Sentiment Analysis of Genshin Impact on Twitter Using Naïve Bayes

1st Maretha Fitrie Puruhita
Faculty of Industrial Engineering
Telkom University
Bandung, Indonesia
maretha@students.telkomuniversity.ac.id

2nd Faqih Hamami
Faculty of Industrial Engineering
Telkom University
Bandung, Indonesia
faqihhamami@telkomunivesity.ac.id

3rd Irfan Darmawan
Faculty of Industrial Engineering
Telkom University
Bandung, Indonesia
irfandaramwan@telkomuniversity.ac.id

***Abstract*—During the COVID-19 pandemic that interferes normal life around the world, people have an obligation to stay at home and quarantine themselves. This has led to an increase in the consumption of entertainment, especially online gaming which is known to be less harmful than other stress and aversive emotions. And Genshin Impact is one of the online games that won Google Play's the Best Game of 2020 award when pandemic happening. Released in September 2020 by China video game developer, miHoYo. Co., Ltd, Genshin Impact has been a hot trend on the microblogging platform, Twitter. The purpose of this research is to provide information regarding people's opinion emotion in their tweets toward Genshin Impact and this information will be a helpful resource for game improvement and can be used as reference of future research. By using sentiment analysis to help analyze the emotion contained in the text, the result will be categorized into three categories: positive, negative, or neutral sentiment. The data is gained through text mining then will be processed as text classified using Naive Bayes algorithm. Thus, the model will be going through evaluation of model's performance to measure how accuracy it is. The result of it stated that the best ratio between training and test set is 60:40 with 71.80% test accuracy, yet the accuracy between 3 others ratio is not much difference. That's why using hyperparameter tuning can find the optimal result. After finding the optimal result, the highest result it can get is 72.14%. Besides that, people on Twitter mostly perceive the game in neutral sentiment.**

***Keywords*—Sentiment Analysis, Genshin Impact, Naïve Bayes Classifier**

I. INTRODUCTION

In pandemic of COVID-19, the virus has disrupted normal activities globally. The obligation to stay at home and be quarantined indeed increases entertainment consumptions and it is part of recreational aspect that people need to make them refreshed, both physically, mentally, and spiritually. Online games are one of many activities of entertainment consumption where they are usually less harmful than other behaviors to cope with stress and aversive emotion like alcohol, drug use, and overeating [1]. This event surely happened in many countries, as well as in Indonesia. Along with the increasing use of the internet in Indonesia, online game players also continue to keep growing. In January 2019, Indonesia's gaming market is ranked 17th in the world and being the first in Southeast Asia. Report of Limelight Networks' states that Indonesians spent around 8.54 hours/week for playing video games, a bit high than global average of 8.45 hours/week In January 2019, Indonesia's gaming market is ranked 17th in the world and being the first in Southeast Asia. Report of Limelight Networks' states that Indonesians spent around 8.54 hours/week for playing video games, a bit high than global average of 8.45 hours/week [2].

And one of the online games that won the Best Game of 2020 award of Google Play annual award is Genshin Impact [3]. Genshin Impact is a cross-platform online game that was released on 28th September 2020 by a video game development and animation studio based in Shanghai, miHoYo. Co., Ltd. Catching a lot of attention from fans its home country, China, and Global fans. This game is an open-world roleplay game that applies Player vs Environment (PvE) type. Based on Active Player via Fiction Horizon, there are around 65,521,480 active players according to the data with the all-time peak registering around 8,500,000 concurrent players [4]. While report from AppMagic site via gamerwk, there are around 6,857,493 downloaders in Indonesia based on Google Play and App store on September 9th, 2022. It makes Indonesia as 4th country with the most Genshin Impact player for mobile platform surpassing Japan [5].

Not only be the winner of the Best Game in 2020, in fact Genshin Impact is in first position as the game that is most talked about on one of the social network services, Twitter, in 2021 [6]. It makes the existence of the Genshin Impact community in Twitter is no need to be doubted. Twitter itself is one of the most widely used microblogging and maintain its status as a popular social media platform [7]. Globally, Twitter has 396.5 million users and at least 500 million tweets are sent every day [8]. Meanwhile in Indonesia, there are approximately 14,75 million users in April 2023, and it makes Indonesia in the sixth rank as a country with the highest number of Twitter user based on databoks [9].

Working on sentiment analysis have been researchers' job to help discovering user attitudes in various cases and on different social media data particularly on Twitter [10]. One case of previous studies about sentiment analysis is Analisis Sentiment pada Twitter untuk Games Online Mobile Legends

dan Arena of Valor dengan Metode Naïve Bayes Classifier (*Sentiment Analysis on Twitter for Online Games Mobile Legend and Arena of Valor using Naïve Bayes Classifier Method*). The purpose of the study is to be able to classify polarization of the positive and negative sentiment to both games online: Mobile Legend and Arena of Valor and get the accuracy from it [11].

The lack of sentiment analysis of Genshin Impact with Twitter user as its main subject, the author proposes a study of sentiment analysis using Naïve Bayes. Many cases of sentiment analysis use Naïve Bayes because the accuracy result is quite high. The author expected this study to help research in developing sentiment analysis and to let people know public opinion on Genshin Impact more.

II. RELATED STUDY

A. Genshin Impact

Genshin Impact is an online game developed and published by miHoyo, Co.Ltd, an animation company from China that was launched on September 28th, 2020. Based on an open-world role-playing game, it was released for PC (Windows), mobile (Android and iOS), Playstation 4, Playstation 5, and other platforms that continue to be added. There are two main available servers, China server and global server. In global server, they are divided into four other servers: Europe, America, Asia, and SAR (Hongkong, Taiwan, Macau). For the global server, they provide dubbing and subtitles from various languages. For its young age, Genshin Impact has already passed $1 billion in player spending in less than 6 weeks according to Sensor Tower [12].

B. Twitter

Twitter is one of the online microblogging platforms which has at 1.3 billion accounts and 336 million active users posting 500 million tweets per day [13]. They are a social network service that can be used by children over the age of 13, allowing them to write their thoughts in 280 characters and called it as tweet. This tweet has information that later can be extracted for business purposes [14].

C. Text Mining, Text Classification, Text Sentiment Analysis

Text mining is described as a process of extracting information directly from textual data. Text itself can be defined as unstructured data consisting of strings called word. Text Classification is defined as the process of assigning data to a category or some categories and when it comes to text classification, basically it's assigning one or some of the predefined categories to each text [15]. While sentiment analysis is a sub-part of Natural Language Processing (NLP) which focuses on detecting feelings from words. This analysis is also known as opinion mining that has an automated process to understand, extract, and process textual data that can retrieve sentiment information from an opinion sentence [16].

D. TF-IDF

TF-IDF stands for two different words, TF and IDF. TF itself is abbreviation of term frequency. As its name suggests, it is used to measure frequency of a term that is present in a document [17].

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D) \quad (1)$$

Where:
tf : term frequency
idf : inverse document frequency

$$\text{tf}(t, d) = \frac{f_{t,d}}{\Sigma_{t' \in d} f_{t',d}} \quad (2)$$

Where:
t : term
d : document
$tf_{(t,d)}$: term frequency
$f_{t,d}$: frequency of term in a document
$\sum_{t' \in d} f_{t',d}$: total number of words in the document.

Meanwhile, IDF is Inverse Document Frequency where it assigns lower weight to frequent words and assigns greater weight for the word that is infrequent.

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \quad (3)$$

Where:
N : total number of documents in the corpus
$|\{d \in D : t \in d\}|$: number of documents that have the term.

E. Naïve Bayes

Naïve Bayes is a simple probabilistic classifier. It calculates a set of probabilities by adding up the frequencies and combinations of values. The method has a minimum error rate compared to other classification algorithms. One of its advantages is the method requires only a small amount of training data to determine estimation of the parameters needed in the classification process. Naïve Bayes often performs much better in most complex real-world situation [18].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (4)$$

Where:
P(A|B) : posterior probability
P(B|A) : likelihood
P(A) & P(B) : prior probability

This study uses distribution function and the most used in NLP as probabilistic learning method based on *Bayesian Reasoning and Gaussian Processes for Machine Learning* by K, Hemachandran et al. (2022) [19], that is Multinomial distribution function, and this variant of Naïve Bayes is a frequency-depended model [20].

III. RESEARCH METHOD

This research identifies the problem and solution by using these steps as Figure 1. There are four mains process: Identification Problem and Solution, Data Preprocessing, Data Processing, and Evaluation.

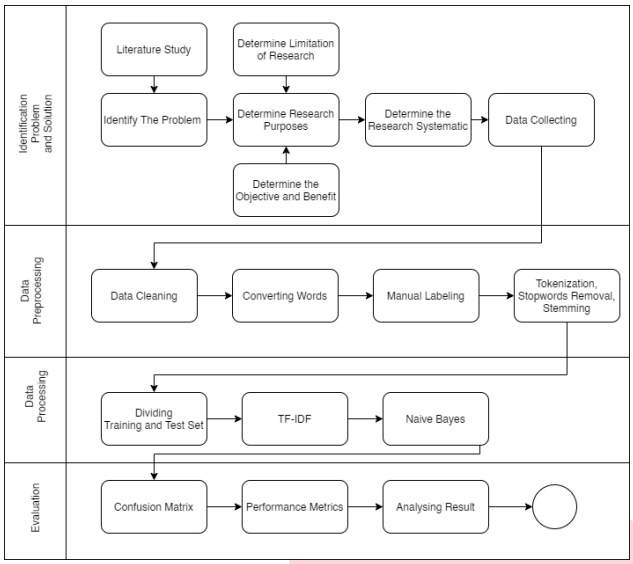

FIGURE 1
Identification Problem and Solution

A. Identification of Problem

Limiting the scope of analysis in this study is crucial in order that this research does not get wider and out from the main objective. In this sentiment analysis research, the main topic is limited to one of famous RPG games, Genshin Impact, in Twitter as the platform. Collecting data by crawling through Twitter database from Genshin Impact player who is also Twitter user, keyword to collect the data is limited to tweet containing #Genshin, #Genshin_Impact, #GenshinImpact, Genshin Impact, and Genshin. Later, these data will be categorized into three different emotions: positive, neutral, and negative. Codes will be performed in online platform, Colaboratory by Google, and python as the programming language. And the data will be processed using Naïve Bayes assisted by TF-IDF.

B. Data Crawling

Data Crawling is the method used in this case study to obtain the data needed. To be specific, the data is crawled tweets from Twitter database using Twitter API and python open-source library, Tweepy. In this method, consumer keys and authentication tokens are required to give authentication to Tweepy to access Twitter database. Data are retrieved every day from December 13th, 2022, to January 25th, 2023. Total retrieved data in 43 days is 1,110,147 raw data and stored in CSV file.

C. Data Preprocessing

Since the retrieved data is still in unstructured raw form, data preprocessing is required to filter the data. The objectives of data preprocessing are to improve the accuracy and reliability of a dataset, make data consistent, and increase the readability of algorithms of data [21].

D. Data Processing

Data processing has a purpose to gain the main information of research. The clean data from the previous stage will be weighted using TF-IDF and classified using Naive Bayes.

TABLE 1
Training and Test Data Ratio

Ratio	Training Set	Test Data
80:20	31,224	7,806
70:30	27,321	11,709
60:40	23,418	15,612
50:50	19,515	19,515

IV. EVALUATION

A. Dataset Evaluation

In data crawling, this study gets 1,110,147 items of raw data. After going through preprocessing stage, the raw data that has been cleaned only left 39,030 items. This data is already manually labeled by the author with subjective view and closest assumptions.

TABLE 2
The Comparison of Amount of Data

The Amount of Data	Before	After
	1,110,147	39,030

This study uses three types of sentiment: positive, neutral, and negative. Thus, the result of manually labeled as in Table 2. The author has 7736 positive sentiment data, 23055 neutral data, 8239 negative data. We can conclude that there is an imbalance in the amount of data. The gap between positive and negative to neutral is quite far.

TABLE 3
The Comparison of Amount of Data in Manual Label

The Amount of Data	Positive	Neutral	Negative
	7736	23055	8239

B. Naïve Bayes Classification by Ratio

In this study, the dataset was divided into two main sets: train and test set. Ideally, the ratio of distribution of train set and test set is 80:20. Table 4 expressed that 80% of 39,030 data is 31,224 for train set and the rest 20% goes to test set in total 7806 test set and the rest of ratio can be seen in Table 3.

TABLE 4
The Comparison of Amount of Train and Test Set

Ratio	Training Set	Test Set	Training Set Accuracy	Test Set Accuracy
80:20	31,224	7,806	83.82%	71.48%
70:30	27,321	11,709	84.59%	71.71%
60:40	23,418	15,612	85.19%	71.80%
50:50	19,515	19,515	85.96%	71.54%

From Table 4, we can conclude that ratio 60:40 has the highest test accuracy even though its training set is lower than training set of 50:50 ratio. Even the result between accuracy makes 60:40 the best ratio for this study best, yet the accuracy between all the ratios is not much different.

C. Naïve Bayes Hyperparameter Tuning

This study also experiments with hyperparameter tuning to get the optimal result. This hyperparameter tuning is using

GridSearchCV. GridSearchCV helps in finding the best way to tune the hyperparameters based on the training set. In this study, there's two parameters that will author manually sets out: alpha and cv (cross validation). The ratio that is used in this calculation is 60:40 based on Table 4 where it is the best ratio. Parameter alpha defines the number of Laplace smoothing. Normally, alpha score is 1,0. But this study sets the alpha into five different six alpha scores: '0,1', '0,25', '0,5', '1.0','2,5', and '5,0'. While cv or cross-validation is a method of resampling on different iteration and different portions. In common, the score of this parameter is 5, but in the study the author sets it into four parameters: 2, 4, 5, and 10.

TABLE 5
Result of Hyperparameter Tuning

CV	Average Accuracy	Best Alpha	Train Accuracy	Test Accuracy
2	70.25%	0.5	87.06%	72.10%
4	71.33%	0.5	87.06%	72.10%
5	71.36%	0.5	87.06%	72.10%
10	71.56%	1.0	83.53%	72.14%

Looking into Table 5 the result of Hyperparameter Tuning, we can make a conclusion that highest average accuracy is in 71.56%, both average accuracies come from cross validation (cv) 10 with the same alpha score in 1.0. Cross validation 2, 4, and 5 have the same train accuracy and test accuracy. While the highest test accuracy is cross validation in 10 with alpha 1.0, the result is 72.14%. And back to Table 4, the score of accuracy between normal parameter in ratio 60:40 compared accuracy after hyperparameter tuning is only increase 0.34%. Last, it can be concluded that score accuracy is close to each other even in different parameter settings.

D. Confusion Matrix and Performance Metric

After searching the best ratio as well as the parameters with hyperparameter tuning, now is time to check how well the created model is using Confusion Matrix and Performance Metric. Looking at Table 6 Confusion Matrix by the best ratio.

TABLE 6
Confusion Matric by Ratio 60:40

		Predicted Values		
		Negative	**Neutral**	**Positive**
Actual Value	**Negative**	2.516	572	232
	Neutral	1.075	6.959	1.127
	Positive	448	948	1.735

From the information we can calculate the accuracy, precision, recall, and the F-1 Score. All calculations are written in Table 7.

TABLE 7
Performance Metric Calculation by Ratio

	Precision (p)	Recall (r)	F1 Score
Negative	$p = \frac{TN}{TN + FN\,neut + FN\,pos} = \frac{2516}{2516 + 1075 + 448} = 0.62$	$r = \frac{TN}{TN + FNt\,neg + FP\,neg} = \frac{2516}{2516 + 572 + 232} = 0.76$	$F1 = \frac{2 \times p \times r}{p + r} = \frac{2 \times 0.62 \times 0.76}{0.62 + 0.76} = 0.68$
Neutral	$p = \frac{TNt}{TNt + FNt\,neg + FNt\,pos} = \frac{6959}{6959 + 572 + 948} = 0.82$	$r = \frac{TNt}{TNt + FN\,neut + FP\,neut} = \frac{6959}{6959 + 1075 + 1127} = 0.76$	$F1 = \frac{2 \times p \times r}{p + r} = \frac{2 \times 0.82 \times 0.76}{0.82 + 0.76} = 0.79$
Positive	$p = \frac{TP}{TP + FP\,neut + FP\,neg} = \frac{1735}{1735 + 1127 + 232} = 0.56$	$r = \frac{TP}{TP + FNt\,pos + FN\,pos} = \frac{1735}{1735 + 948 + 448} = 0.55$	$F1 = \frac{2 \times p \times r}{p + r} = \frac{2 \times 0.52 \times 0.55}{0.56 + 0.55} = 0.56$

Precision represents the ratio of correctly classified. From Table 7, can be seen that Precision of Neutral sentiment is higher than Negative and Positive sentiment. This is obtained because neutral sentiment has the biggest value for True Neutral which the actual value same with the predicted one. While the positive sentiment only has 1735 items where the actual and predicted value same.

Recall represents the ratio between the correctly classified with the total number of results that have its presence. In short, it's the ability to detect the characteristic, the higher recall score, the higher chance it can be detected. Both Neutral and Negative sentiment has 0,76 score in recall means that the ability to detect its characters is way excellent than detect item with Positive sentiment.

F1-score represents how balance between precision and the recall. As expected, Neutral sentiment has the biggest score of F1-score.

TABLE 8
Confusion Matrix by Hyperparameter Tuning

		Predicted Values		
		Negative	**Neutral**	**Positive**
Actual Value	**Negative**	1.997	1.204	119
	Neutral	502	8.161	498
	Positive	272	1755	1.104

From Table 8, there is a slightly different score with Confusion Matrix by the ratio. This will also affect the value of its Performance Metric. Looking at Table 9, especially the score of Positive—Recall, the score is lower than Positive—Recall from Table 7. This low Positive—Recall score in Table 9 can lead the mobility of detecting the positive sentiment later.

TABLE 9
Performance Metric by Hyperparameter Tuning

Sentiment	Precision	Recall	F1-Score
Negative	0,72	0,60	0,66
Neutral	0,73	0,89	0,80
Positive	0,64	0,35	0,46

In Table 10, we can conclude that accuracy by hyperparameter tuning is higher than setting the ratio only. The highest score that can hyperparameter tuning reach is 72,14% while with normal parameter and changing the ratio, the highest score that this model can get is 71,80%.

TABLE 10
Accuracy Comparison

By Ratio Only	By Hyperparameter Tuning
$ACC = \frac{TP + TNet + TN}{Ntotal}$	$ACC = \frac{TP + TNet + TN}{Ntotal}$

$= \frac{1.735 + 6.959 + 2.516}{15.612}$ $= 0{,}7180$	$= \frac{1.104 + 8.161 + 1.997}{15.612}$ $= 0{,}7214$
71,80%	72,14%

Some public perception about Genshin Impact mostly have neutral sentiment, even the model can detect way easier to neutral sentiment by looking at highest score of both Neutral—Recall score. But it will be harder when it comes to positive sentiment. With all this accuracy, people can use this model for later research.

E. Visualization Result

Word of cloud is kind of visualization that can present the text of document as graphical representation. Mostly displaying words that most often appear in the document. The bigger its word, the bigger its frequency.

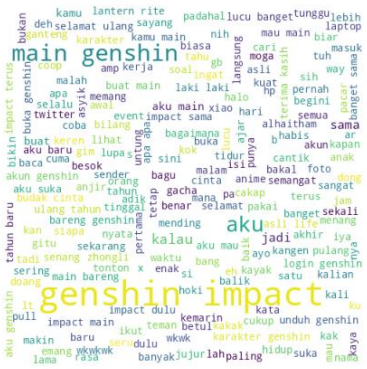

FIGURE 2
Word of Cloud from Positive Sentiment

Like in Figure 2, we can see that people mostly talk about the Genshin Impact itself and followed by the main Genshin. The positive sentiment words that we can catch up here are *asyik, cantik, enak, jujur, cinta, semangat, kuat*. But even in positive sentiment, we can see a word that has negative meaning like *habis*. The rest of the words are quite irrelevant. In Figure 4, it shows the word cloud from negative sentiment. Negative sentiment identically with profanities that may not be suitable with minor age. Even Geshin Impact itself has restricted player's age—it's restricted to ages 12 and below—parent still need to supervise their kids.

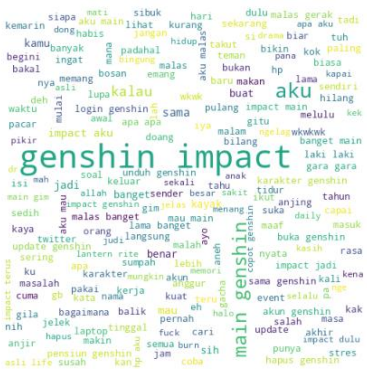

FIGURE 3
Word of Cloud from Negative Sentient

There are some negative sentiments like *jelek, sakit, aneh, takut, etc*. There is word for sakit indicates that player may complained about their health while playing this game. This is why Genshin Impact put warning screen while loading that this game may experience epileptic seizure, eye soreness, altered vision, and others. People need to be aware if some symptoms appear, they need to seek professional help and stop playing immediately.

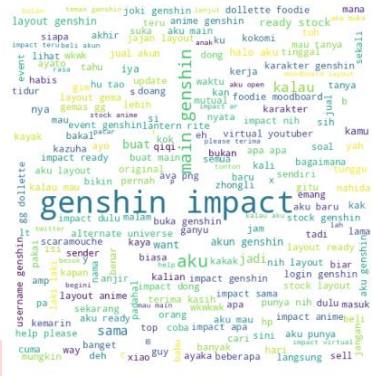

FIGURE 4
Word of Cloud from Neutral Sentiment

Most neutral sentiment consists of no sentiment word. In short, it doesn't have any special meaning or specific sentiment: positive or negative. Some words that considered in neutral sentiment, like *ready stock, joki, jual akun, jajan layout*, can be used as a sign of how big this Genshin Impact in buying and selling area. Some people use this chance to increase their income by selling layout with Genshin Impact theme or provide automation services.

V. CONCLUSION

From the data of this study and some calculations of it, people on Twitter mostly perceive Genshin Impact in neutral sentiment. Most of the time, neutral sentiments are not helping with the reputation of the games. But in some cases, it can lead to positive sentiment if the developer wants to give extra attention to their product. On the other side, the amount of negative sentiment is bigger than the positive one. The author recommends to people who want play Genshin Impact and join the community on Twitter to be more careful, especially minor (or underage). People need to be aware of the number of profanities that can be found. And some may get sick from playing this game because there are many words related to pain found in the sentiment as Genshin Impact already give the warning screen before the loading page.

The highest accuracy for this model of Naïve Bayes can be up to 72,14% by using hyperparameter turning. The setting of the parameter is alpha or Laplace smoothing in 1.0 with cross-validation (cv) around 10-20. But, in normal parameter the highest accuracy can be up to 71,80% for ratio 60:40. Even though it's quite high, this accuracy is considered smaller than some previous studies.

REFERENCES

[1] D. L. King, P. H. Delfabbro, J. Billieux and M. N. Potenza, "Problematic Online Gaming and the COVID-19 Pandemic," *Journal of Behavioral,* vol. 9, no. 2, pp. 184-186, 2020.

[2] InCorp Editorial Team, "A Window Opportunity in Indonesia Gaming for Chinese Developers," 13 March 2023. [Online]. Available: https://www.cekindo.com/blog/indonesia-gaming-market. [Accessed 5 August 2023].

[3] D. Takahashi, "Google Play's Best Games of 2020 — Genshin Impact is the Big Winner," 1 December 2020. [Online]. Available: https://venturebeat.com/2020/12/01/google-plays-best-games-of-2020-genshin-impact-is-the-big-winner/. [Accessed 5 August 2021].

[4] H. Milakovic, "How Many People Play Genshin Impact in 2023? (User & Growth Stats)," 4 January 2023. [Online]. Available: https://fictionhorizon.com/how-many-people-play-genshin-impact/. [Accessed 2 August 2023].

[5] Fadhil, "Salip Jepang, Indonesia Adalah Negara ke-4 dengan Pemain Genshin Impact Terbanyak," 9 September 2022. [Online]. Available: https://gamerwk.com/salip-jepang-indonesia-adalah-negara-ke-4-dengan-pemain-genshin-impact-terbanyak/. [Accessed 02 August 2023].

[6] R. Chadha, "Twitter Gaming and Esport Insights for the First Half of 2021," 21 June 2021. [Online]. Available: https://blog.twitter.com/en_us/topics/insights/2021/twitter-gaming-and-esports-insights-for-the-first-half-of-2021. [Accessed 5 August 2021].

[7] M. Ahlgren, "50+ Twitter Statistics & Facts For 2020," 25 July 2021. [Online]. Available: https://www.websitehostingrating.com/twitter-statistics/. [Accessed 8 August 2021].

[8] J. Shepherd, "23 Essential Twitter Statistics You Need to Know in 2023," 16 May 2023. [Online]. Available: https://thesocialshepherd.com/blog/twitter-statistics. [Accessed 2 August 2023].

[9] C. M. Annur, "Jumlah Pengguna Twitter di Indonesia Capai 14,75 Juta per April 2023, Peringkat Keenam Dunia," 31 May 2023. [Online]. Available: https://databoks.katadata.co.id/datapublish/2023/05/31/jumlah-pengguna-twitter-di-indonesia-capai-1475-juta-per-april-2023-peringkat-keenam-dunia. [Accessed 3 August 2023].

[10] M. A. Kausar, A. Soosaimanickam and M. Nasar, "Sentiment Analysis on Twitter Data during COVID-19 Outbreak," *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 12, pp. 415-422, 2021.

[11] H. Simorangkir and K. M. Lhaksamana, "Analisis Sentimen pada Twitter untuk Games Online Mobile Legends dan Arena of Valor," *e-Proceeding of Engineering,* vol. 5, pp. 8130-8140, December 2018.

[12] C. Chapple, "Genshin Impact Races Past $1 Billion on Mobile in Less Than Six Months," 23 March 2021. [Online]. Available: https://sensortower.com/blog/genshin-impact-one-billion-revenue.

[13] A. Karami, M. Lundy, F. Webb and Y. Dwivedi, "Twitter and Research: A Systematic Literature Review Through Text Mining," *IEEE Access ,* vol. 8, pp. 67698 - 67717, 26 March 2020.

[14] F. Eibensteiner, V. Ritschl and F. A. Nawaz, "People's Willingness to Vaccinate Against COVID-19 Despite Their Safety Concerns: Twitter Poll Analysis," *J Med Internet Res 2021;23(4):e28973,* vol. 23, no. 4, 29 April 2021.

[15] T. Jo, Text Mining: Concepts, Implementation, and Big Data Challenge, Seoul: Springer, 2018.

[16] D. A. Nurdeni, A. B. Santoso and I. Budi, "Sentiment Analysis on Covid19 Vaccines in Indonesia: From The perspective of Sinovac and Pfizer," *2021 3rd East Indonesia Conference on Computer and Information Technology ,* 2021.

[17] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Document," *International Journal of Computer Application,* vol. 18, pp. 25-29, July 2018.

[18] M. Anggraini, R. A. Tyas, I. A. Sulasiyah and Q. Aini, "Implementasi Algoritma Naive Bayes dalam Penentuan Rating Buku," *SISTEMASI ,* vol. 9, no. 3, pp. 557-566 , September 2020.

[19] H. K, S. Tayal, P. M. George, P. Singla and U. Kose, Bayesian Reasoning and Gaussian Processes for Machine Learning Application, Boca Raton, Florida: CRC Press, 2022.

[20] M. B. Ressan and R. F. Hassan, "Naive-Bayes family for Sentiment Analysis during COVID-19 Pandemic and Classification Tweets," *Indonesian Journal of Electrical Engineering and Computer Science,* pp. 375-383, 2022.

[21] Indeed Editorial Team, "What Is Data Preprocessing? (With Importance and Examples)," 13 March 2023. [Online]. Available: https://ca.indeed.com/career-advice/career-development/data-preprocessing. [Accessed 21 July 2023].