

Sentiment Classification of Hate Speech Against Islam on Twitter Platform Using Multinomial Naïve Bayes

^{1*}Zul Ifflah Al Juhaeda, ²Muhammad Faisal, ³Suhartono

^{1,2,3}Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia

*aljuhaeda@gmail.com

ABSTRACT

Twitter is widely used by public figures, politicians, celebrities, and organizations to communicate with the public. However, the platform's freedom of speech is often misused, leading to conflicts such as hate speech, especially against Islam. This study aims to develop a text classification system for detecting hate speech against Islam and to evaluate the performance of Multinomial Naïve Bayes (MNB) in this task. The data was obtained through Twitter data crawling and processed through several pre-processing steps: cleaning, case folding, tokenizing, stop words removal, and stemming. The processed data was then transformed using Bag of Words to compute word frequency, which was input into MNB. The first test compared the ratio of training to test data, adjusting the alpha hyperparameter within its minimum and maximum ranges. The second test involved k-fold cross-validation for model validation. The results showed the highest accuracy of 85% at a 90:10 training-to-test data ratio with the maximum alpha value. Using 10-fold cross-validation, the model achieved an average accuracy of 79.09%, with the highest accuracy of 85.05% in the 4th iteration. This study demonstrates that the training/test data ratio, alpha parameter, and cross-validation influence MNB's performance in classifying hate speech.

Keywords: Hate Speech, Sentiment Classification, Multinomial Naïve Bayes

1. INTRODUCTION

In the era of globalization, social media platforms like Twitter have experienced significant growth, bringing both positive and negative impacts to society. Twitter serves as a digital platform that offers various benefits in terms of communication and sharing online life. It enables the exchange of information through the sending and reading of short messages or tweets among users, encompassing text, images, links, videos, and GIFs [1]. With Twitter, users can access information anytime and anywhere on a wide range of topics, from global news to popular trends. During crises or major events, Twitter becomes a primary news source as users can share updates and information instantaneously. This platform has evolved into one of the most influential social media channels frequently used by public figures, politicians, celebrities, and organizations to communicate their messages directly to the public from various perspectives [2].

However, the rapid expansion and widespread use of Twitter also bring significant challenges. The excessive exercise of freedom of speech on social media leads to conflicts, notably hate speech. Hate speech is a severe conflict that undermines social harmony. It encompasses direct or indirect expressions of hatred in the form of provocation, incitement, or insults towards individuals or groups based on gender, sexual orientation, race, nationality, skin colour, and even religion [3]. Cases of hate speech against religion, particularly Islam, have become increasingly prevalent in society. These instances often manifest as insults towards Islam, desecration of religious symbols, defamation of religious figures, and other actions related to Islamic beliefs. Hate speech not only disrupts social harmony but also incites violence, leading to division and a loss of mutual respect among individuals [4].

Recent statistics highlight a troubling increase in hate speech incidents on Twitter. According to data from Robinopsnal Bareskrim Polri, the police handled 33 cases of hate speech from January to May 2022, with 10 cases reported in January alone [5]. Research conducted by Monash University and the Independent Journalists Alliance (AJI) Indonesia revealed that from September 1, 2023, to January 31, 2024, there were 182,118 hate speech posts on social media, with Twitter accounting for the highest number at 120,381 posts. Additionally, an analysis by Kumparan from the Supreme Court's decision directory recorded 60 cases of hate speech against religion on social media from 2012 to 2021, with the highest number of cases (20) occurring in 2020 [6]. Among the numerous recorded cases of hate speech against religion, Islam remains the most targeted in the past decade, with 51 cases accounting for 83.6% of total religious defamation cases [7].

Given the negative impact of spreading hate speech on social media, especially against Islam, effective preventive measures are necessary to create a safer and more harmonious environment. One of the preventive efforts to mitigate hate speech against Islam is the classification of text into hate speech (negative) and non-hate speech (positive) [8]. This classification aims to enable digital platforms to proactively detect circulating hate speech and prevent its dissemination. Additionally, it raises public awareness about the impact of hateful language, encouraging more cautious use of language that could harm or offend others [9].

Machine learning algorithms, specifically the Multinomial Naïve Bayes (MNB) classifier, have demonstrated efficacy in accurately categorizing text as hate speech (negative) or non-hate speech (positive) [10] [11]. The MNB algorithm is particularly suitable for text classification tasks due to its ability to handle large datasets and calculate word probabilities based on their frequencies, achieving relatively high accuracy [12]. Previous studies have shown that MNB, when combined with feature extraction techniques like Bag of Words, can effectively classify hate speech with substantial accuracy rates [13]. This study aims to evaluate the performance of the MNB algorithm in classifying hate speech against Islam on the Twitter platform using the Bag of Words feature extraction method. By assessing metrics such as accuracy, precision, recall, and F1-score, this research seeks to enhance the effectiveness of hate speech detection mechanisms, thereby contributing to the creation of a safer and more harmonious digital community [14].

2. RESEARCH METHOD

2.1 Data Collection

This study utilizes a dataset obtained from tweets on the Twitter platform through data crawling related to hate speech against Islam. The selected sentiment data consists of text only, excluding images, with a total of 1,190 data points collected between January 1, 2024, and July 31, 2024. The dataset labelling was performed manually through a questionnaire designed based on public assessment, resulting in outputs categorized as positive (containing non-hate speech elements against Islam) and negative (containing hate speech elements against Islam). The labelling process via the questionnaire ensures that the categorized data aligns with the actual context of hate speech [15]. The labelled dataset is then stored in a .csv file, serving as the research foundation for training and testing the classification model to detect hate speech against Islam on social media platforms, specifically Twitter. An example of the research data is shown in Table 1. follows.

Table 1. Sample Research Data

Id.	Tweet	Class
1	@Hilmi28 Negara2 Islam sdh seharusnya bersatu menyuarakan / melawan Zionis dan sekutunya...	Positif
2	@ProfOnline_id betul islam itu rohmatan lillalamin bukan radikal	Positif
3	@ProfOnline_id Itu hanya segilintir oknum ngak semua orang islam ...	Positif
4	@Sarcophagus12 @VKunyak Tidak ada adabnya dan tidak ada hukumnya. Islam membuat muslim jadi biadab. ...	Negatif
5	@sakkusaitna Kurang biadab apalagi ustad ini mendiskreditkan islam dengan persepsi liarnya sendiri	Negatif

2.2 System Design

The system design outlines the workflow from data input to generating output that fulfills the research objectives. The system architecture comprises four main stages: data input, preprocessing, Bag of Words (BoW) feature extraction, and classification using the Multinomial Naïve Bayes (MNB) algorithm. The design flow is illustrated in Figure 1.

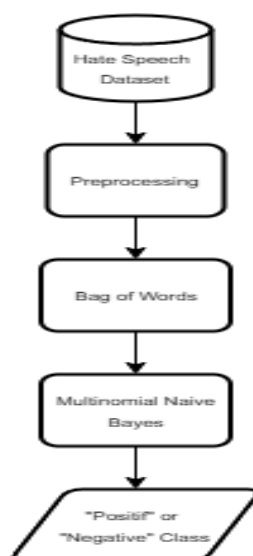


Figure 1. Design Flow

The first stage involves obtaining the dataset from tweets on the Twitter platform through data crawling related to hate speech against Islam. Following data collection, the pre-processing stage is conducted to clean and prepare the data for analysis. Pre-processing encompasses several steps: cleaning, case folding, tokenization, stop word removal, and stemming [16]. Cleaning removes irrelevant elements (noise) from the text, ensuring that each tweet contains only pertinent words. Case folding converts all characters to lowercase to maintain uniformity and avoid ambiguities caused by capitalization. Tokenization splits the cleaned and case-folded text into individual words or tokens based on punctuation marks and spaces. Stop word removal eliminates common words that do not carry significant meaning in the context of hate speech detection, thereby reducing noise in the data. Finally, stemming reduces words to their base or root form, minimizing variations and ensuring consistency in word representation [17].

After pre-processing, the Bag of Words (BoW) method is employed to transform the text data into numerical features by counting the frequency of each unique word in the dataset. This technique involves mapping each unique word in the corpus to an index and calculating the occurrence of each word in every document (tweet). In the BoW representation, each document is depicted as a vector in a high-dimensional space where each dimension corresponds to a unique word in the vocabulary [18]. The value in each dimension represents the frequency of the corresponding word in the document. This representation effectively captures the presence and frequency

of words, making it suitable for text classification tasks such as hate speech detection [19]. The BoW feature extraction process is illustrated in Figure 2.

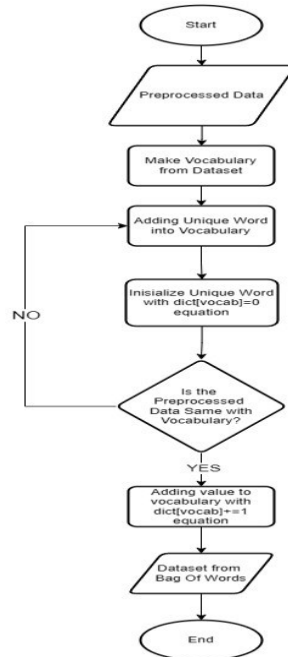


Figure 2. Bag of Words Feature Extraction

The final stage of the system design involves classification using the Multinomial Naïve Bayes (MNB) algorithm. The MNB classifier is chosen for its efficiency and effectiveness in handling large-scale text classification tasks [20]. The algorithm assumes that the features (words) are conditionally independent given the class label and follows a multinomial distribution to model word frequencies. The classification process begins with calculating prior probabilities for each class based on the training data. Subsequently, likelihood probabilities of words occurring within each class are determined. These probabilities are combined to compute the posterior probabilities, and the class with the highest posterior probability is assigned as the predicted class for each tweet [21]. Hyperparameter tuning, specifically adjusting the alpha parameter, is conducted to optimize the classifier's performance and improve accuracy in categorizing hate speech [22]. The Multinomial Naïve Bayes (MNB) algorithm process is illustrated in Figure 3.

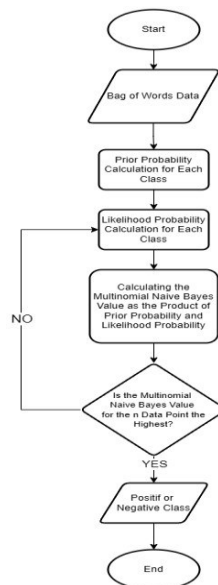


Figure 3. Multinomial Naïve Bayes

2.3 Experiment Design

The experimental setup aims to evaluate the performance of the MNB classifier in accurately classifying hate speech against Islam. The experiment design encompasses several key components to ensure a comprehensive assessment of the model's effectiveness. Data splitting involves dividing the dataset into training and testing sets using different split ratios (e.g., 70:30, 80:20, 90:10) to assess the impact of data distribution on model performance [23]. The MNB classifier is trained on the training dataset, learning the patterns and

word distributions associated with each class. Hyperparameter tuning focuses on adjusting the alpha parameter to identify the optimal value that maximizes classification accuracy. Additionally, k-fold cross-validation (e.g., 10-fold) is implemented to validate the model's performance and ensure its generalizability across different subsets of data [24]. The performance of the classification system is evaluated using four key metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's ability to detect hate speech accurately and reliably. The experiment design is summarized in Table 2.

Table 2. Experiment Design

Model	Training Data	Testing Data	Hyperparameter Tuning	Split Ratio	K-fold Cross Validation	Random Seed
Model A	833	357	Minimal	70:30	10 Fold	42
Model B	833	357	Maximal	70:30	10 Fold	42
Model C	952	238	Minimal	80:20	10 Fold	42
Model D	952	238	Maximal	80:20	10 Fold	42
Model E	1071	119	Minimal	90:10	10 Fold	42
Model F	1071	119	Maximal	90:10	10 Fold	42

The experiments aim to identify the best combination of parameters that yield the highest accuracy in hate speech classification, ensuring the robustness and reliability of the MNB classifier in real-world applications.

3. RESULT AND DISCUSSION

In this section, we present the results of the experiments conducted to evaluate the performance of the Multinomial Naïve Bayes (MNB) classifier in classifying hate speech against Islam on the Twitter platform. The evaluation encompasses various models (Model A to Model F) with different configurations of training-testing split ratios and hyperparameter settings. The performance is analyzed based on key metrics: accuracy, precision, recall, and F1-score.

3.1 Experiment Setup and Execution

In this section, we present the results of the experiments conducted to evaluate the performance of the Multinomial Naïve Bayes (MNB) classifier in classifying hate speech against Islam on the Twitter platform. The evaluation encompasses various models (Model A to Model F) with different configurations of training-testing split ratios and hyperparameter settings. The performance is analyzed based on key metrics: accuracy, precision, recall, and F1-score.

3.2 Model Performance Results

The classification experiments involved six different models (Model A to Model F) with varying training-testing split ratios and hyperparameter settings. Each model was evaluated based on its accuracy, precision, recall, and F1-score, as summarized in Table 3.

Table3. Performance Metrics of Classification Models

Model	Training Data	Testing Data	Hyperparameter Tuning	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Model A	833	357	Minimal (0.6)	83	80	83	80
Model B	833	357	Maximal (1.0)	84	81	84	80
Model C	952	238	Minimal (0.8)	83	79	83	79
Model D	952	238	Maximal (1.0)	84	81	84	80
Model E	1071	119	Minimal (0.4)	82	79	82	80
Model F	1071	119	Maximal (1.0)	85	82	85	81

Another illustration of the testing scenario using 10-Fold Cross Validation, which divides the dataset into ten subsets (folds) to ensure each subset is used for both training and testing, thereby enhancing the reliability of the evaluation. Is depicted in Figures 4.

K-Fold Cross Validation									
Iterasi	iterasi-1	testing	training	training	training	training	training	training	training
	iterasi-2	training	testing	training	training	training	training	training	training
	iterasi-3	training	training	testing	training	training	training	training	training
	iterasi-4	training	training	training	testing	training	training	training	training
	iterasi-5	training	training	training	training	testing	training	training	training
	iterasi-6	training	training	training	training	training	testing	training	training
	iterasi-7	training	training	training	training	training	training	testing	training
	iterasi-8	training	training	training	training	training	training	training	testing
	iterasi-9	training	training	training	training	training	training	training	testing
	iterasi-10	training	training	training	training	training	training	training	testing
		Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Fold-6	Fold-7	Fold-8
		Fold							

Figure 4. Testing Scenario Using 10-Fold Cross Validation

3.3 Detailed Model Performance

Model A utilized a 70:30 training-testing split with a minimal alpha value of 0.6. The confusion matrix (Figure 4.2) indicates that out of 357 test tweets, 48% of positive sentiments were correctly identified (precision), while 19% were correctly recalled. For negative sentiments, 86% were accurately classified with a recall rate of 96%. These results are summarized in Table 4.

Table 4. Calculation Results of Precision, Recall, and F1-Score of Model A

Class	Precision (%)	Recall (%)	F1-Score (%)
Positive	48	19	27
Negative	86	96	91

Model B also employed a 70:30 split but with a maximal alpha value of 1.0. The confusion matrix (Figure 4.3) reveals a precision of 56% and a recall of 16% for positive sentiments, while negative sentiments achieved an 86% precision and a 98% recall. These results are detailed in Table 5.

Table 5. Calculation Results of Precision, Recall, and F1-Score of Model B

Class	Precision (%)	Recall (%)	F1-Score (%)
Positive	56	16	24
Negative	86	98	91

Model C utilized an 80:20 split with an alpha of 0.8. The confusion matrix (Figure 4.4) shows that 45% of positive sentiments were precisely identified, with a recall of 13%. Negative sentiments exhibited an 85% precision and a 97% recall, as summarized in Table 6.

Table 6. Calculation Results of Precision, Recall, and F1-Score of Model C

Class	Precision (%)	Recall (%)	F1-Score (%)
Positive	45	13	20
Negative	85	97	91

Model D applied an 80:20 split with an alpha of 1.0. According to the confusion matrix (Figure 4.5), positive sentiments achieved a 62% precision and a 13% recall, while negative sentiments maintained an 85% precision and a 98% recall, as depicted in Table 7.

Table 7. Calculation Results of Precision, Recall, and F1-Score of Model D

Class	Precision (%)	Recall (%)	F1-Score (%)
Positive	62	13	21
Negative	85	98	91

Model E utilized a 90:10 split with a minimal alpha value of 0.4. The confusion matrix (Figure 4.6) indicates that 38% of positive sentiments were accurately classified, with a recall of 26%. Negative sentiments achieved an 87% precision and a 92% recall, as outlined in Table 8.

Table 8. Calculation Results of Precision, Recall, and F1-Score of Model E

Class	Precision (%)	Recall (%)	F1-Score (%)
Positive	38	26	31
Negative	87	92	89

Model F employed a 90:10 split with a maximal alpha of 1.0. The confusion matrix (Figure 4.7) demonstrates a 60% precision and a 16% recall for positive sentiments, while negative sentiments maintained an 86% precision and a 98% recall, as detailed in Table 9.

Table 9. Calculation Results of Precision, Recall, and F1-Score of Model F

Class	Precision (%)	Recall (%)	F1-Score (%)
Positive	60	16	25
Negative	86	98	92

Table 10 presents the results of the 10-fold cross-validation, with an average accuracy of 79.09% and the highest accuracy observed e model's performance consistent in the fourth iteration at 85.05%. The variability across folds highlights and areas for potential improvement.

Table 10. 10-Fold Cross Validation Results

Iteration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	77.78	51.84	51.11	50.76
2	75.70	49.71	49.77	49.60
3	75.70	54.54	54.54	54.54
4	85.05	71.78	69.64	70.60
5	78.50	55.13	53.82	54.13
6	80.37	60.80	57.18	57.42
7	79.44	59.30	56.62	57.36
8	78.38	53.25	53.25	53.25
9	84.11	71.02	68.36	66.10
10	80.37	62.32	59.39	60.39
Average	79.09	58.98	56.48	57.48

In The experimental results indicate that the Multinomial Naïve Bayes (MNB) classifier demonstrates a commendable ability to classify hate speech against Islam on Twitter, albeit with varying degrees of effectiveness across different models and configurations. Among the six models tested, Model F (90:10 split, alpha = 1.0) achieved the highest accuracy of 85%, along with a precision of 82%, recall of 85%, and F1-score of 81%. This suggests that allocating a larger portion of data to training (90%) enhances the model's ability to learn and recognize patterns associated with hate speech, thereby improving classification performance. Additionally, the maximal alpha value of 1.0 in smoothing appears to contribute positively by mitigating the impact of rare words and reducing overfitting, leading to more stable predictions.

The comparison between minimal and maximal alpha settings across different split ratios consistently shows that the maximal alpha (1.0) yields better performance metrics. This trend underscores the importance of hyperparameter tuning in optimizing the classifier's performance. The smoothing parameter (alpha) plays a crucial role in handling zero-frequency problems and ensuring that the probability estimates are more robust, especially in datasets with imbalanced classes. Furthermore, the training-testing split ratio significantly influences the model's performance. Models with a 90:10 split generally outperformed those with 70:30 and 80:20 splits. A higher proportion of training data allows the model to better capture the underlying patterns and nuances in the data, thereby enhancing its predictive capabilities. However, extremely imbalanced splits (e.g., 90:10) may lead to insufficient testing data, potentially affecting the generalizability of the results.

The 10-fold cross-validation results, with an average accuracy of 79.09%, provide a more comprehensive evaluation of the model's performance. The variability across folds indicates that while the model performs well on certain data subsets, there is room for improvement in achieving consistent performance across all subsets. The high accuracy in specific iterations (e.g., Iteration 4 and 9) suggests that the model is capable of leveraging patterns in the data effectively, but it may also be susceptible to fluctuations based on the distribution of data in each fold. While the overall accuracy is high, the precision and recall metrics reveal a trade-off between correctly identifying positive instances (hate speech) and minimizing false positives. For instance, Model F achieves a high recall of 85%, indicating a strong ability to identify actual hate speech instances. However, maintaining high precision is equally important to avoid misclassifying non-hate speech as hate speech, which can have significant implications for user experience and platform moderation policies.

Overall, the findings validate the efficacy of the Multinomial Naïve Bayes classifier, particularly when optimized with appropriate parameter settings and sufficient training data. Future research could explore the integration of more sophisticated feature extraction techniques, such as TF-IDF or word embeddings, and the application of hybrid models combining MNB with other machine learning algorithms to further enhance classification accuracy and robustness.

4. CONCLUSION

This study successfully implemented the Multinomial Naïve Bayes (MNB) classifier to detect hate speech against Islam on the Twitter platform, utilizing a dataset of 1,190 tweets categorized into positive and negative sentiments. By experimenting with different

training-testing split ratios of 70:30, 80:20, and 90:10, along with hyperparameter tuning of the alpha parameter within minimal and maximal ranges, the classifier achieved its highest accuracy of 85% using a 90:10 split and an alpha value of 1.0. Additionally, the application of 10-fold cross-validation yielded an average accuracy of 79.09%, affirming the model's reliability and generalizability across diverse data subsets. These results underscore the significance of optimal data partitioning and hyperparameter settings in enhancing the performance of sentiment classification models. For future research, it is recommended to explore advanced feature extraction techniques such as TF-IDF or word embeddings, experiment with alternative classification algorithms like Support Vector Machines (SVM) and ensemble methods, and address class imbalance through techniques like SMOTE. These steps can further improve the accuracy and robustness of hate speech detection systems, contributing to more effective moderation and fostering a safer online environment.

REFERENCES

- [1] B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela, "Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection," Dec. 2020, [Online]. Available: <http://arxiv.org/abs/2012.15761>
- [2] A. Afdhaluzzikri, H. Mawengkang, and O. S. Sitompul, "Performance analysis of Naive Bayes method with data weighting," *Sinkron*, vol. 7, no. 3, pp. 817–821, Jul. 2022, doi: 10.33395/sinkron.v7i3.11516.
- [3] D. C. Asogwa, C. I. Chukwuneke, C. C. Ngene, and G. N. Anigbogu, "Hate Speech Classification Using SVM and Naive BAYES," Mar. 2022, doi: 10.9790/0050-09012734.
- [4] Kementerian Agama RI, *Tafsir Wajiz Jilid I*. Jakarta, 2016.
- [5] Kepolisian Negara Republik Indonesia, "SURAT EDARAN KAPOLRI MENGENAI PENANGANAN UJARAN KEBENCIAN," 2015.
- [6] Bitsanie Zujajmahir Ramanda, "Makna Surah An-Nisa' Ayat 148 Dalam Al-Qur'an Perspektif Tafsir Al-Azhar Karya Buya Haji Abdul Malik Karim Amrullah," *Al-Mizan: Journal of Islamic Studies*, vol. 1, no. Nomor 1, pp. 15–16, 2024, [Online]. Available: <https://www.ejournal.stiq-kepri.ac.id/index.php/Al-Mizan/index>
- [7] N. A. Saputra, K. Aeni, and N. M. Saraswati, "Indonesian Hate Speech Text Classification Using Improved K-Nearest Neighbor with TF-IDF-ICSpF," *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 21–30, Feb. 2024, doi: 10.15294/sji.v11i1.48085.
- [8] M. M. Munir, M. A. Fauzi, and R. S. Perdana, "Implementasi Metode Backpropagation Neural Network berbasis Lexicon Based Features dan Bag of Words Untuk Identifikasi Ujaran Kebencian Pada Twitter," 2018. [Online]. Available: <http://jptiik.ub.ac.id>
- [9] A. Ilham, N. Azmi Verdikha, and A. J. Latipah, "Jurnal Explore IT|64 Klasifikasi Ujaran Kebencian di Twitter Menggunakan Fitur Ekstraksi Glove dengan Support Vector Machine (SVM)," 2023, doi: 10.35891/explorit.
- [10] M. M. Munir, M. A. Fauzi, and R. S. Perdana, "Implementasi Metode Backpropagation Neural Network berbasis Lexicon Based Features dan Bag of Words Untuk Identifikasi Ujaran Kebencian Pada Twitter," 2018. [Online]. Available: <http://jptiik.ub.ac.id>.
- [11] A. Sidqi and Suhartono, Bibliometric Analysis Using Publish or Perish and VOS Viewer: Machine Learning in Relation to Cardiovascular Disease, *Journal of Informatics and Science Media*, vol. 1, no. 1, pp. 34–38, Jul. 2024, [Online]. Available: <https://jismandiri.com/index.php/jisman/article/view/11/8>
- [12] A. Rahman and A. Doewes, "Online News Classification Using Multinomial Naive Bayes." [Online]. Available: www.kompas.com
- [13] L. Efrizoni, S. Defit, M. Tajuddin, and A. Anggrawan, "Komparasi Ekstraksi Fitur dalam Klasifikasi Teks Multilabel Menggunakan Algoritma Machine Learning," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 653–666, Jul. 2022, doi: 10.30812/matrik.v21i3.1851.
- [14] T. T. A. Putri, S. Sriadhi, R. D. Sari, R. Rahmadani, and H. D. Hutahaean, "A comparison of classification algorithms for hate speech detection," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, May 2020. doi: 10.1088/1757-899X/830/3/032006.
- [15] I. And and D. Expert, "Deteksi Ujaran Kebencian dengan Metode Klasifikasi Naïve Bayes dan Metode N-Gram pada Dataset Multi-Label Twitter Berbahasa Indonesia INFORMASI ARTIKEL ABSTRAK," 2022. [Online]. Available: <http://index.unper.ac.id>
- [16] C. P. Chai, "Comparison of text preprocessing methods," *Nat Lang Eng*, vol. 29, no. 3, pp. 509–553, May 2023, doi: 10.1017/S1351324922000213.
- [17] M. Anandarajan, C. Hill, and T. Nolan, "Text Preprocessing," 2019, pp. 45–59. doi: 10.1007/978-3-319-95663-3_4.
- [18] D. Qiu, H. Jiang, and S. Chen, "Fuzzy information retrieval based on continuous bag-of-words model," *Symmetry (Basel)*, vol. 12, no. 2, Feb. 2020, doi: 10.3390/sym12020225.
- [19] M. M. Munir, M. A. Fauzi, and R. S. Perdana, "Implementasi Metode Backpropagation Neural Network berbasis Lexicon Based Features dan Bag of Words Untuk Identifikasi Ujaran Kebencian Pada Twitter," 2018. [Online]. Available: <http://jptiik.ub.ac.id>
- [19] B. Harjito, K. N. Aini, A. Wijayanto, and B. Murtiyasa, "Comparison of Multinomial Naïve Bayes with K-Nearest Neighbors, Support Vector Machine and Random Forest for Classification of 'Network Attacks' Document," 2020. [Online]. Available: www.sciencedirect.com
- [20] R. M. Yazid, F. R. Umbara, and P. N. Sabrina, "Deteksi Ujaran Kebencian dengan Metode Klasifikasi Naïve Bayes dan Metode N-Gram pada Dataset Multi-Label Twitter Berbahasa Indonesia," 2022. [Online]. Available: <http://index.unper.ac.id>

- [21] N. Yusliani, S. A. Q. Aruda, M. D. Marieska, D. M. Saputra, and A. Abdiansah, "The effect of Chi-Square Feature Selection on Question Classification using Multinomial Naïve Bayes," *Sinkron*, vol. 7, no. 4, pp. 2430–2436, Oct. 2022, doi: 10.33395/sinkron.v7i4.11788.
- [22] A. Sabrani, I. W. Gede Putu Wirarama Wedashwara, and F. Bimantoro, "METODE MULTINOMIAL NAÏVE BAYES UNTUK KLASIFIKASI ARTIKEL ONLINE TENTANG GEMPA DI INDONESIA (Multinomial Naïve Bayes Method for Classification of Online Article About Earthquake in Indonesia)." [Online]. Available: <http://jtika.if.unram.ac.id/index.php/JTIKA/>
- [23] L. Yance Nanlohy, E. Mulyanto Yuniarno, and S. Mardi Susiki Nugroho, "Classification of public complaint data in sms complaint using naive bayes multinomial method," in *4th International Conference on Vocational Education and Training, ICOVET 2020*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 241–246. doi: 10.1109/ICOVET50258.2020.9229941.
- [24] Yuyun, Nurul Hidayah, and Supriadi Sahibu, "Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 820–826, Aug. 2021, doi: 10.29207/resti.v5i4.3146.