Journal of Emerging Information System and Business Intelligence
ISSN: 2774-3993
Journal homepage: https://ejournal.unesa.ac.id/index.php/JEISBI/

A Comparative Analysis Opinion Mining Sea Games On Social Media

Lathifatuz Zuhroh[1], I Kadek Dwi Nuryana[2]

[1,2] *Information System, Faculty of Engineering,State University of Surabaya*

lathifatuz.19071@mhs.unesa.ac.id, dwinuryana@unesa.ac.id

ABSTRACT

Twitter is a social media platform that is freely accessible to everyone. As a result, numerous social phenomena and issues quickly emerge and spread globally through users' tweets. The 2023 SEA Games in Cambodia was no exception, as it sparked public opinion, particularly among Indonesian citizens, due to several incidents during the event that were considered unethical or inappropriate. This study aims to analyze public sentiment regarding these events using the C5.0 and Naïve Bayes algorithms. The performance of both algorithms will be compared to determine which one yields better results. The dataset consists of 1,000 tweets collected between March 14 and April 14, 2023. The findings indicate that Naïve Bayes outperforms C5.0, achieving an accuracy of 70%, compared to 67% for C5.0.

Keyword: Twitter, Sentiment Analysis, Naïve Bayes, C5.0, Sea Games.

Article Info:
Article history:
Received February 10,2025
Revised March 21,2025
Acceted April 11,2025

Corresponding Author
Lathifatuz Zuhroh
State University of Surabaya, and Indonesia
Lathifatuz.19071@mhs.unesa.ac.id

1. INTRODUCTION

The internet serves as a key medium for rapidly sharing opinions and information. Social media platforms enable people to interact easily without limitations of time and place. According to the National Information and Communication Technology Council, digital access increased during the pandemic. Today, social media is widely used to express opinions on trending topics or issues. Twitter, in particular, has seen a rise in users. Based on a report by We Are Social, the number of Twitter users in Indonesia reached 18.45 million in 2022, making Indonesia the fifth-largest Twitter user base globally [1].

Twitter offers various features, allowing users to post tweets in the form of text, images, GIFs, videos, and links. A single tweet can contain up to 280 characters, while media posts can include up to four images, GIFs, or videos [2]. Many issues are actively discussed on Twitter, with responses ranging from positive and negative to neutral. In this study, user opinions regarding the 2023 SEA Games will be collected, analyzed, and used as an evaluation reference for future event planning.

For the analysis, the Naïve Bayes and C5.0 algorithms will be applied to measure accuracy in assessing public sentiment. Since both algorithms are known for their high accuracy, they will be compared to determine which one performs better.

Based on this background, the study titled "Sentiment Analysis of the SEA Games on Twitter Using the C5.0 and Naïve Bayes Algorithms" will be conducted.

2. METHODS

The outline of this study begins with identifying problems in the surrounding environment. After that, relevant literature is reviewed, focusing on sentiment analysis, the C5.0 algorithm, and the Naïve Bayes algorithm. The research then formulates the problem, collects data from Twitter, preprocesses the data, assigns weights, splits the dataset into training and testing sets, and classifies the model using the C5.0 and Naïve Bayes algorithms. The final steps involve evaluating the processed data and comparing the accuracy levels of the two algorithms. The data used in this study consists of user opinions from Twitter regarding the SEA Games 2023.

Sentiment labeling is conducted to classify tweets as either positive or negative, ensuring the analysis proceeds correctly. This process utilizes a sentiment lexicon, which contains words or phrases associated with specific sentiments. Case folding is the process of converting uppercase letters into lowercase. This is the first step in data preprocessing to ensure consistency before moving on to tokenization. Tokenization involves breaking text into smaller units, such as words, phrases, or sentences, while removing punctuation and irrelevant characters. This step prepares the data for further analysis by structuring it into manageable components. In some cases, textual data may contain unnecessary duplicates. This stage involves removing duplicate entries to ensure the dataset remains clean and representative. This step applies stemming or lemmatization techniques to convert words into their root forms. For example, "walking," "ran," and "running" are reduced to their base form, such as "walk."

The data is transformed into a format compatible with the RapidMiner software. At this stage, the dataset is split into two parts: 20% for testing and 80% for training. The data is then analyzed and grouped into related variables. The classification process is carried out using RapidMiner, applying both the Naïve Bayes and C5.0 algorithms. The discussion presents the findings of the study, while the evaluation aims to provide insights for future researchers who intend to study similar topics, helping improve subsequent research.

3. RESULTS AND DISCUSSION

The data for this study was obtained through web crawling on Twitter using Google Colab with Python programming language. A total of 1,000 tweets related to the SEA Games 2023 were collected. Table 1. presents the raw data retrieved from the crawling process. The dataset includes several columns such as created_at, full_text, in_reply_to_screen_name, lang, location, tweet_url, and username. The "lang" column refers to the language used in the tweets, which in this study is Bahasa Indonesia.

Table 1.Crawling data

Create_at	Full__text	In_reply_to_screen_name	lang	location	username
Wed May 31, 2023	@idextratime Malah kaya final seagames	Idextratime	indonesia	Rembang, Indonesia	ayudhabayuuuu
Wed May 31, 2023	Wah gelut. Mau nyaingin final SEA Games ya	Sony Andrio Ranhas	Indonesia	Jakarta, Indonesia	SonyAndrio
Wed May 31, 2023	Kontroversi SEA Games 2023: Ketidaksetaraan Perlakuan Atlet dari	Kompasiana	Indonesia	Indonesia	Kompasiana

	Cabang Olahraga Berbeda https://t.co/xJvxxIteb T				

Preprocessing data is performed to clean and prepare the data for further analysis. The preprocessing steps in this study include: 1) Case folding; 2) Tokenizing; 3) Filtering; 4) Stemming.

Table 2. Example of a Tweet Before and After Preprocessing

Raw Tweet	
@idextratime Malah kaya final seagames	
Preprocessing Data	
Case Folding	idextratime malah kaya final seagames
Tokenizing	[idextratime, malah, kaya, final, seagames]
Filtering	[kaya, final, seagames]
Stemming	[kaya, final, seagames]

After processing the collected tweets, sentiment classification was performed. The distribution of sentiments in the dataset is as follows:

Table 3. Distribution of sentiments

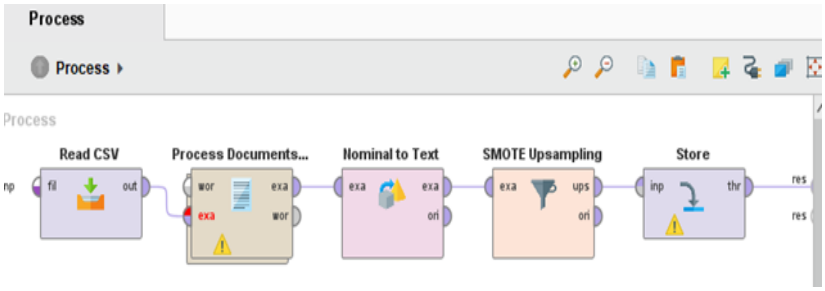
Sentiment	Count
Positive	426
Negative	374

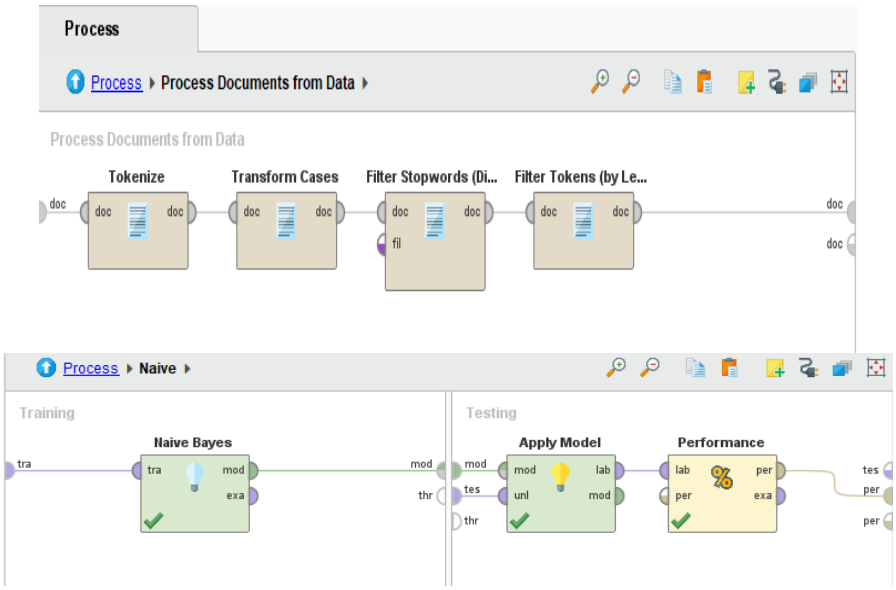
These results indicate that the SEA Games 2023 event in Cambodia received predominantly positive sentiment from Twitter users.

Table 4. Sentiment Result

	Data	Sentiment
1	kaya final seagames	Positive
2	gelut nyaingin final games	Negative
3	Kontroversi games ketidaksetaraan laku atlet cabang olahraga	Negative

The sentiment classification was conducted using the Naïve Bayes algorithm, implemented with the RapidMiner tool.





The process resulted in the following confusion matrix:

Table 5. Classification Result

	True positive	True negative	Class precision
Pred. positive	480	355	57.49%
Pred. negative	0	125	100.00%
Class recall	100%	26.04%	

Accuracy :63.02% +/-2.04% (micro average : 63.02%)
Weighted mean recall : 63.02% +/-2.04% (micro average : 63.02%)
Weighted mean precision : 78.76% +/- 0.69% (micro average : 78.74%)

With C5.0 Algorithm resulted in the following confusion matrix:

Table 6. Confusion Matrix

	True positive	True negative	Class precision
Pred. positive	480	480	50.00%
Pred. negative	0	0	00.00%
Class recall	100.00%	00.00%	

Accuracy : 50% +/- 0.00% (micro average: 50.00%)

In evaluating model performance, a higher accuracy generally indicates a more reliable and effective algorithm. Precision and recall were also considered in the comparison. High precision implies fewer false positive predictions, whereas high recall suggests that the model successfully identifies more positive instances.

Based on the testing results, the Naïve Bayes algorithm outperformed the C5.0 algorithm, achieving an accuracy of 63.02% compared to 50.00% for C5.0. This demonstrates that Naïve Bayes is a more suitable approach for sentiment analysis in this study.

CONCLUSION

The results of the sentiment analysis study indicate that the Naïve Bayes algorithm has a higher accuracy rate compared to the C5.0 algorithm, with an accuracy of 63.02% versus 50%. Each algorithm has its own strengths and weaknesses; therefore, it is advisable to use the algorithm that best suits the research needs. This research still requires further development, particularly in enhancing its accuracy and usefulness for future studies.

Provide a statement on the extent to which the research findings have addressed the formulated research questions. Present a conclusion based on the results and discussion, highlighting the theoretical contributions. Additionally, potential further development of the research findings and prospects for their application in future studies can also be included based on the obtained findings.

REFERENCES

[1] Rizaty, M. A. (2022, August 10). Pengguna Twitter di Indonesia Capai 18,45 Juta pada 2022. Retrieved May 25, 2023, from DataIndonesia.id.

[2] Twitter help.com Ui P, Aplikasi UX, Wisata D. Tempat Kuliner Berbasis Android Menggunakan Metode User-Centered Design Ui / Ux Design of Tourism Destination and Culinary Places Application Based on Android Using User-Centered Design. 2021;8(5):6574-658.

[3] Gissely Fiqih Harso (2022). Analisis Komparasi Kualitas Layanan Dan Harga Antara Shopee Dan Tokopedia. GEMAH RIPAH: Jurnal Bisnis. Volume 02, No 02.

[4] Bobby Kurniadi Widodo, N. H. (2022). Penerapan Algoritma Naive Bayes untuk Analisis Sentimen Penggunaan APlikasi Jobstreet. Techno.COM, 523-533.

[5] Munasatya, Nico., & Novianto, Sendi (2020) Natural Language Processing untuk Analisis Sentimen Presiden Jokowi Menggunakan Multi Layer Perceptron. Techno.COM, Vol. 19, No. 3, 237-244.

[6] Hakim, Bustony (2021) Analisa Sentimen data text preprocessing pada Data Mining dengan menggunakan Machine Learning. Journal of Business and Audit Information Systems 202;4(2):16-22.

[7] News, B. (2023, March 13). SEA Games 2023 'dihujani kritik', apa yang sebenarnya terjadi di Kamboja? Retrieved May 06, 2023, from BBC News Indonesia.

[8] X Trends FAQ, Retrieved May 06, 2023 from https://help.x.com/en/using-x/x-trending-faqs.

[9] Silitonga, W. H., & Sihotang, J. I. (2019). Analisis Sentimen Pemilihan Presiden Indonesia Tahun 2019 Di Twitter Berdasarkan Geolocation Menggunakan Metode Naïve Bayesian Classification. TeIKa, 9(2), 115-127.

[10] Rahmanita Widiyanti, C. S. (2022). Implementasi Algoritma C5.0 Untuk Klasifikasi Kepuasan Masyarakat Terhadap Pelayanan Kantor Kecamatan. JURIKOM (Jurnal Riset Komputer), 1200-1209.