Analysis and Development of Information System for Cyberbullying Tendency on Twitter Social Media Using the Naïve Bayes Approach

***Yulius Hari, Maharani Kusuma Putri, Darmanto**

Informatics Department, Widya Kartika University, Indonesia

***Corresponding Author**

DOI : https://dx.doi.org/10.47772/IJRISS.2024.806116

Received: 22 May 2024; Revised: 05 June 2024; Accepted: 11 June 2024; Published: 10 July 2024

ABSTRACT

Social media news becomes information read by thousands of individuals worldwide. Social media users in society have the freedom to post comments based on their perspectives, and the larger community can see whether these remarks are positive or negative. However, many comments are not constructive, and many of them lead to bullying. Unsupervised communication in the social realm can lead to a variety of deviations, which are commonly referred to as cyberbullying; several incidents of cyberbullying have happened. This research can also be used to express negative feelings about someone by writing them down and sharing them on social media. Methodology used in this research is following the System Development Life Cycles and based on naïve Bayes approach to classify the expression whenever it's a bullying or not. From the research finding the system can help to mitigate the bullying tendencies in social media, although the system cannot predict whenever its actual bullying or a pun made by friends.

Keywords*:* bullying, cyber bullying, SDLC, naïve bayes classifier

INTRODUCTION

In the current technological era, the development of information technology is very rapid. This can be proven by the ease of accessing the internet, one of which is using the internet as a communication medium [1]. Currently, social media is widely known by the public and is also used to carry out daily life. This social media can usually be used individually or simultaneously by various groups ranging from children, teenagers to adults and also the elderly[2]. Because everything we need is available quickly and practically on online service, therefore social media also impacting our view on certain topics. The use of social media can also influence the way we socialize, and can also have positive and negative impacts if it is misused by various parties who have bad wishes for someone[3].

One of the social media that is currently popular is Twitter, in this research we will discuss the "Cyberbullying Tendency Analysis System on Twitter Social Media Using the Naïve Bayes Approach" which often occurs on social media, especially on Twitter. Looking at the existing data, there needs to be real and immediate efforts to prevent the spread of cyberbullying practices in Indonesia, especially for children who are most vulnerable to becoming victims. Most cases of cyberbullying occur on social media, one of which is via Twitter[4].

Cyberbullying is an incident where someone experiences bullying behavior, who is insulted, intimidated, humiliated by other people, this is because cyberbullying itself can intimidate anyone, wherever and whenever the victim is. Because access via the internet or cell phone on social media, especially Twitter[5]. Cyberbullying in question includes negative comments containing cyberbullying elements in certain posts in unfriendly personal messages on Twitter social media[6].

The forms and methods of cyberbullying are very diverse. This could be in the form of a message or comment on the victim's post, for example, "Hey, your face looks like a pig's showing off," commenting by making fun of the victim on other people's social networking accounts to threaten the victim and cause problems for access[7]. The motivations used by the perpetrators also vary, for example because they are angry and want revenge, frustrated, want to get attention, maybe the perpetrators even do it to fill their free time, and there are also several motivations that are carried out, namely for the purpose of joking.

The effects of prolonged cyberbullying can also kill a person's character, making the victim someone who is gloomy, worried, has no sense of security within himself, always feels guilty and a failure[8]. Some victims of cyberbullying even want to end their lives because they can no longer stand the intimidating words given by the perpetrator[9]. Cyberbullying perpetrators usually harass victims who don't like to fight, weak victims who can't defend themselves, and usually the perpetrators are people who feel the most powerful, think they are great, and have a high position/status, while the victims are often teased because of their appearance. victims, skin color, victims' families. However, it is possible that the victim is someone who has outstanding abilities so that the person who is the perpetrator feels jealous[10].

LITERATURE REVIEW

Naïve Bayes Classifier

The Naive Bayes Classifier is a machine learning algorithm used to classify cyberbullying in social media platforms, particularly on Twitter and Instagram. This algorithm is based on Bayes' theorem, which calculates the probability of an event given prior knowledge of conditions that might be related to the event. The Naive Bayes Classifier works by analyzing the text data from social media platforms, such as tweets or comments, and classifying them into two categories: cyberbullying and non-cyberbullying [11].

The algorithm uses a set of predefined rules and a statistical model to identify the likelihood of a comment or tweet being cyberbullying based on its content. Before the classification process, the algorithm performs preprocessing on the text data, which includes tokenizing, case folding, removing stop words, and stemming [12]. This step is crucial in preparing the data for analysis and ensuring that the algorithm can accurately identify the features of the text that are relevant to cyberbullying.

The algorithm then selects the most relevant features from the preprocessed data. These features are the words or phrases that are most commonly associated with cyberbullying. The algorithm uses these features to create a set of parameters or criteria for classification [13]. The Naive Bayes Classifier uses the selected features to classify the text data into cyberbullying or non-cyberbullying. The algorithm calculates the probability of each comment or tweet being cyberbullying based on the features it has identified. It then compares these probabilities to a predefined threshold to determine whether the comment or tweet is classified as cyberbullying or not [14].

The performance of the Naive Bayes Classifier is evaluated using metrics such as accuracy, precision, recall, and error rate. The algorithm's accuracy is measured by comparing its predictions to the actual labels of the data. Precision measures the proportion of true positives (correctly classified cyberbullying comments) out of all positive predictions. Recall measures the proportion of true positives out of all actual cyberbullying comments. Error rate measures the proportion of incorrect predictions [16].

Studies have shown that the Naive Bayes Classifier can achieve high accuracy in classifying cyberbullying on social media platforms. For example, one study reported an accuracy of 98.5% in detecting cyberbullying comments on Instagram using the Naive Bayes Classifier. Another study reported an accuracy of 82.12% in detecting cyberbullying tweets on Twitter using the same algorithm [15]. These results demonstrate the effectiveness of the Naive Bayes Classifier in identifying cyberbullying on social media platforms.

METHODOLOGY

In one study, the Naive Bayes algorithm classification method was used to discuss cyberbullying. The general description of the system that will be created in this research is a system that will be used to classify a comment. The system is created based on the system's requirements to be able to classify a message whether it is categorized as bullying or neutral. The topic raised here is the topic of bullying comments which will later be processed by the system. The following is a schematic of the system that will be built:

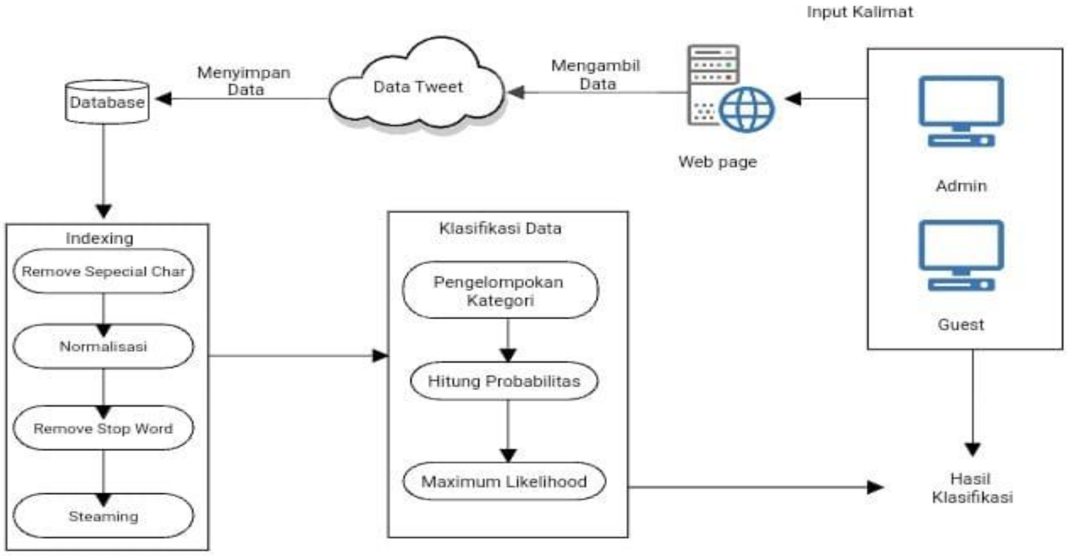


Fig 1. system schematic (source: author research, 2024)

Training data is obtained from tweets data sentences which are carried out using existing data. By inputting sentences and retrieving data in the database, the system will later classify existing sentences and carry out steaming. Next, data classification will be carried out which is used to determine the weight of words or sentences. entered in a system into each existing category [16] until finally it will produce model data that contains probabilities in each vocabulary per category. To avoid the zero value (0), a smoothing technique will be used or smoothing our data by adding numbers, known as additive smoothing, also called Laplace smoothing.

In the classification process two data types are needed, namely model data and testing data. Testing data is obtained from user input where this data is in the form of words or sentences and there is no minimum character. The next process is data grouping: here the testing data entered by the user will be sorted into individual words which will later be checked for probability in the model that has been created previously. Then the maximum likelihood is calculated by changing it to logarithmic form to make calculations easier and reduce errors in calculating numbers below zero (0). For classification, determining whether a sentence falls into one of the categories, look for the maximum likelihood algorithm value that is close to null (0).

Use Case Diagram is a type of UML (Unified Modelling Language) diagram that is used to show the relationship between systems and actors. This diagram can explain the types of interactions that occur between system users and the system. A usage examples are something that is easy to learn. To start modelling, one must create a diagram that can depict the actor's actions with the actions of the system itself, as seen in the Use Case.

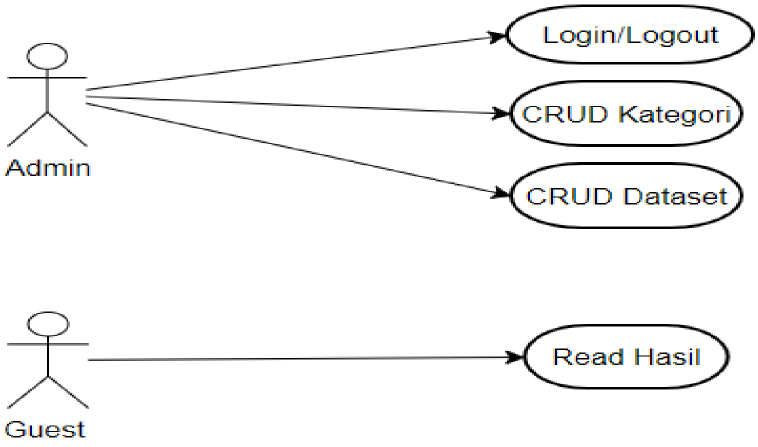


Fig 2. Use case Diagram

The admin actor is the person responsible for managing the website system required as follows:

1. The login feature is only available for admins, where admins can display datasets and categories.
2. Admin can also add datasets or categories that you want to input
3. Then the admin can log out

Guest actors are people who can only see and input sentences on a website, whether the sentences entered are in the bullying or neutral category.

RESULT AND DISCUSSION

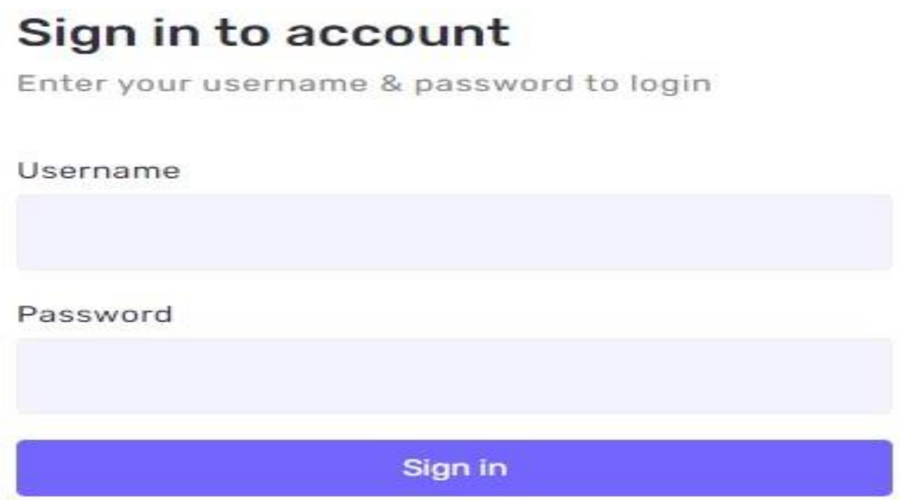
The research is conducted and tested using various computer and devices whenever have a web browser in it. the system is developed in using VS Code with AMPP as server. The system is developed for web-based application, and build with PHP as server-side language and Phyton as data processing language. After the system design is created, the next step is how to implement the system plan that has been created in the form of coding and produce a program that can be used by guests. The features that can be used by guests can be seen in the use case diagram. The following are the results of system implementation and design:

1. Creation of applications for admin

- Admin logs in
- Admin can add the desired category or dataset
- Admin can delete and view datasets and categories

2. Creation of applications for user benefit

- Users can input the desired words/sentences
- Users can see the calculations and classification results whether the word/sentence falls into the bullying or neutral category.

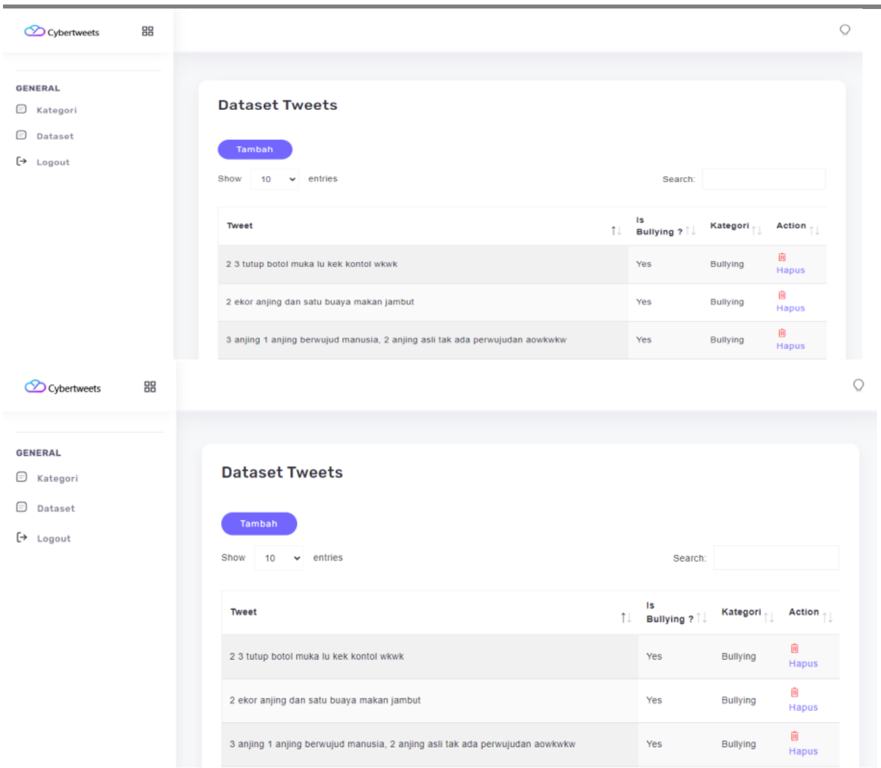


Sign in to account

Enter your username & password to login

Fig 3. Admin Login View (source: author research 2024)

Admin pages can add or expand the data set for a better result. For this research we used 500 sample sentences for each category. The data is collected from various source and edited before used.



Dataset Tweets

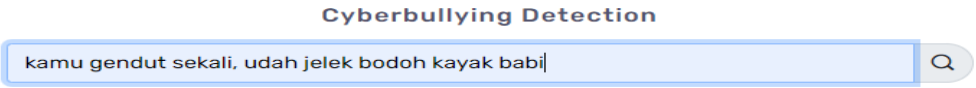
Tweet	Is Bullying ?	Kategori	Action
2 3 tutup botol muka lu kek kontol wkwk	Yes	Bullying	Hapus
2 ekor anjing dan satu buaya makan jambut	Yes	Bullying	Hapus
3 anjing 1 anjing berwujud manusia, 2 anjing asli tak ada perwujudan aowkwkw	Yes	Bullying	Hapus

Dataset Tweets

Tweet	Is Bullying ?	Kategori	Action
2 3 tutup botol muka lu kek kontol wkwk	Yes	Bullying	Hapus
2 ekor anjing dan satu buaya makan jambut	Yes	Bullying	Hapus
3 anjing 1 anjing berwujud manusia, 2 anjing asli tak ada perwujudan aowkwkw	Yes	Bullying	Hapus

Fig 4. Admin page and dataset (source: author research 2024)

For searching the sentences and checking whenever it can be classified as bullying or not, can be shown in Fig 5.

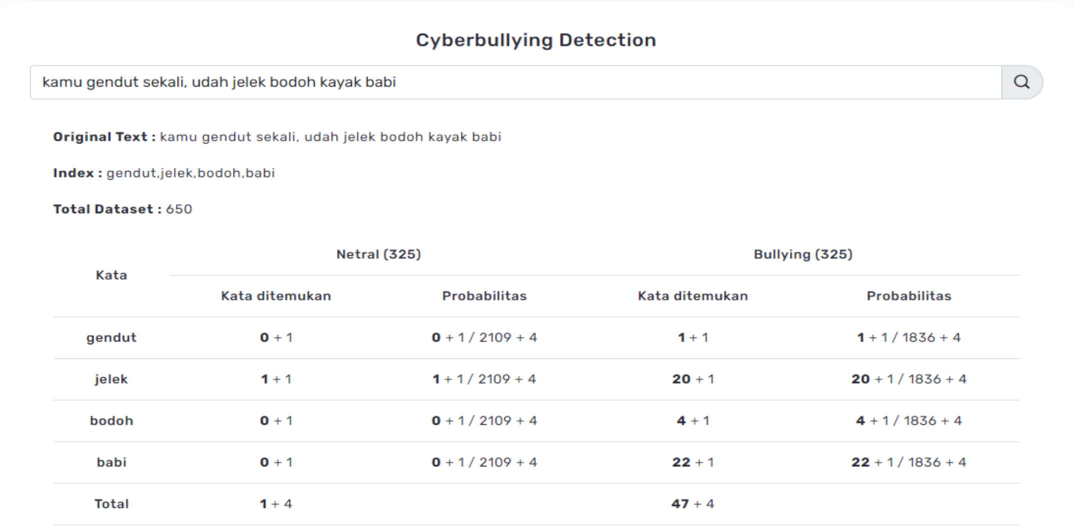


Cyberbullying Detection

kamu gendut sekali, udah jelek bodoh kayak babi

Fig 5 Search Page (source: author research 2024)

The result after we search the system will be calculated and make the category based on the result of the classification. The example of result page and Naïve Bayes calculation can be shown on Fig 6. and Fig. 7 respectively.



Cyberbullying Detection

kamu gendut sekali, udah jelek bodoh kayak babi

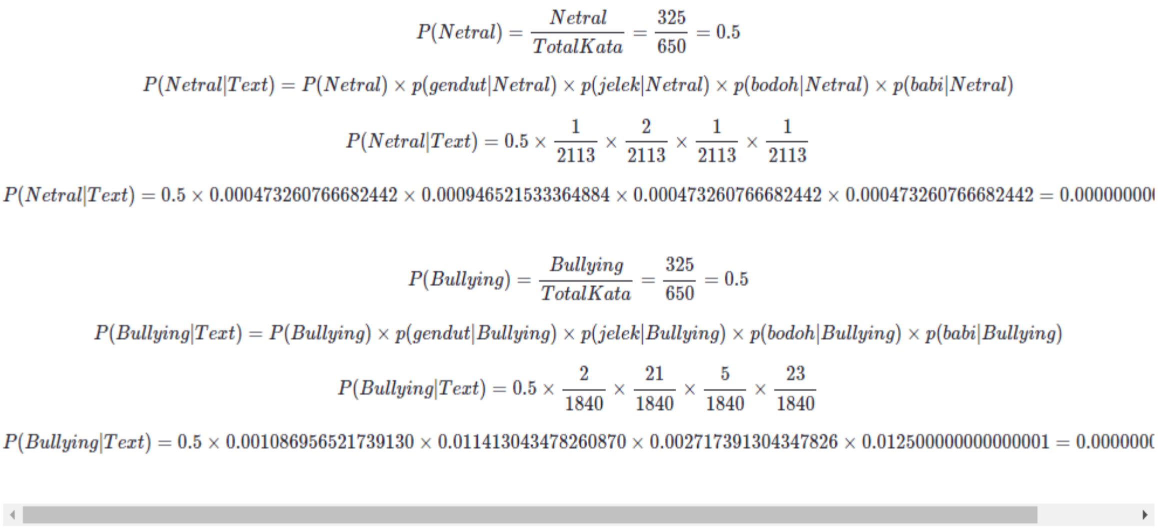
Original Text : kamu gendut sekali, udah jelek bodoh kayak babi

Index : gendut,jelek,bodoh,babi

Total Dataset : 650

Kata	Netral (325)		Bullying (325)	
	Kata ditemukan	Probabilitas	Kata ditemukan	Probabilitas
gendut	**0** + 1	**0** + 1 / 2109 + 4	**1** + 1	**1** + 1 / 1836 + 4
jelek	**1** + 1	**1** + 1 / 2109 + 4	**20** + 1	**20** + 1 / 1836 + 4
bodoh	**0** + 1	**0** + 1 / 2109 + 4	**4** + 1	**4** + 1 / 1836 + 4
babi	**0** + 1	**0** + 1 / 2109 + 4	**22** + 1	**22** + 1 / 1836 + 4
Total	**1** + 4		**47** + 4	

Fig 6. Category Result Page (source: author research 2024)



$$P(Netral) = \frac{Netral}{TotalKata} = \frac{325}{650} = 0.5$$

$$P(Netral|Text) = P(Netral) \times p(gendut|Netral) \times p(jelek|Netral) \times p(bodoh|Netral) \times p(babi|Netral)$$

$$P(Netral|Text) = 0.5 \times \frac{1}{2113} \times \frac{2}{2113} \times \frac{1}{2113} \times \frac{1}{2113}$$

$$P(Netral|Text) = 0.5 \times 0.000473260766682442 \times 0.000946521533364884 \times 0.000473260766682442 \times 0.000473260766682442 = 0.00000000$$

$$P(Bullying) = \frac{Bullying}{TotalKata} = \frac{325}{650} = 0.5$$

$$P(Bullying|Text) = P(Bullying) \times p(gendut|Bullying) \times p(jelek|Bullying) \times p(bodoh|Bullying) \times p(babi|Bullying)$$

$$P(Bullying|Text) = 0.5 \times \frac{2}{1840} \times \frac{21}{1840} \times \frac{5}{1840} \times \frac{23}{1840}$$

$$P(Bullying|Text) = 0.5 \times 0.001086956521739130 \times 0.011413043478260870 \times 0.002717391304347826 \times 0.012500000000000001 = 0.000000$$

Fig 7 Naïve Bayes Calculation Result Page (source: author research 2024)

Feedback User

Table 1 Questionnaire

o	**Question**	**Percentage**	**Average**
1	What is your opinion regarding the website that you have simulated?	83%	4,18
2	Are the features of this application easy to understand?	81%	4,09
3	Have you encountered any problems on this website?	83%	4,15
4	Is the operation familiar?	80%	4
5	Is this application able to help you categorize news?	87%	4,39
6	Are the results of this application understandable?	85%	4,27
7	What do you think about the appearance of this website?	80	4

The average feasibility result from this questionnaire is 83%. With a total of 33 respondents. Data was collected with the help of a questionnaire from respondents ranging in age from 16 – 22 years.

Table 2 Rating scale

Percentage	**Information**
0% - 19.99%	Very less
20% - 39,99%	Not enough
40% - 59,99%	Enough
60% - 79,99%	Good
80% - 100%	Very Good

CONCLUSION

From the results of creating this system, it can be concluded that the system created can be classified into 2 categories, namely bullying and neutral. The results of the classification can be determined by how large the existing dataset is, where the Naïve Bayes classification method uses weight calculations per category. The average result of the feasibility of this questionnaire is 83% with a total of 33 respondents.

REFERENCES

1. Khairunnisa, S., Adiwijaya, A., & Al Faraby, S. (2021). Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19). Jurnal Media Informatika Budidarma, 5(2), 406-414
2. Alifia, A., Cholissodin, I., & Adikara, P. P. Analisis Sentimen terhadap Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) Level 3 berdasarkan Data Twitter menggunakan Algoritma Naïve Bayes. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, 2548, 964X.
3. Renault, Thomas. "Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages." Digital Finance 2.1-2 (2020): 1-13.
4. Samsir, Samsir, et al. "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes." Jurnal Media Informatika Budidarma 5.1 (2021): 157-163.
5. ABDULLOH, Nassharih; HIDAYATULLAH, Ahmad Fathan. Deteksi Cyberbullying pada Cuitan Media Sosial Twitter. AUTOMATA, 2020, 1.1.
6. Maulana, F. A., & Ernawati, I. (2020, November). Analisa sentimen cyberbullying di jejaring sosial twitter dengan algoritma naïve bayes. In Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya (Vol. 1, No. 2, pp. 529-538).
7. Hutagalung, A. S., Negara, A. B. P., & Pratama, E. E. (2021). Aplikasi Pendeteksi Cyberbullying Terhadap Komentar Postingan Media Sosial Instagram dengan Metode Naïve Bayes Classifier Berbasis Website. JUSTIN (Jurnal Sistem Dan Teknologi Informasi), 9(3), 364-371.
8. Pebrianto, J. (2023). SENTIMENT ANALYSIS OF SERVICE PROVIDER ON TWITTER TWEET USING NAIVE BAYES CLASSIFIER WITH PHP. Journal of Innovation and Future Technology (IFTECH), 5(2), 13-23.
9. Rahman, O. H., Abdillah, G., & Komarudin, A. (2021). Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 5(1), 17-23.
10. Nurhuda, F., Sihwi, S. W., & Doewes, A. (2016). Analisis sentimen masyarakat terhadap calon Presiden Indonesia 2014 berdasarkan opini dari Twitter menggunakan metode Naive Bayes Classifier. ITSmart: Jurnal Teknologi dan Informasi, 2(2), 35-42.
11. Kurniawan, T. (2017). Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Media Mainstream Menggunakan Naïve Bayes Classifier Dan Support Vector Machine. Institut Teknologi Sepuluh Nopember. Kurniawan, T. (2017). Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Media Mainstream Menggunakan Naïve Bayes Classifier Dan Support Vector Machine. Institut Teknologi Sepuluh Nopember.
12. Riadi, I., & Kom, M. (2017). Analisis Bukti Digital Cyberbullying pada Jejaring Sosial Menggunakan Naïve Bayes Classifier (NBC). Riadi, I., & Kom, M. (2017). Analisis Bukti Digital Cyberbullying pada Jejaring Sosial Menggunakan Naïve Bayes Classifier (NBC).
13. Cahyono, A. S. (2016). Pengaruh media sosial terhadap perubahan sosial masyarakat di Indonesia. Publiciana, 9(1), 140-157.
14. Putri, W. S. R., Nurwati, N., & Santoso, M. B. (2016). Pengaruh media sosial terhadap perilaku remaja. Prosiding Penelitian dan Pengabdian kepada Masyarakat, 3(1).
15. Putri, A., & Muzakir, A. (2022). ANALISIS SENTIMEN CYBERBULLYING KPOP DI MEDIA SOSIAL TWITTER MENGGUNAKAN METODE NAIVE BAYES. Journal of Syntax Literate, 7(9).
16. Hari, Y., Hermawan, B., & Paramitha, M. (2022, October). INTERPRETATION OF STUDENTS ABILITY TO IDENTIFY HOAXES AND INFORMATION DISORDER DURING THE COVID-19 PANDEMIC. In Proceeding of the International Conference on Innovation in Open and Distance Learning (Vol. 2).