246 tweets were obtained over the time period beginning on January 1 and ending on January 31, 2022.

B. Data Preparation

At this stage, data preparation is carried out after the tweets have been successfully retrieved, namely labeling according to the tweet's sentiment, whether the sentiment is positive, neutral, or negative.

C. Data Pre-processing

At this stage, data pre-processing is carried out before being used to perform sentiment analysis. Data preprocessing is a technique that is applied to unprocessed data to get it ready for further processing [18], namely changing the raw data into a

DOI : http://dx.doi.org/10.25139/inform.v8i1.5222

form that is easier to understand [19]. The operators used can be seen below, followed by an explanation.

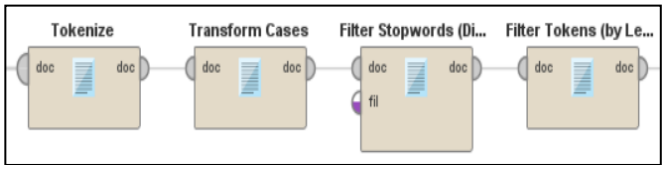

Figure 2. Pre-processing Operators

The operators used in Figure 2, namely:

a) Data Cleaning: To make use of the data, first, it must be cleaned. This involves deleting any dots, links, or URLs that appear in the tweets more than once and any symbols.

b) Tokenization: Tokenization is the division of a text into tokens. Due to the fact that entire texts are too specific to be used for any useful computations, this is essential for computational text analysis. Words are the most frequent tokens, because they are the most frequent semantically meaningful textual elements [20].

c) Case Folding Case folding is the process of making all the letters in a document lowercase. The capitalization of letters is not always used consistently in text documents. To put all of the text in a document into a standard form, case folding is therefore necessary [21].

d) Stopword: A stop word is a word that frequently appears in a text but offers nothing in the way of information. Stop words can be omitted over because they have no impact on classifying items [22].

e) Token Filtering: At this stage, tokens are filtered based on the length or number of characters to eliminate less meaningful words.

D. Feature Extraction

Extraction converts terms from text into numbers that a computer can read. Extraction is done using TF-IDF [23]. Term Frequency-Inverse Document Frequency (TF-IDF) is a feature extraction technique that assigns a value to each word in the training data [24].

Words are weighted or counted to determine how important they represent a sentence. The TF-IDF score is based on how often the words appear in the document [24]. The operator to be used can be seen below in Figure 3, with the create word vector option checked and TF-IDF vector creation schema selected.

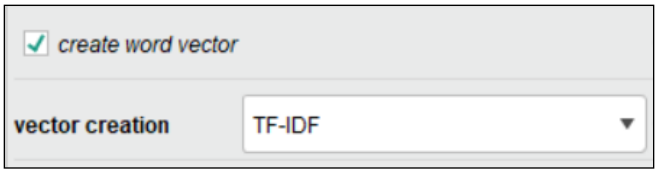

Figure 2. Feature Extraction

E. Classification

Classification evaluates a data object to be included in a particular class among several available classes [23]. Classification is done using Naïve Bayes and K-Nearest Neighbor algorithms, with the operators used as shown in Figures 4 and 5. The performance operator is used for statistical performance evaluation of classification tasks. This operator delivers a list of performance criteria values of the classification task.

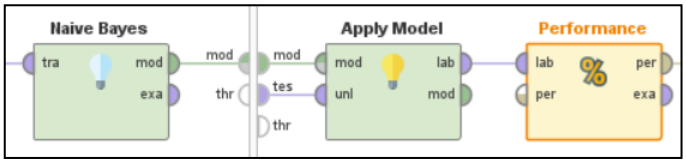

Figure 4. Naïve Bayes Classification Operators

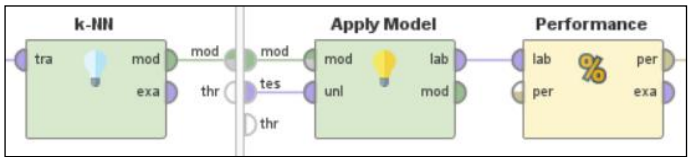

Figure 5. K-Nearest Neighbor Classification Operators

F. Evaluation

At this point in the process, an assessment of the Model's usefulness and accuracy is carried out, and the results of the classification algorithm's prediction are obtained. After that, the prediction outcomes' accuracy is evaluated by applying the confusion technique [25]. The next step is to draw a conclusion based on the gathered results.

III. RESULT AND DISCUSSION

After retrieving tweet data from Twitter, the amount of data successfully obtained was 246 tweets. The data is then divided into 2, namely training and test data, with a total of 123 data each. After going through the process of data preparation, data pre-processing, feature extraction, and classification. Sentiment analysis can be done using the Rapid Miner software and operators in Figure 6.

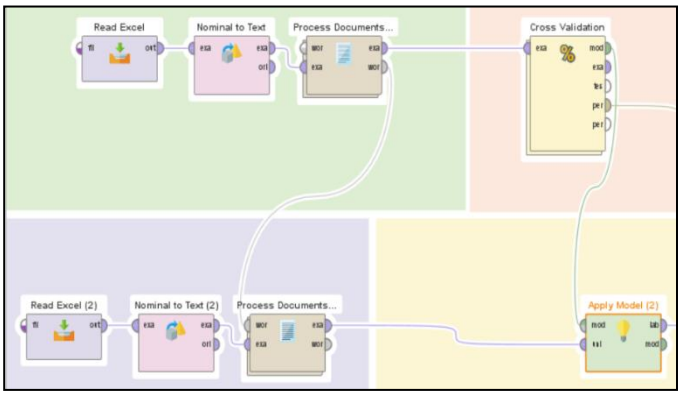

Figure 6. Sentiment Analysis Operators

DOI : http://dx.doi.org/10.25139/inform.v8i1.5222

As shown in Figure 6, the operators used are:

- Read Excel is an operator used to read excel formatted data.
- Nominal to text is an operator that changes the type of selected nominal attributes to text. It also maps all values of these attributes to corresponding string values.
- Process Documents from Data is an operator used to generate word vectors from string attributes
- Cross Validation is an operator used to perform cross-validation to estimate the statistical performance of a learning model.
- Apply Model is the operator used to apply the Model to the data.

Following the completion of sentiment analysis with the operators mentioned above, both algorithms' results are comparable in Figure 7. This is followed by a confusion matrix demonstrating how accurately the predictions were made. According to the data presented in Figure 7, Naive Bayes achieves an accuracy of 64.74 percent, with a class recall of 88.24 percent for positive, 38.24 percent for neutral, and 33.33 percent for negative, and a class precision of 75.95 percent for positive, 59.09 percent for neutral, and 31.82 percent for negative.

accuracy: 64.74% +/- 11.01% (micro average: 65.04%)

	true Positive	true Neutral	true Negative	class precision
pred. Positive	60	8	11	75.95%
pred. Neutral	6	13	3	59.09%
pred. Negative	2	13	7	31.82%
class recall	88.24%	38.24%	33.33%	

Figure 7. Naïve Bayes

As can be seen in Figure 8, the K-Nearest Neighbor algorithm achieves an accuracy of 65.64 percent, with a class recall of 88.24 percent for positive, 41.18 percent for neutral, and 33.33 percent for negative, as well as a class precision of 69.77 percent for positive, 63.64 percent for neutral, and 46.67 percent for negative.

accuracy: 65.64% +/- 12.10% (micro average: 65.85%)

	true Positive	true Neutral	true Negative	class precision
pred. Positive	60	15	11	69.77%
pred. Neutral	5	14	3	63.64%
pred. Negative	3	5	7	46.67%
class recall	88.24%	41.18%	33.33%	

Figure 8. K-Nearest Neighbor

When all of the findings have been collated, a chart can show how tweets are distributed according to the feelings they express.

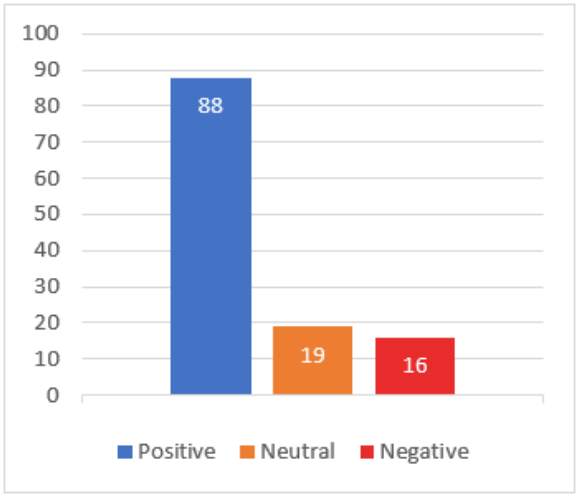

Figure 9. Naïve Bayes Sentiment Sum Chart

According to the findings of the classification carried out by Naive Bayes in Figure 9, the total number of tweets containing a happy feeling comes to 88, the total number of tweets containing a neutral mood comes to 19, and the total number of tweets containing a negative feeling comes to 16.

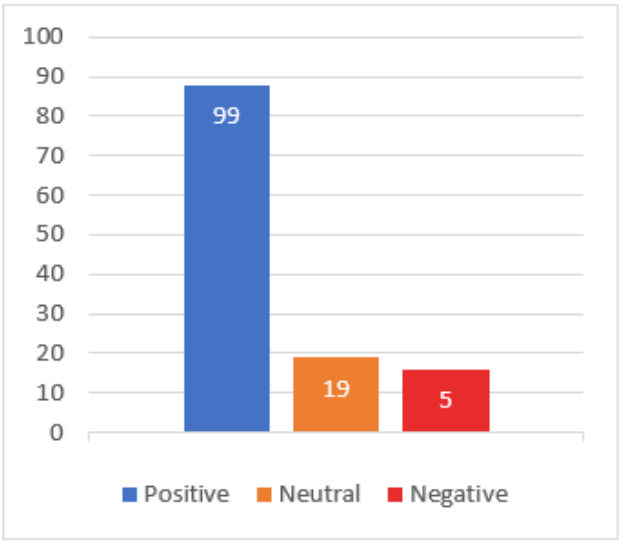

Figure 10. K-Nearest Neighbor Sentiment Sum Chart

The number of tweets with a positive sentiment is 99, the number of tweets with a neutral emotion is 19, and the number of tweets with a negative sentiment is 5. Figure 10 displays the results of the categorization using K-Nearest Neighbor. Once the numerical diagram has been finished, a second diagram may be made to establish the % of tweet distribution based on the sentiment. This can be done after the first diagram has been finished. The percentage of tweets with positive sentiment is 71.54 percent, neutral sentiment is 15.45 percent, and negative sentiment is 13.01 percent for Naive Bayes in Figure 11.

DOI : http://dx.doi.org/10.25139/inform.v8i1.5222

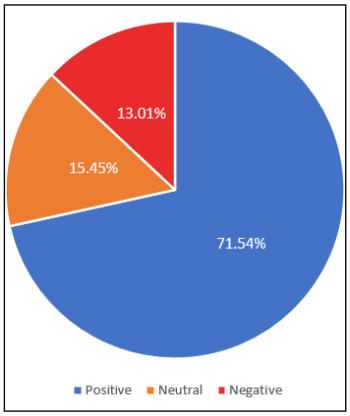

Figure 11. Naive Bayes Sentiment Percentage Chart

The percentage of tweets with positive sentiment is 80.49 percent, neutral sentiment is 15.45 percent, and negative sentiment is 4.06 percent for K-Nearest Neighbor in Figure 12. The results of the two algorithms used do not differ much, although K-Nearest Neighbor has slightly higher accuracy than Naive Bayes.

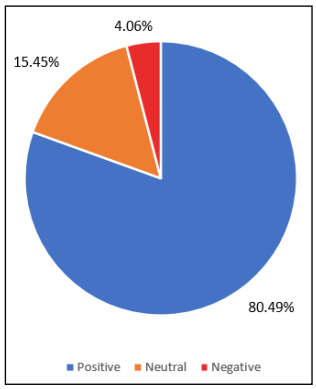

Figure 12. Naive Bayes Sentiment Percentage Chart

After going through all the processes and the results of sentiment analysis that have been obtained, a word cloud in Figure 13 below is created to find out what words often appear in tweets about the Indonesian National Team in the 2020 AFF Cup, with words that often arise in tweets totaling 20.

Figure 13. Word Cloud

The next step was to create Table I based on the word cloud in order to determine the number of occurrences of each word, with the frequency of a word being more than 10 times. Indonesia and AFF are the two terms that come up most frequently in this article.

TABLE I
NUMBER OF WORDS

Word	Total
indonesia	123
aff	123
timnas	69
piala	62
thailand	30
final	29
juara	27
pemain	26
cup	20
runner	19
bangga	16
shin	14
muda	13
leg	13
suzuki	13
gol	12
tae	12
tim	11
garuda	11
kalah	11

IV. CONCLUSION

The accuracy of the Naive Bayes method may be measured at 64.74 percent, whereas the accuracy of the K-nearest Neighbor method is measured at 65.64 percent. In comparison, the percentage of positive tweet sentiment for Naive Bayes is 71.54 percent, the percentage of neutral tweet sentiment is 15.45 percent, and the percentage of negative tweet sentiment is 13.01 percent. However, the percentage of positive tweet sentiment for K-Nearest Neighbor is 80.49 percent, the percentage of neutral tweet sentiment is 15.45 percent, and the percentage of negative tweet sentiment is 4.06 percent.

The two algorithms used have the highest accuracy compared to other algorithms in the Rapid Miner software when sentiment analysis is performed, with K-Nearest Neighbor having slightly higher accuracy than Naive Bayes.

When the general public's reactions are analyzed, it is clear that the percentage of tweets expressing a happy attitude is significantly larger than the number expressing a neutral or negative sentiment. In other words, most tweets about the Indonesian National Team competing in the 2020 AFF Cup have favorable views about the two algorithms employed. The word "*bangga*" carries a pleasant connotation and is used rather frequently.

Even though they had to concede defeat, Shin Tae Yong's team could still go through to the championship round. Despite this, it is possible to conclude, based on the findings that have been obtained from the analysis, that the public's reaction to the team is favorable, or, to put it another way, the Indonesian National Team that Shin Tae Yong coaches have been met with favorable reactions from the public.

DOI : http://dx.doi.org/10.25139/inform.v8i1.5222

REFERENCES

[1] A. B. Rogers, *Soccer: Science on the Pitch*. Greenhaven Publishing LLC, 2017.
[2] T. Dunmore and S. Murray, *Soccer For Dummies*. Wiley, 2022.
[3] J. Luxbacher, *Soccer: Steps to Success*. Human Kinetics, 2013.
[4] T. Strudwick, *Soccer Science*. Human Kinetics, 2016.
[5] I. Youssef Abdelhamid, H. Yahaya, N. Z. Ahmad, and M. Z. Mohammad Nazmi, "Foreign Language Learning Through Social Media: A Review Study," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 12, no. 6, 2022, doi: 10.6007/ijarbss/v12-i6/13910.
[6] H. R. Alhakiem and E. B. Setiawan, "Aspect-Bas1ed Sentiment Analysis on Twitter Using Logistic Regression with FastText Feature Expansion," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 5, pp. 840–846, 2022, doi: 10.29207/resti.v6i5.4429.
[7] N. Gulnazir and K. Salehuddin, "Investigating Lexical Variation and Change in Malaysian Twitter: A Conceptual Paper," *GEMA Online J. Lang. Stud.*, vol. 22, no. 4, pp. 90–107, 2022, doi: 10.17576/gema-2022-2204-06.
[8] L. Fitton, A. Hussain, and B. Leaning, *Twitter For Dummies*. Wiley, 2014.
[9] D. Murthy, *Twitter*. Polity Press, 2018.
[10] G. Kent, *You Are What You Tweet: Harness the Power of Twitter to Create a Happier, Healthier Life*. Star Stone Press, 2015.
[11] AFF, "AFF Cup - About the Tournament," *AFF*. https://www.affmitsubishielectriccup.com/2022/about/about-the-tournament (accessed Dec. 10, 2022).
[12] A. E. Pratama, A. Ariesta, and G. Gata, "Analisis Sentimen Masyarakat terhadap Tim Nasional Indonesia pada Piala AFF 2020 Menggunakan Algoritma K-Nearest Neighbors," *J. TICOM Technol. Inf. Commun.*, vol. 10, no. 3, pp. 187–196, 2022.
[13] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2020.
[14] I. B. A. Peling, I. N. Arnawan, I. P. A. Arthawan, and I. G. N. Janardana, "Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm," *Int. J. Eng. Emerg. Technol.*, vol. 2, no. 1, p. 53, 2017, doi: 10.24843/ijeet.2017.v02.i01.p11.
[15] K. Chomboon, P. Chujai, P. Teerarassammee, K. Kerdprasop, and N. Kerdprasop, "An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm," pp. 280–285, 2015, doi: 10.12792/iciae2015.051.
[16] M. Arnaldo, "Rapidminer | What is Rapidminer | Rapidminer for Beginners," *Analytics Vidhya*, Oct. 04, 2021. https://www.analyticsvidhya.com/blog/2021/10/intro-to-rapidminer-a-no-code-development-platform-for-data-mining-with-case-study/ (accessed Dec. 28, 2022).
[17] JustAnotherArchivist, "snscrape: A social networking service scraper," 2022. https://github.com/JustAnotherArchivist/snscrape (accessed Dec. 28, 2022).
[18] F. Isik, G. Ozden, and M. Kuntalp, "Importance of data preprocessing for neural networks modeling: The case of estimating the compaction parameters of soils," *Energy Educ. Sci. Technol. Part A Energy Sci. Res.*, vol. 29, no. 2, pp. 871–882, 2012.
[19] A. Aloysius and P. Nikil, "Data Preprocessing in Sentiment Analysis Using Twitter Data," *Int. Educ. Appl. Res. J.*, vol. 03, no. July, pp. 89–92, 2019.
[20] K. Welbers, W. Van Atteveldt, and K. Benoit, "Text Analysis in R," *Commun. Methods Meas.*, vol. 11, no. 4, pp. 245–265, 2017, doi: 10.1080/19312458.2017.1387238.
[21] R. Novendri, A. S. Callista, D. N. Pratama, and C. E. Puspita, "Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes," *Bull. Comput. Sci. Electr. Eng.*, vol. 1, no. 1, pp. 26–32, 2020, doi: 10.25008/bcsee.v1i1.5.
[22] P. Daowadung and Y. H. Chen, "Stop word in readability assessment of Thai text," *Proc. 12th IEEE Int. Conf. Adv. Learn. Technol. ICALT 2012*, pp. 497–499, 2012, doi: 10.1109/ICALT.2012.9.
[23] V. D. Antonio, "Analisis Kinerja Ekstrasi Fitur TF-IDF (Term Frequency – Inverse Document Frequency) Untuk Algoritma Klasifikasi Stochastic Gradient Descent Pada Analisis Sentimen Teks Indonesia," *Tesis*, pp. 1–82, 2021.
[24] 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, [Online]. Available: http://j-ptiik.ub.ac.id
[25] N. Yahya and A. Jananto, "Komparasi Kinerja Algoritma C.45 Dan Naive Bayes Untuk Prediksi Kegiatan Penerimaanmahasiswa Baru (Studi Kasus : Universitas Stikubank Semarang)," *Pros. SENDI*, no. 2014, pp. 978–979, 2019, [Online]. Available: https://www.unisbank.ac.id/ojs/index.php/sendi_u/article/view/7389/2369

DOI : http://dx.doi.org/10.25139/inform.v8i1.5222