**Hate Speech Detection on Indonesian Tweet Using Multinomial Naïve Bayes**

**First Author1, Second Author2, Third Author3 (10 pt)**

1,3National Chung Cheng University, Taiwan (9 pt)

2Universitas Bumigora, Indonesia (9 pt)

|  |  |  |
| --- | --- | --- |
| **Article Info** |  | **ABSTRACT** (10 PT) |
| ***Article history:***  Received mm dd, yyyy  Revised mm dd, yyyy  Accepted mm dd, yyyy |  | Detecting hate speech on social media is a significant challenge, particularly considering the increasing volume of data. Currently, the task of monitoring the spread of hate speech is officially carried out by the Ministry of Communication and Information (Kominfo) and the National Cyber and Encryption Agency (BSSN). However, it must be acknowledged that the dissemination of hate speech continues to rise over time. In order to identify content containing hate speech on social media, the Indonesian National Police (Polri) has formed a team known as the "Cyber Patrol." On the other hand, combating the spread of hate speech requires active participation from all segments of society. This research aims to develop a system capable of detecting sentences containing hate speech using the Multinomial Naive Bayes algorithm and the TF-IDF feature extraction technique. A total of 300 data from the Twitter platform were used as training data, while 30 data were used as testing data. The research results showed an accuracy rate of 76% with a 90%:10% data split between training and testing. These testing results illustrate that the developed system can identify tweets containing hate speech using the Multinomial Naïve Bayes method |
| ***Keywords:***  Text-Mining  Hate Speech  Multinomial Naïve Bayes  TF-IDF |
| *This is an open access article under the* [*CC BY-SA*](https://creativecommons.org/licenses/by-sa/4.0/) *license.* |
| ***Corresponding Author:***  Name of Corresponding Author,  Department of Engineering And Design,  Universitas Bumigora.  Email: corresp-author@universitasbumigora.ac.id | | |

1. **INTRODUCTION (10 PT)**

The failure of Indonesia being host the U-20 World Cup has caused disappointment among some people. Many people blame politically charged narratives as the main cause of this failure. The escalation of issues on social media further increased on March 29, 2023, after FIFA officially withdrew Indonesia's status as the host of the U-20 World Cup. Since then, there has been an increase in confirmed cases of hate speech over time.

Hate speech can be delivered both orally and in writing. If delivered orally, it occurs directly, face-to-face with the interlocutor. Meanwhile, when someone conveys hate speech in writing, they express their feelings and thoughts through written media, such as books or social media [1] . Since 2018, the Indonesian Ministry of Communication and Information Technology (Kominfo) has handled 3,640 pieces of content related to hate speech based on ethnicity, religion, race, intergroup relations (SARA), and politics. An example of a hate speech case in Indonesia was the case involving the artist Ahmad Dhani. He was considered to have spread hatred towards a particular group through his Twitter account. His tweet contained the following statement: "Anyone who supports the blasphemer of religion is a scoundrel who deserves to be spat on their face."

There have been other studies conducted previously on hate speech classification. The research focused on Sentiment Analysis of Hate Speech on Twitter users using the Naïve Bayes classifier (NBC) method with 5000 data implemented using Python. That study achieved the highest accuracy of 80% with data split of 70% training and 30% testing[2]. willianto et al. [3] using data sourced from Facebook, consisting of 500 data with a data split of 80% training and 20% testing, Obtained an accuracy score of 83%. Ihsan et al. [4] added abusive word classes to their research, using a data split proportion of 90%:10% and the Decision Tree Algorithm, achieving an accuracy score of 70.48%. (Hakiem et al.) [5] using Multinomial Naïve Bayes on N-Gram with Feature Selection of Information Gain The accuracy rate achieved is 84% with a total of 250 data.

In this current research, a system has been devised to identify instances of hate speech in Indonesian tweets about PSSI (Indonesian Football Association) and political matters. Three classes will be used in this research non hs (non-hate speech), penghinaan (contempt), and provokasi (provocation). The method used involves web scraping for data collection.The data split is utilized with a ratio of 90% : 10% for data training and testing and Multinomial Naïve Bayes combined with Term-Frequency-Inverse Document Frequency (TF-IDF) for algorithm and feature extraction.

1. **RESEARCH METHOD (10 PT)**

Figure 1 illustrates the processes to be conducted in this research. The first step involves data collection of tweets through crawling, which will become the dataset. The gathered data will be split into two distinct sections: the training data and the testing data. The training data contain five steps, The first step is manual labeling consisting of three classes, namely " non hs (non-hate speech), penghinaan (contempt), and provokes (provocation),The second step is preprocessing to obtain clean data, and the third step is inputting the data into the database, The fourth step is creating a word corpus, and the last step is feature extraction. The testing data contains two steps,The first is preprocessing to obtain cleaned data. In the subsequent stage, the multinomial Naïve Bayes algorithm is applyed to attain classification results.

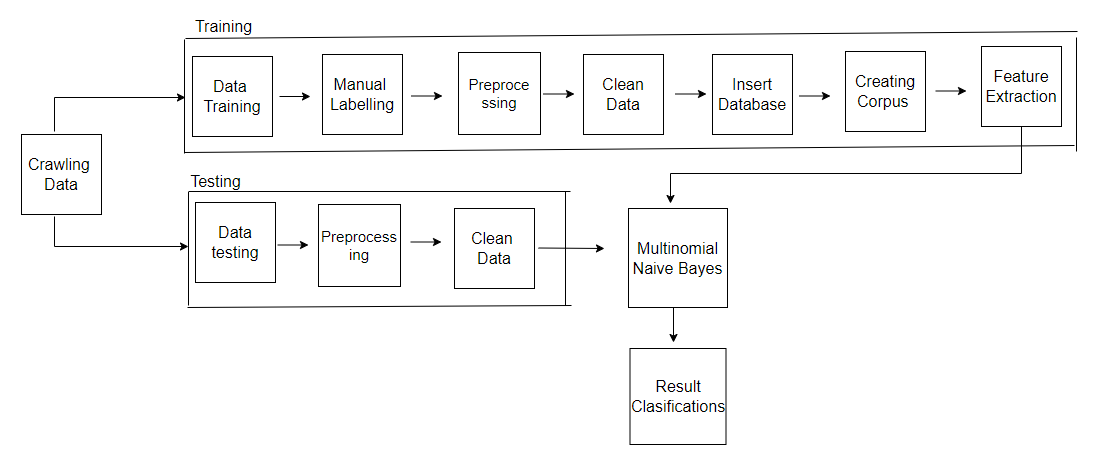


Figure 1 The program's overall process flow.

**2.1. Data Collection**

This research is conducted through several stages. The study starts with data collection, which involves using a website that can scrape data from Twitter called Apify and Netlytic. Specific keywords such as PSSI (persantuan sepakbola seluruh indonesia) and politics are used in the scraping process. The data acquired through the scraping procedure will be stored in an Excel file format.

**2.2 Labeling Data**

After collecting the data in the data collection phase, subsequently, the data will be segregated into two categories: training data and testing data.. The training data is used to train the model or algorithm that will be used in the data analysis. Conversely, the testing data assesses the model's performance on data it has yet to encounter previously. After the data division is done, the training data will undergo a manual labeling process. The labeling in this research involves three classes: non-hs (non-hate speech), penghinaan (contempt), and provokasi (provocation). The labeling is done manually by analyzing the meaning of each tweet. In analyzing the tweet data in this research, a linguistic meaning approach is used, using the concept of conceptual meaning. This conceptual meaning approach refers to the meaning of words or sentences based on grammatical meaning without considering the context[6]

**2.2 Preprocessing**

Preprocessing is the initial stage of selecting and transforming data structurely[7]. The preprocessing stage is performed to prepare the data so that the text data becomes more structured by removing noise present in the data, thus facilitating the classification process[8]. In this research, the preprocessing stage includes several steps, namely case folding, data cleaning, slang word handling, stemming, and stopword removal.

a) Case Folding

Case Folding is a step to convert all uppercase letters to lowercase letters. This step is performed to ensure uniformity in the text and make it easier to process

b) Data Cleaning

Cleansing the data by removing extraneous elements such as emoticons, punctuation, and numerical values in order to streamline and reduce unnecessary information [9].

c) Slang word

The Slang Word stage aims to convert non-standard words into standard words that are ready for processing. Non-standard words may include abbreviations or slang terms. Examples of non-standard words commonly seen on social media are “begajulan," which means “nakal," and “bumil," which means”ibu hamil." In this stage, these non-standard words will be converted into standard words for better processing.

d) Stemming

Stemming is a word-processing process that aims to obtain the base form of a word after removing affixes according to certain rules. This process is done because words with prefixes, suffixes, or infixes can complicate the matching of related words[10]

e) Stopword

The Stopword stage involves extracting important words and removing words that do not carry meaning in the classification process. The words in question include common words such as "yang," "di," "dan," "diri," and so on

**2.3 Feature Extraction**

TF-IDF stands as one of the methods for weighting words. It combines Term Frequency (TF) and Inverse Document Frequency (IDF). Term Frequency, denoted as TF, quantifies the number of times a term appears within a document. Certain terms could occur more frequently in longer documents compared to shorter ones. On the other hand, IDF computes the occurrence of terms across the entire document collection. IDF assigns higher values to terms that appear less frequently, while more common terms receive lower values [11].TF-IDF is a method that combines two-word weighting approaches, namely by calculating word frequency and applying the inverse of the number of documents containing that word[12]. The TF-IDF calculation can be performed using equations (1), (2), and (3).

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |
|  | (3) |

Where :

(𝑑,𝑡) : Frequency of the term t appearing in document d.

𝐼𝐷𝐹(𝑡) : Inverse document frequency in term t

𝑁𝑑 : Total Number of documents

𝑑(𝑡) : Number of documents containing term t.

n(d,t) : total number of terms in a document d

TF(t) : Term frequency in term t

**2.4 Multinomial Naïve Bayes**

Multinomial Naïve Bayes Classification is a form of Bayes algorithm commonly used in text classification. Multinomial Naïve Bayes documents are considered as "bag of words," where the sequence of word occurrences in the paper is disregarded, and each word is processed using a multinomial distribution[13]. The Multinomial Naïve Bayes classification process encompasses a tripartite sequence. Initially, probabilities are computed for each class. Subsequently, probabilities for each word within every class are computed. Lastly, these calculations are juxtaposed to ascertain the anticipated class. The calculation of Multinomial Naïve Bayes can be seen in equation (4).

|  |  |
| --- | --- |
| (𝑐, 𝑑) = 𝑥 𝑃(𝑡1, 𝑐) 𝑥 … 𝑥 𝑃(𝑡𝑛, 𝑐) | (4) |

Where :

(𝑐, 𝑑) : Probability of document d belonging to class c

𝑁 : Count of class c document

: Total number of documents

𝑡 : The n word in document d

(𝑡𝑛, 𝑐): Probability of word tn given class c

The formula for calculating the probability of the n word using TF-IDF word weighting can be seen in the equation (5) [12]

|  |  |
| --- | --- |
| 𝑃 (𝑡𝑛, 𝑐) = | (5) |

Where:

𝑊𝑐 : The TF-IDF weighting value or W of term t in category c

∑𝑊′ : Total sum of W values from all terms in category c

𝐵 : Total sum of W values for unique terms (IDF value not multiplied by tf) across all documents

**2.5 Confusion Matrix**

Confusion Matrix determines how effectively a model classifies data[14].

There exist four terms signifying the results of the classification procedure: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The True Negative (TN) value signifies the number of negative data instances correctly identified, whereas False Positive (FP) represents the negative data instances incorrectly identified as positive [15]. Applying the Confusion Matrix makes it possible to calculate Accuracy, Precision, and Recall values. The calculations for Accuracy, Precision, and Recall can be seen in equations (6), (7), and (8).

|  |  |
| --- | --- |
| 𝐴𝑐𝑐𝑢𝑟𝑎𝑐𝑦 = | (6) |
| 𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 = | (7) |
| 𝑅𝑒𝑐𝑎𝑙𝑙 = | (8) |

1. **RESULTS AND ANALYSIS (10 PT)**

**3.1. Dataset**

This research utilizes a Dataset sourced from the Twitter social media platform. The data was obtained through a crawling process facilitated by the websites "Apify" and "Netlic." The total amount of successfully collected data is 300 tweets, all generated between March 29, 2023, and June 30, 2023. An instance of the collected data is shown in Table 1

Table 1 Example of Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| Created At | Username | Tweets | Tweets (Bahasa) |
| 28-05-2023 | @Mira\_SasMiita | Good news: Indonesia has officially been selected by the Asian Football Confederation (AFC) as one of the 11 host countries for the qualifying round of the 2024 AFC U-23 Asian Cup. This achievement is attributed to the role of the Indonesian Football Association (PSSI) Chairman @erickthohir, who has international relations. | Good news, Indonesia resmi didapuk oleh Konfederasi Sepak Bola Asia (AFC) menjadi salah satu dari 11 negara tuan rumah babak kualifikasi Piala Asia U-23 2024. Hal ini tak terlepas dari peran Ketum PSSI @erickthohir yang memiliki relasi internasional |

**3.2. Labeling**

The next stage is data labeling. In the labeling phase, the acquired data will be classified according to the predefined categories in this study. This research applies three classes: non-hs (nonhate speech), penghinaan (contempt), and provokasi (provocation). The labeling is conducted manually by analyzing the meaning of each tweet. In the analysis of data tweets in this study, a linguistic meaning approach is utilized, applying conceptual concepts. The labeling process is shown in Table 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Created At | Username | Tweets | Tweets (Bahasa) | Label |
| 28-05-2023 | @Mira\_SasMiita | Good news: Indonesia has officially been selected by the Asian Football Confederation (AFC) as one of the 11 host countries for the qualifying round of the 2024 AFC U-23 Asian Cup. This achievement is attributed to the role of the Indonesian Football Association (PSSI) Chairman @erickthohir, who has international relations. | Good news, Indonesia resmi didapuk oleh Konfederasi Sepak Bola Asia (AFC) menjadi salah satu dari 11 negara tuan rumah babak kualifikasi Piala Asia U-23 2024. Hal ini tak terlepas dari peran Ketum PSSI @erickthohir yang memiliki relasi internasional | Non Hs (non hate speech) |

Table 12 presents an example tweet labeled as non hs (non hate speech), as the respective tweet content does not fall within the categories of penghinaan (contempt) or provokasi (provocation). In the context of tweets categorized as defamation, they encompass content that evokes emotions and belittles the dignity of individuals or institutions. Conversely, tweets classified as provocation aim to foment hatred or enmity towards specific individuals and/or societal groups

**3.3. Sub section 3**

xx

3.3.1. Subsub section 3

yy

3.3.2. Subsub section 2

yy

1. **CONCLUSION (10 PT)**

Provide a statement that what is expected, as stated in the "Introduction" chapter can ultimately result in "Results and Analysis" chapter, so there is compatibility. Moreover, it can also be added the prospect of the development of research results and application prospects of further studies into the next (based on result and discussion).

**ACKNOWLEDGEMENTS (10 PT)**

The Acknowledgments section is optional. Research sources can be included in this section.

**REFERENCES (10 PT)**

The main references are international journals and proceedings. All references should be to the most pertinent, up-to-date sources **and the minimum of references are 15**. References are written in IEEE style. Please use a consistent format for references – see examples below (9 pt):

1. X. S. Li*, et al.*, "Analysis and Simplification of Three-Dimensional Space Vector PWM for Three-Phase Four-Leg Inverters," *IEEE Transactions on Industrial Electronics,* vol. 58, pp. 450-464, Feb 2011.
2. R. Arulmozhiyal and K. Baskaran, "Implementation of a Fuzzy PI Controller for Speed Control of Induction Motors Using FPGA," *Journal of Power Electronics,* vol. 10, pp. 65-71, 2010.
3. D. Zhang*, et al.*, "Common Mode Circulating Current Control of Interleaved Three-Phase Two-Level Voltage-Source Converters with Discontinuous Space-Vector Modulation," *2009 IEEE Energy Conversion Congress and Exposition, Vols 1-6,* pp. 3906-3912, 2009.
4. Z. Yinhai*, et al.*, "A Novel SVPWM Modulation Scheme," in *Applied Power Electronics Conference and Exposition, 2009. APEC 2009. Twenty-Fourth Annual IEEE*, pp. 128-131, 2009.