Electronic Integrated Computer Algorithm Journal
Vol. 2, No. 1, October 2024, pp 52~60

ISSN: 3031-0350, DOI: 10.62123/enigma.v2i1.39

Sentiment Analysis of Hate Speech against DPR-RI on Twitter Using Naive Bayes and KNN Algorithms

Joy Lousia Brigitha Munthe[1*], Kristin Sinaga[1], Santi Prayudani[1]

[1] *Department of Computer Engineering and Informatics*, *Politeknik Negeri Medan,* Indonesia

DOI : 10.62123/enigma.v2i1.39

Received : September 09, 2024
Revised : October 23, 2024
Accepted : October 26, 2024

Keywords:
Twitter, Sentiment, K-Nearest Neighbours, Crawling, Naïve Bayes Classification.

ABSTRACT

According to surveys, the rise in social media users has resulted in an increase of hate speech, and Twitter is one of the most popular platforms for this type of speech. The tweet feature on Twitter enables users to make repeated instances of hate speech, making Twitter data very intriguing to analyze. This study aims to investigate whether a tweet contains hate speech towards the Indonesian House of Representatives (DPR-RI). The research employed crawling techniques to gather data from Twitter using the Twitter API feature. The Naïve Bayes algorithm was applied, and the results were compared with the accuracy of the K-Nearest Neighbor. After preprocessing, the total data obtained was 1,494, with 956 test data and 538 training data. The study revealed that Twitter users' sentiment towards DPR-RI was 49.2% positive and 50.8% negative sentiment when tested using Naïve Bayes. Meanwhile, KNN showed 23.4% positive and 76.6% negative sentiment. The high negative sentiment in both classifiers suggests that Twitter users frequently express hate speech towards DPR-RI. Naïve Bayes algorithm showed the highest prediction accuracy at 98.32%, while the K-Nearest Neighbor algorithm had an accuracy of only 62.84%.

1. INTRODUCTION

Social media is one of the platforms used for users to interact with each other. Currently, there are a large number of users who use social media, including Twitter [1]. Twitter provides a feature called Tweet which is similar to a post with a limit of 280 characters. Usually, tweets are concise and straightforward [2]. Tweets can contain expressions, opinions, and even criticism about a particular issue. One of the issues that has become trending on Twitter is about the Indonesian House of Representatives (DPR-RI). Many tweets contain opinions and criticism about the DPR-RI that lead to hate speech. Twitter is a place where people can freely express their opinions. There are many analytical methods that can be used to analyze public opinion based on information available on social media such as Twitter. One of them is sentiment analysis method [3].

Detecting a tweet that contains hate speech is not an easy task, as it requires in-depth indications to avoid it turning into accusations or controversy. This requires a special approach to analyzing tweets from year to year. The data can be quite extensive and requires a significant amount of time to search, process, and develop concrete algorithms to process the data [4]. Meanwhile, the classification of research that applies the KNN and Naive Bayes algorithms to detect hate speech in general on social media Twitter expects that the KNN and Naive Bayes Algorithms can learn from the data collected from Twitter and obtain accuracy.

2. LITERATURE REVIEW

2.1 Sentiment Analysis

Sentiment anlysis, also known as opinion mining, is a computational study to recognize and express opinions, sentiments, evaluations, atitudes, emotions, subjectivity, judgements, or view that are found in a text [5]. The analysis's sources are social media platforms, such as websites that host reviews, forum discussions, blogs, microblogs, Twitter, and more. Due to its opinionated data, which allows users to obtain reviews for any service that is helpful for their everyday life, this study subject is becoming increasingly popular. Digital formats are used to store the vast quantity of biased data. Sentiment analysis, which links data mining operations to specific topics or opinions, produces results. Research on feeling, mind extraction, and emotion-based summarization is done for sentiment analysis. Because sentiment analysis is so well-known, it may also be helpful in surveys and marketing campaigns by determining the success rate of any product or service based on people's suggestions or opinions.It also provides information on what customers like and dislike, which helps the company come up with much clearer ideas about the qualities of its products [6].

2.2 Data Mining

Data mining is the analysis of data (usually large amounts of data) to discover clear relationships and previously unknown insights in a new way that is understood and useful to the owner of the data [7] [8]. The process of employing certain tools or

*Corresponding Author Email: joymunthe1@gmail.com

approaches to search through chosen data for interesting patterns or information is known as data mining. Data mining is a process that uses one or more machine learning algorithms to automatically analyse and extract knowledge. A technique for extracting knowledge from an existing database is called Knowledge Discovery in Databases (KDD). A few tables in the database are related to one another. The information gathered from this approach can serve as a knowledge base for making decisions. In order to discover new patterns or models in a big database that are legitimate (perfect), practical, and comprehensible, data mining is an interactive, iterative process [9]. The following are significant aspects of data mining:

1. Data mining is an automated process using preexisting data.
2. There are enormous amounts of data that need to be processed.
3. Finding links or patterns that could offer helpful hints is the goal of data mining.

Certain tools identify these trends and can offer insightful and helpful data analysis. These findings can then be investigated further, possibly utilizing additional decision support tools.

2.3 K-Nearest Neighbors classification

Groups data based on the distance to its nearest neighbors [10] [11] [12]. The distance calculation is performed using the Euclidean Distance method. The formula used is:

$$D\,(X,Y) = \sqrt{\sum_{k\,=\,1}^{n}(Xk - Yk)^2} \quad (1)$$

Where:
D = Distance between two points, X and Y
X = Test data
Y = Sample data
n = Dimension of the data [13]

2.4 Naïve bayes classification

Naive Bayes Classifier works by calculating the posterior probability of each possible class based on the observed feature values in the instance [14] [15]. he class with the highest posterior probability is then selected as the predicted class for the instance. This method is well-suited for sentiment classification in this journal due to its several advantages, including simplicity, speed, and high accuracy.

The formula used for Naive Bayes Classifier is as follows:

$$P(X|H) = \frac{P(X|H)P(H)}{P(X)} \quad (2)$$

Where:
X = Data with unknown class
H = Hypothesis that data X belongs to a specific class
P(H|X) = Probability of hypothesis H given the condition X
P(H) = Probability of hypothesis H
P(X|H) = Probability of X given hypothesis H
P(X) = Probability of X

2.5 *Rapid Miner*

Rapid miner s an open source application used for data processing, including data mining, text mining, and predictive analysis. It provides various algorithms for data analysis, such as Naïve Bayes, k-nearest neighbors, and fuzzy clustering, making it unnecessary for users to have coding skills for data processing. RapidMiner also provides formulas used in the available algorithms, which eliminates the need for manual calculation by users [16] [17] [18] [19] [20]. In this journal, the calculation for the Naïve Bayes algorithm was performed using RapidMiner.

3. RESEARCH METHODS

To analyze and classify hate speech directed towards the Indonesian parliament (DPR-RI) in tweets, a method was developed that enables automatic text data classification through several stages to produce the best classification results. As shown in Figure 1 Research Structure, the process starts with crawling, followed by text pre-processing and process and classification.

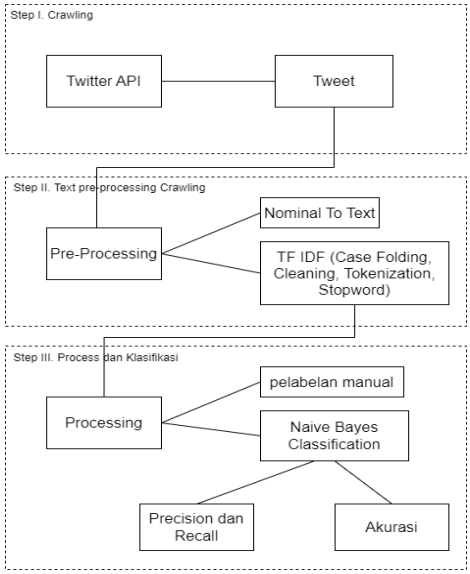

Figure 1. Research Structure

Explanation of Figure 1 Research Structure:

1. Step 1. Crawling, his step is performed to collect data based on an index. In this study, the crawling process used is the 'Twitter API'. Twitter API is very helpful in collecting data on Twitter by simply typing in the targeted keyword, then the data will be collected based on its index[21].
2. Step 2. ext. pre-processing Crawling, after the Crawling step, the process continues with the extraction process by counting words. In this study, preprocessing is used to filter the data to be analyzed, to avoid inaccurate data to become more easily understood data. The following are the stages in preprocessing:
   a) Converting nominal to text.
   b) Case Folding, a stage used to change all letters in the text to lowercase.
   c) Cleaning, this stage is used to eliminate inconsistent data by removing characters other than letters, such as hashtags, URLs, hashtags, and emojis.
   d) Tokenization, in this stage, each word in a text is divided. The result of this stage will display words without characters other than letters[22].
   e) Stop word, this stage is used to select every word listed on the stop list and filter words. The stop list used in this study is taken from Kaggle (https://www.kaggle.com/datasets/oswinrh/indonesian-stoplist).
   f) TF-IDF, the five stages above are part of the TF-IDF process in preprocessing, which is to weight words used to separate characteristics from a text.
3. Step 3. Process and Classification, the classification process uses Naïve Bayes to classify words and the accuracy of the truth of a word. The following are the stages of the Naïve Bayes classification:
   a) Manually labeling data for use as training data.
   b) Reading training data and creating a data model.
   c) Performing automatic classification on data that has no label based on training data.
   d) The classification is performed by calculating how often data appears or, more specifically, calculating the probability of a word that appears in a tweet.

4. RESULTS AND DISCUSSIONS

The data mining process was conducted using the naive Bayes classification algorithm, which is commonly used in sentiment analysis to obtain accuracy, patterns, and human behavior based on the probability of word occurrence.

4.1 *Crawling Process*

The first step before conducting classification was crawling. In this study, the main keyword used for crawling was 'DPR RI.' The crawling process was conducted by connecting a Twitter account with the Twitter API and accessing the token.

Due to the limit of Twitter API, 1500 tweets related to the keyword 'DPR RI' were collected, focusing on the latest and most popular posts from March 2023. This specific period was chosen because it marked a time of heightened political discussions and controversies involving DPR-RI, which is likely to generate relevant public sentiment and hate speech. By focusing on March 2023, the study aims to capture a snapshot of recent public discourse. Moreover, this timeframe allows the analysis to stay within the technical constraints of the API while ensuring the data reflects significant and timely events. Some of the results can be seen in table 1.

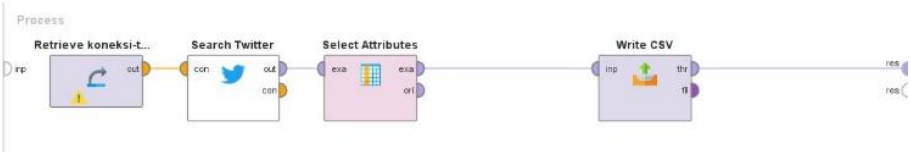

Figure 2. Crawling Process

Table 1. Result of Crawling

Username	***Tweet***
akuluthfi	Shame on you @DPR_RI
haruasakamol	@MARQUEZ__93 @iqbalbaqo @KPK_RI @KejaksaanRI @Kemenkumham_RI @PolhukamRI @DPR_RI Sampai mahluk halus seperti setanpun takut sama one ABUD
andiebs	@YanHarahap @DPR_RI @Edhie_Baskoro @PDemokrat @AgusYudhoyono betul....mas Agus pingin jadi cawapres
usman_alfath	@DirtyTime2019 Hahaha kq DPR RI...biar apa..biar bs dduk disenayan?? Ngajak taruhan nya gk rasional..mana masih jauh lagi pemilunya...hdeeuuuhh cebong emang otaknya kwadrat tololnya.
…..	.....
…..	.....
opinirakyat2024	@VIVAcoid Bukti lemahnya negara tak mampu melarang impor pakaian bekas cu @DPR_RI @MUIPusat @muhammadiyah @nahdlatululama https://t.co/mzdIVqbMot
suhermanidham	@03__nakula @DPR_RI Orang ini korupsinya apa ya? Mungkin penyidikan @KPK_RI lebih diperdalam karena jangan2 org ini adalah sebagai penghubung.

4.2 Pre-processing

The data generated from the crawling process is still messy and contains many characters that could disrupt the classification process. Therefore, preprocessing is needed to perform filtering, case folding, cleaning, tokenization, and stopword removal, all of which are part of the TF-IDF process. The results can be seen in Figure 3.

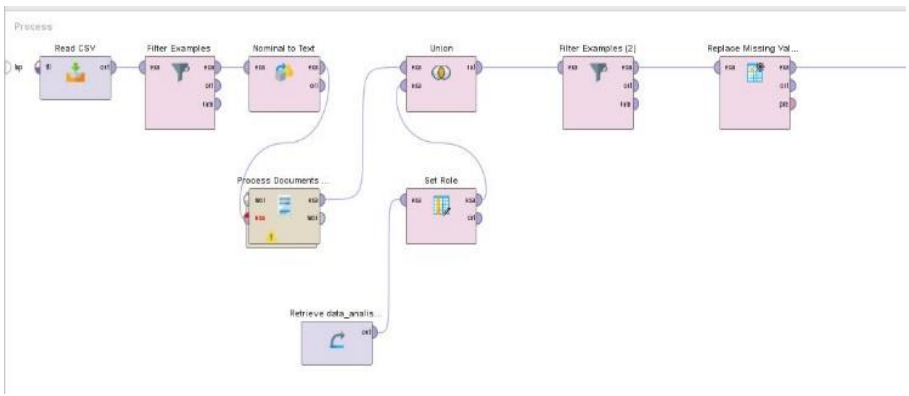

Figure 3. TF-IDF Process

Table 2. Result of Pre-processing

Username	***Tweet***
akuluthfi	Shame on you DPRRI
haruasakamol	Sampai mahluk halus seperti setanpun takut sama one ABUD
andiebs	betul....mas Agus pingin jadi cawapres
usman_alfath	Hahaha kq DPR RI...biar apa..biar bs dduk disenayan?? Ngajak taruhan nya gk rasional..mana masih jauh lagi pemilunya...hdeeuuuhh cebong emang otaknya kwadrat tololnya.
jgnblgsyp2	Negeri wakanda lagi demam nasionalisme, makanya org kyk gini dipuji-puji terus

…..	…..
…..	…..
kicausunyi	Moga tunjangan untuk DPR dan pejabat RI segera naik....
adipriyono95	Anaknya, sepupunya, iparnya, tanya jg dong...

4.3 K-Nearest Neighbor Classification

When performing classification in K-Nearest Neighbor (KNN), it is essential to have training data that will be used as a measure for the data to be classified. In this study, 538 tweets were collected, and manual labeling was performed to become the training data. The data was divided into two sentiment labels, positive and negative. Positive sentiment contains supportive sentences and positive words, while negative sentiment contains criticisms, comments that lead to hate speech. After filtering and grouping, the total data became 534 tweets. The results are shown in Table 3, which displays the manual labeling.

Table 3. Manual Labeling

Sentimen	***Tweet***
negatif	Shame on you DPRRI
positif	betul....mas Agus pingin jadi cawapres
negatif	Sampai mahluk halus seperti setanpun takut sama one ABUD
…..	…..
…..	…..
positif	Kamu nanya kamu bertanya tanya
positif	Keren ...salute rakyat buat kalian
negatif	Orang ini korupsinya apa ya? Mungkin penyidikan lebih diperdalam karena jangan2 org ini adalah sebagai penghubung.

The data that has been manually labeled has become the training data and has been divided based on sentiment, out of 534 tweets there are 245 positive sentiments and 289 negative sentiments. This can be seen in Figure 4, the Data Training graph.

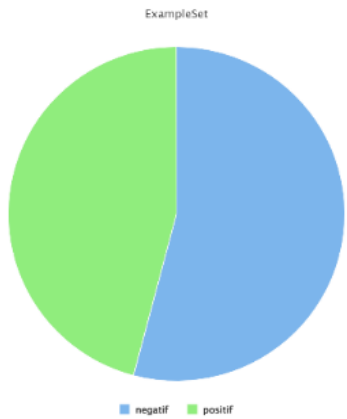

Figure 4. Data Training Graph

The data that has been labeled manually will be classified into KNN and the model will be created through the classification that has been done. This model can serve as a bridge to perform sentiment classification automatically. There are 956 tweet data that will be used as test data and given automatic classification.

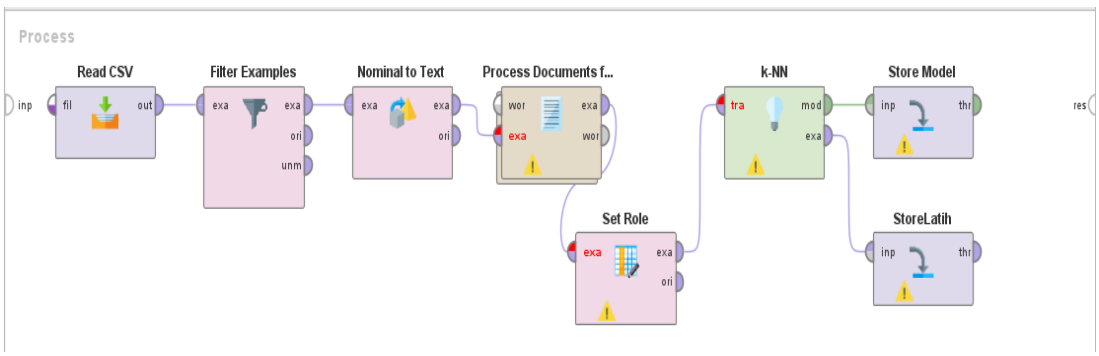

Figure 5. KNN Classification Automatically

The training data and test data will be intersected to find attributes that are interconnected to each other and then classified through KNN using the model that has been created when classifying the test data. The results are shown in table 4.

Table 4. Table Result of KNN Classification Automatically

Text	***Confidence (negatif)***	***Confidence (positif)***	***Predection (sentimen)***
coba periksa ganjar ktp keterangan mantan dpr setya novanto	0.627	0.373	negatif
maaf maaf maaf	0.812	0.188	negatif
pls jelasin gue orang anti pacaran nomor ipb	0.420	0.580	positif
mobil komando sekelas carry butut phk presiden ente klo ngelawak liat jam kek susah ketawa jam segini mah	0.613	0.387	negatif
gua bener bener biro jodoh maaf banget	0.808	0.192	negatif

The accuracy level resulting from the prediction can be tested by comparing the training data and test data and performing reclassification using the K-Nearest Neighbor Algorithm to obtain the percentage of accuracy of the prediction generated by K-Nearest Neighbor. There are a total of 958 data sets after reclassification. The formula for measuring precision and recall :

$$Precision = \frac{true\ positive}{true\ positive + false\ positive}\ x\ 100\% \quad (4)$$

$$Recall = \frac{true\ positive}{true\ positive + false\ negative}\ x\ 100\% \quad (5)$$

The results of precision are shown in table 4 Precision and class recall :

Table 5. Precision and Recall

	True negatve	***True positive***	***Class Predection***
Pred. negative	466	346	57,39 %
Pred. positive	10	136	93,15 %
Class recall	97,90%	28,22%	

According to the table in table 4 it can be observed that :

1. the ratio of negative correct predictions compared to the overall negative predicted results is 57.39%.
2. the ratio of positive sentiment correct predictions compared to the overall positive predicted results is 93.15%.
3. the ratio of negative true prediction compared to the overall negative true data is 97.90%.
4. the ratio of positive true predictions compared to the overall positive true data is 28.22%.

Based on the Precision and Recall results mentioned earlier, the accuracy of the Naïve Bayes classification can be calculated using the provided formula :

$$Accuracy = \frac{true\ positif + true\ negatif}{true\ positif + false\ positif + true\ negatif + false\ positif}\ x\ 100\% \quad (6)$$

$$Accuracy = \frac{602}{958}\ x\ 100\%$$

$$Accuracy = 62{,}84\%$$

From the calculation above, it can be concluded that the accuracy of KNN classification is 62.84%. This indicates that the ratio of correct predictions for both positive and negative sentiments compared to the total data is 62.84%.

4.4 Naïve Bayes Classification

Similar to KNN in classifying naïve bayes also really needs training data. data that has been manually labeled is taken the same as in KNN so that the level of accuracy can be compared.

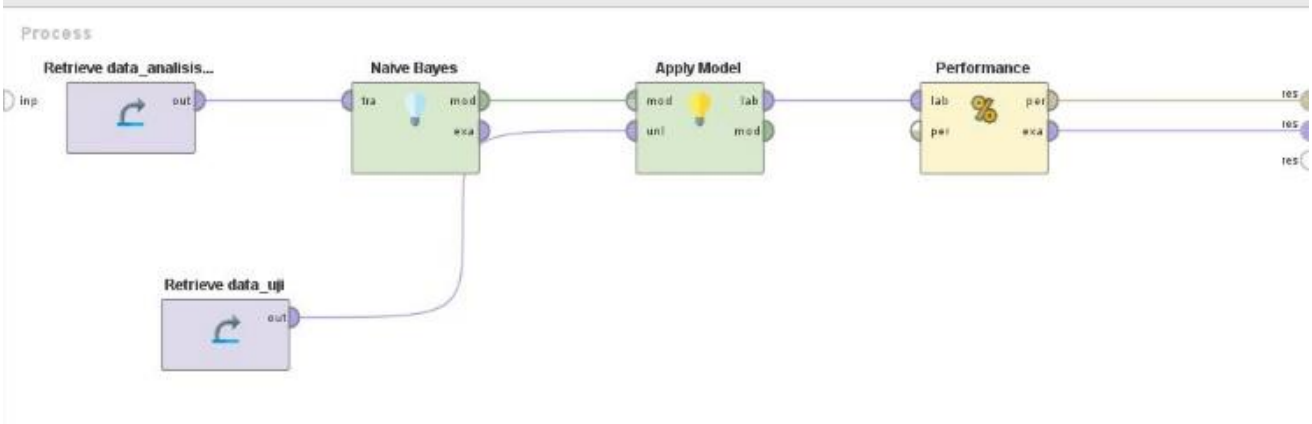

Figure 6. Model linkage and naïve bayes algorithm

There are 956 tweet data that will be used as test data and given automatic classification. The training data and test data will be intersected to find attributes in the form of continuous connections and then classified through naive bayes using the model that has been created when classifying the test data. The result will be as shown in Table 6.

Table 6. Table of Naïve Bayes Classification Automatically

Text	*Confidence (negative)*	*Confidence (positive)*	*Predection (sentiment)*
coba periksa ganjar ktp keterangan mantan dpr setya novanto	0.0	1.0	positive
maaf maaf maaf	1.0	0.0	Negative
pls jelasin gue orang anti pacaran nomor ipb	0.0	1.0	positive
mobil komando sekelas carry butut phk presiden ente klo ngelawak liat jam kek susah ketawa jam segini mah	1.0	0.0	negative
gua bener bener biro jodoh maaf banget	1.0	0.0	negative

Using the same training data, the accuracy level of the classification results can be tested by comparing the training data and test data, and performing classification again using the Naïve Bayes algorithm. The precision and class recall results are shown in Table 7.

Table 7. Precision and Recall

	True negative	*True positive*	*Class Predection*
pred. negative	468	9	98,11 %
pred. positive	8	473	98,34 %
class recall	98,32%	98,13%	

The table in Table 7 shows that :

1. The ratio of negative correct predictions compared to the overall negative predicted results is 98.11%.
2. The ratio of positive sentiment correct predictions compared to the overall positive predicted results is 98,34%.
3. The ratio of negative true prediction compared to the overall negative true data is 98,32%
4. The ratio of positive true predictions compared to the overall positive true data is 98,13%.

From the Precision and Recall results above, the accuracy of the naïve bayes classification can be calculated using the formula :

$$Accuracy = \frac{true\ positive + true\ negative}{true\ positive + false\ positive + \ true\ negative + \ false\ positive}\ x\ 100\%$$

$$Accuracy = \frac{941}{958}\ x\ 100\%$$

$$Accuracy = 98{,}23\%$$

So, from the above calculations, the accuracy is 98.23%. Indicates that the ratio of positive and negative correct predictions with all data has a value of 98.23%.

5. CONCLUSSION

This research has used the K-Nearest Neighbour and Naïve Bayes algorithms for sentiment classification of hate speech towards the Indonesian Parliament (DPR-RI). The sentiment analysis was conducted using data obtained from Twitter through the Twitter. API with the main keyword "DPR RI". The crawling process was conducted to obtain the data, which was then pre-processed and classified to obtain the training and test data. The training data was used to perform automatic classification on the unclassified test data. The level of accuracy of the K-Nearest Neighbour and Naïve Bayes algorithms was compared based on the results of the automatic classification. The K-Nearest Neighbour algorithm classified 23.4% of the sentiments as positive and 76.6% as negative, with a precision of 57.39% for negative sentiments and 93.15% for positive sentiments. The recall rate, which is the ratio of correctly predicted instances to the total instances, was 97.90% for negative sentiments and 28.22% for positive sentiments. Thus, the overall accuracy of the classification was 62.84%.

The Naïve Bayes algorithm classified 49.7% of the sentiments as positive and 50.3% as negative, with a precision of 98.11% for negative sentiments and 98.34% for positive sentiments. The recall rate was 98.32% for negative sentiments and 98.13% for positive sentiments. Thus, the overall accuracy of the classification was 98.23%. The high number of negative sentiments classified by both the K-Nearest Neighbour and Naïve Bayes algorithms indicates that there is a significant amount of hate speech towards DPR-RI on Twitter. The Naïve Bayes algorithm had the highest accuracy rate of 98.32%, while the K-Nearest Neighbour algorithm had an accuracy rate of only 62.84%. Therefore, the Naïve Bayes algorithm is more accurate for analysing hate speech sentiment towards DPR-RI on Twitter.

REFERENCES

[1] B. Auxier and M. Anderson, "Social Media Use in 2021." Accessed: Oct. 06, 2024. [Online]. Available: https://www.pewresearch.org/

[2] K. S. Nugroho, F. A. Bachtiar, and W. F. Mahmudy, "Detecting Emotion in Indonesian Tweets: A Term-Weighting Scheme Study," *Journal of Information Systems Engineering and Business Intelligence*, vol. 8, no. 1, pp. 61–70, Apr. 2022, doi: 10.20473/jisebi.8.1.61-70.

[3] Z. Mansur, N. Omar, and S. Tiun, "Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities," 2023, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2023.3239375.

[4] A. F. Hidayatullah, S. Cahyaningtyas, and A. M. Hakim, "Sentiment Analysis on Twitter using Neural Network: Indonesian Presidential Election 2019 Dataset," *IOP Conf Ser Mater Sci Eng*, vol. 1077, no. 1, p. 012001, Feb. 2021, doi: 10.1088/1757-899x/1077/1/012001.

[5] S. Pandya and P. Mehta, "A Review On Sentiment Analysis Methodologies, Practices And Applications," *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, vol. 9, p. 2, 2020, [Online]. Available: www.ijstr.org

[6] R. Marcec and R. Likic, "Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines," *Postgrad Med J*, vol. 98, no. 1161, pp. 544–550, Jul. 2022, doi: 10.1136/postgradmedj-2021-140685.

[7] P. Bist and A. Prambudi, "Implementation Of Data Mining On Glasses Sales Using The Apriori Algorithm," *International Journal of Cyber and IT Service Management (IJCITSM)*, vol. 1, no. 2, pp. 159–172, 2021, doi: 10.34306/ijcitsm.v1i1.46.

[8] T. H. Sinaga, A. Wanto, I. Gunawan, S. Sumarno, and Z. M. Nasution, "Implementation of Data Mining Using C4.5 Algorithm on Customer Satisfaction in Tirta Lihou PDAM," *Journal of Computer Networks, Architecture, and High-Performance Computing*, vol. 3, no. 1, pp. 9–20, Jan. 2021, doi: 10.47709/cnahpc.v3i1.923.

[9] H. 1, T. Wahyuningsih , and E. Rahwanto, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer." [Online]. Available: http://archive.ics.uci.edu/ml.

[10] H. Wisnu, M. Afif, and Y. Ruldevyani, "Sentiment analysis on customer satisfaction of digital payment in Indonesia: A comparative study using KNN and Naïve Bayes," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Feb. 2020. doi: 10.1088/1742-6596/1444/1/012034.

[11] F. M. J. M. Shamrat *et al.*, "Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 1, pp. 463–470, Jul. 2021, doi: 10.11591/ijeecs.v23.i1.pp463-470.

[12] I. Prayoga, M. D. Purbolaksono, and A. Adiwijaya, "Sentiment Analysis on Indonesian Movie Review Using KNN Method With the Implementation of Chi-Square Feature Selection," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 7, no. 1, p. 369, Jan. 2023, doi: 10.30865/mib.v7i1.5522.

[13] Z. Rais, R. N. Said, and R. Ruliana, "Text Classification on Sentiment Analysis of Marketplace SHOPEE Reviews On Twitter Using K-Nearest Neighbor (KNN) Method," *JINAV: Journal of Information and Visualization*, vol. 3, no. 1, pp. 1–8, Jul. 2022, doi: 10.35877/454ri.jinav1389.

[14] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, "The feature extraction for classifying words on social media with the Naïve Bayes algorithm," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 3, pp. 1041–1048, Sep. 2022, doi: 10.11591/ijai.v11.i3.pp1041-1048.

[15] Samsir *et al.*, "Naives Bayes Algorithm for Twitter Sentiment Analysis," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jun. 2021. doi: 10.1088/1742-6596/1933/1/012019.

[16] M. H. Santoso, "Application of Association Rule Method Using Apriori Algorithm to Find Sales Patterns Case Study of Indomaret Tanjung Anom," *Brilliance: Research of Artificial Intelligence*, vol. 1, no. 2, pp. 54–66, Dec. 2021, doi: 10.47709/brilliance.v1i2.1228.

[17] A. A. Aldino, D. Darwis, A. T. Prastowo, and C. Sujana, "Implementation of K-Means Algorithm for Clustering Corn Planting Feasibility Area in South Lampung Regency," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1742-6596/1751/1/012038.

[18] M. Uska, R. Wirasasmita, U. Usuluddin, and B. Arianti, "Evaluation of Rapidminer-Aplication in Data Mining Learning using PeRSIVA Model," *Edumatic: Jurnal Pendidikan Informatika*, vol. 4, no. 2, pp. 164–171, Dec. 2020, doi: 10.29408/edumatic.v4i2.2688.

[19] N. Baharun, N. F. M. Razi, S. Masrom, N. A. M. Yusri, and A. S. A. Rahman, "Auto Modellingfor Machine Learning: A Comparison

Implementation between Rapid Miner and Python," *International Journal of Emerging Technology and Advanced Engineering*, vol. 12, no. 5, pp. 15–27, May 2022, doi: 10.46338/ijetae0522_03.

[20] S. Kurniawan, W. Gata, D. A. Puspitawati, I. K. S. Parthama, H. Setiawan, and S. Hartini, "Text Mining Pre-Processing Using Gata Framework and RapidMiner for Indonesian Sentiment Analysis," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, May 2020. doi: 10.1088/1757-899X/835/1/012057.

[21] T. D. Dikiyanti, A. M. Rukmi, and M. I. Irawan, "Sentiment analysis and topic modeling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Mar. 2021. doi: 10.1088/1742-6596/1821/1/012054.

[22] Y. Deta Kirana and S. Al Faraby, "Sentiment Analysis of Beauty Product Reviews Using the K-Nearest Neighbor (KNN) and TF-IDF Methods with Chi-Square Feature Selection," *OPEN ACCESS J DATA SCI APPL*, vol. 4, no. 1, pp. 31–042, 2021, doi: 10.34818/JDSA.2021.4.71.