HATE SPEECH DETECTION ON INDONESIA LANGUAGE TWITTER USING THE NAÏVE BAYES ALGORITHM

Compiled as one of the requirements to complete the Strata I Study Program in the Informatics Study Program, Faculty of Communication and Informatics

By:

NADIA ZERLINDA TAUFIK PUTRI

L200194009

INFORMATION ENGINEERING STUDY
PROGRAM FACULTY OF COMMUNICATION
AND INFORMATIONMUHAMMADIYAH
SURAKARTA UNIVERSITY

2022

APPROVAL PAGE

HATE SPEECH DETECTION ON INDONESIA LANGUAGE TWITTER USING THE NAÏVE BAYES ALGORITHM

by:

NADIA ZERLINDA TAUFIK PUTRI
L200194009

Checked and approved for
testing by: Dosen Pembimbing

Endang Wahyu Pamungkas, S.Kom., M.Kom., PhD

NIK.1704

HALAMAN PENGESAHAN

HATE SPEECH DETECTION ON INDONESIA LANGUAGE TWITTERUSING THE NAÏVE BAYES ALGORITHM

BY

NADIA ZERLINDA TAUFIK PUTRI

L200194009

Telah dipertahankan di depan Dewan Penguji
Fakultas Komunikasi dan Informatika
Universitas Muhammadiyah Surakarta
Pada hari Senin, 13 Januari 2022
dan dinyatakan telah memenuhi syarat

Dewan Penguji:s

1. **Endang Wahyu Pamungkas, S.Kom., M.Kom., Ph.D** (.....)
   (Ketua Dewan Penguji)

2. **Fajar Suryawan, S.T., M.Eng.Sc., Ph.D** (.....)
   (Anggota I Dewan Penguji)

3. **Nurgiyatna, S.T., M.Sc., Ph.D** (.....)
   (Anggota II Dewan Penguji)

Dekan
Fakultas Komunikasi dan Informatika

Nurgiyatna, S.T., M.Sc., Ph.D
NIK. 881

Ketua Prodi Informatika

Dedi Gunawan, S.T., M.Sc, Ph.D
NIK. 1305

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam naskah publikasi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan orang lain, kecuali secara tertulis diacu dalam naskah dan disebutkan dalam daftar pustaka.

Apabila kelak terbukti ada ketidakbenaran dalam pernyataan saya di atas, maka akan saya pertanggungjawabkan sepenuhnya.

Surakarta, 10 November 2022

Penulis

NADIA ZERLINDA TAUFIK PUTRI

L200194009

HATE SPEECH DETECTION ON INDONESIA LANGUAGE TWITTERUSING THE NAÏVE BAYES ALGORITHM

Abstract

Social media is a medium for disseminating information online and is very easy to do because there are no rules for writing in online media. Therefore, anyone with access to social media can disseminate information without prior filtering, which can lead to hate speech. Hate speech is any form of communication in the form of harsh words that provoke both groups and individuals. However, these harsh words often cause hate speech to be banned in public spaces such as social media. The social media used in this study is Twitter, because data can be retrieved through the Twitter library. This study aims to create a tool to detect hate speech on Instagram comments in Indonesian. This study uses the Naive Bayes algorithm method, which has classification types such as Gaussian, Bernoulli, and Multinomial. This is used to calculate the probability value of words in text documents that are included in hate speech or non-hate speech with its level of accuracy. This classification is made with the Python programming language. At the Classification data collection stage, there were 4393 data collected randomly on social media Twitter. The data was divided using a ratio of 80% - 20%, where 3514 text data were used as training data and 879 data were used as test data. This research resulted in an accuracy level of classification of hate speech texts of 80% resulting from the Multinomial Naive Bayes classification with a pre-processing process.

Keywords: hate speech, naïve bayes, social media, twitter.

1. INTRODUCTION

Indonesia is one of the countries with the most social network users, where social networks are used for various purposes, such as finding and sharing information, establishing communication, and much more (Luthfi & Riasti, 2013). The use of social networks on a large scale is often found in uncontrolled communication and leads to abusive language.

Twitter is one of the social networking media that is quite popular in Indonesia and has been used byvarious agencies, organizations, and individuals. People can communicate through tweets posted onTwitter, these tweets can have two types, some are positive and some are negative (Dwitama & Hidayat,2021). Not a few people who are members of Twitter write words that are SARA (Ethnicity, Religion,Race, and Intergroup) or even write harsh words. Harsh words in Indonesian are usually written to express feelings of disappointment, attack certain parties, or express emotions about certain events.

Currently, active users of social media are mostly teenagers, they are used to commenting, sharing, and giving criticism on social media (Rahmadhany et al., 2021) The uncontrolled use of harsh wordson social networks is due to the absence of an effective tool for filtering abusive language on social networks, too free coverage on social media, lack of awareness from the public, and possibly a lack of parental supervision. The need for filtering harsh words on social networks so that no children andteenagers learn abusive language from the social networks they use. However, it is very difficult to filter out profanity on social

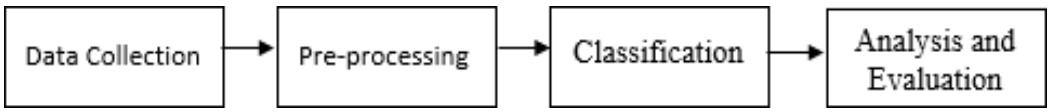
networks manually because it is very easy to find abusive language writing.

Several studies have been carried out to solve speech classification cases using various algorithms. The classification methods used are quite popular including Decision Tree, Naive Bayes Classifier (NBC), and K-nearest Neighbor (k-NN) (Alwiah Musdar et al., 2021). (Athira et al., 2018) conductedan analysis of cyberbullying sentiment on Instagram comments using the Support Vector Machine (SVM) classification method and obtained the best accuracy of 90%. (N. T. Romadloni & S. Budilaksono, 2019) compared the NBC, k-NN, and Decision Tree methods to complete the sentimentanalysis of commuter line KRL transportation. (Fauzi et al., 2018) determine public sentiment towards AMIK BSI Tegal which is conveyed on Instagram with the classification of the Naïve Bayes Algorithm, thus helping to conduct marketing research on public opinion. In this study, the text of the comment will detect whether it contains hate speech or does not contain hate speech. Thetext document used is an Indonesian language document obtained from Twitter social media, using the Naive Bayes method.

The purpose of the research "Hate Speech Detection On Indonesian Language Twitter Using The Naïve Bayes Algorithm" is to implement the Naïve Bayes method to classify hate speech in Twitter comments. As well as measuring the performance of the Naïve Bayes method in classifying hate speech text on Twitter social media. This research consists of several sections that explain the introduction, background, and research objectives. The second part explains about containing journals related to research work that has beendone previously and is used as reference material by the author in conducting research. The third section describes the method which contains an explanation of the system and how the system works.The fourth section describes the discussion of the results obtained after following the methods and systems that have been determined. The last section describes the closing, which includes the expected results and whether they are appropriate and have met the goals desired by the author.

2. METHOD

In this study, the hate speech classification system in Indonesia on Twitter uses Naive Bayes with a workflow that can be seen in Picture 1.



Picture 1. Method flow

2.1. Data collection

Data collection is a desirable target in manufacturing by means of various types of sensors (Xu et al., 2020). Tests on the implementation use data collection from social media such as Twitter. This data consists of 4393 text data, from previous research can be seen in "https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection". And also can be seen in "https://github.com/NakulLakhotia/Hate-Speech-Detection-in-Social-Media-using-Python" where all text data is collected and stored in csv file format. There are 1735 data labeled HS, and 2658 data labeled Non_HS.

```
Non_HS    2658
HS        1735
Name: Label, dtype: int64
```

Picture 2. Amount of Data on The label Non_HS and HS

2.2. Pre-processing

The pre-processing process is a stage for the process of eliminating several problems that can interfere with data processing because a lot of data has an inconsistent format. In this pre-processing process, the were carried out with the following:

2.2.1. Case Folding

Is to generalize capital letters or case folding, and remove punctuation, so the system can clean the data. The data is free from double spaces, mentions, links, and punctuation which are less informative.

2.2.2. Tokenization

The text data generated from the pre-processing process in the form of sentences will be broken down into words (Sari, Dedy, Wahyu, & Rathimala, 2021).

2.2.3. Filtering

This process uses a list of Indonesian stopwords. The process of removing less descriptive words that have no meaning at all in a document

2.2.4. Stemming

This process is a very important process for finding the root word of a derivative word, serves to cut the word affixes from each word and make it a basic word.

In retrieving data on text labels there are examples as in table 1.

Table 1. Example Tweet Before Pre-processing and After Pre-processing

No	Label	Before Pre-processing	After Pre-processing
1	Non_HS	Hamdalah. Kelar juga fitur keparat.’	hamdalah. Kelar juga fitur keparat
2	Non_HS	USER udah siap didemo berjilid2 sama bani cin…	User udah demo berjilid bani ingkarang…

2.3. TF-IDF Weighting

Term Frequency Inverse Document Frequency (TF-IDF) is an algorithm that calculates the weight of a value in each word in the document. The weight value is the value of the importance of a word in the document.

2.4. Classification

The classification used in this study uses Naïve Bayes, because it has good performance in carrying out the text classification process. The Naive Bayes classification is based on the Bayes theorem for conditional probabilities which is formulated as follows:

$P(A|B) = P(B|A)P(A)P(B)$

Description:

P(A|B) : The probability that A will occur given the evidence that B has occurred

P(B|A) : The probability that B will occur given the evidence that A has occurred

P(A) : Probability of occurrence of A

P(B) : Probability of occurrence of B

This research uses three Naïve Bayes algorithms namely Naïve Bayes Gaussian, Naïve Bayes Bernoulli, and Naïve Bayes Multinomial.

2.4.1 Gaussian Naive Bayes

The Model is a probabilistic classification algorithm based on Bayes' theorem that works by:

1. Calculate the previous probability of each class based on the frequency of that class in the training data.
2. Calculate the average and variance of each feature in each class.
3. Calculate the probability of each feature value for each class using the

Gaussian probability density function.

4. Multiply the probability for all the features in the observations in each class, then multiply the result by the previous probability of the same class.
5. Normalize the results by dividing each class probability by the sum of all class probabilities.
6. Finally, assign observations to the class with the highest posterior probability.

2.4.2 Bernoulli Naive Bayes

This model works with calculations the amount of data containing the word term, not the frequency of occurrence of the word (Ashari, 2020).

2.4.3 Multinomial Naïve Bayes

This model works by calculating the frequency of each word that appears in certain documents where the feature value has more than two occurrences. The difference between Multinomial and Gaussian lies in that multinomial fits discrete data whereas Gaussian Naïve Bayes fits continuous data (Gravita, 2022).

Text data that has passed preprocessing is used totrain the Naïve Bayes classification method using the Python programming language. This also requires label data, because when the label is removed, it is very difficult to clean data (Arboleda, 2019). In this process classification there are several stages as follows:

1. Featuring

   This process initializes the vector object used to convert the raw text document set to a matrix. Then use transform to fit the vector to the text data in the 'Tweet' column of the Data frame data set to produce a representation of the text data. In this process, there is a fit method that is used to train models on certain datasets. This method takes training data as input and trains the model based on the information provided.

2. Split data in training and test data

   This process is separated into training data and test data and can be used to assist the classification process by determining the keywords for a document that are needed in a system to group the same documents (Thamrin & Sabardilla, 2015).

3. Scanning Process

The scanning process for all data reading processes. Furthermore, these transactions will be grouped into several classes according to the similarity of the items in each transaction (Gunawan, 2016).

Classification in this system performs the process of data from the results of the previous pre-processing process. This will produce a level of accuracy in the text, where there is a comparison of the accuracy of the Naive Bayes algorithms.

2.5. Analysis and Evaluation

The total data of hate speech tweets totaled 4393 data, with the total share of hate speech tweet data namely 80% for training data and 20% for testing data. This test aims to see the effect of word pre-processing and compare the results of accuracy, precision, recall, and f-1 score in the pre-processing process with pre-processing and without the pre-processing. The results of testing using the Gaussian method are in Table 2.

Table 2. Accuracy Results Using Gaussian Classification

Evaluation	***Pre-processing***	***Without Pre-processing***
Precission	68%	68%
Recall	62%	63%
F1-score	62%	64%
Accuracy	62%	63%

The results of accuracy in the Gaussian classification using pre-processing is 62%. The results of testing using the Bernoulli method are in Table 3.

Table 3. Accuracy Results Using Bernoulli Classification

Evaluation	***Pre-processing***	***Without Pre-processing***
Precission	78%	79%
Recall	78%	79%
F1-score	78%	79%
Accuracy	78%	79%

The accuracy results in the Bernoulli classification using pre-processing is 78%. The results of testing using the Multinomial method are in Table 4.

Table 4. Accuracy Results Using Multinomial Classification

Evaluation	***Pre-processing***	***Without Pre-processing***
Precission	80%	79%
Recall	80%	79%
F1-score	79%	79%
Accuracy	80%	79%

The results of accuracy in the Multinomial classification using pre-processing is 80%. From this case, the highest or most optimal accuracy results are obtained by Multinomial classification with an accuracy of 80% by going through the pre-processing process. This is because data cleaning through pre-processing can produce good tweet data and does not interfere with the process of running system accuracy. And thus, the system succeeded in eliminating things that were less informative in the data.

3. FINDING AND DISCUSSION

In the classification of hate speech, the polarity of the text document used is a document that containshate speech or does not contain hate speech. Determining the polarity of a text document can be carried out using machine algorithm learning which is included in supervised learning. This algorithmrequires training data, namely pre-labeled text data, then the algorithm will classify the text from thetraining data, Naïve Bayes is an example of a supervised learning algorithm.

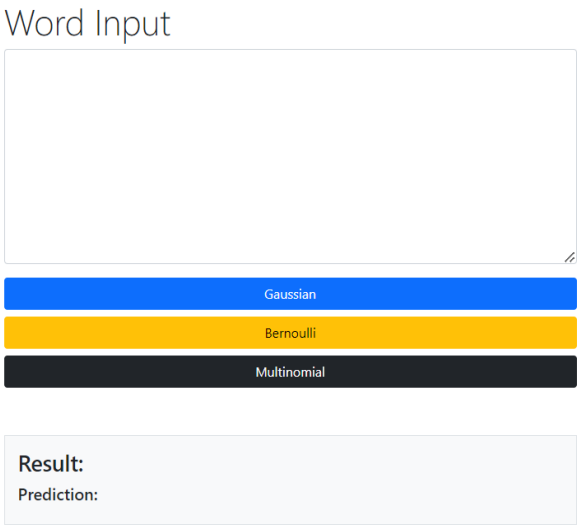
Naïve Bayes Classifier is a classification method derived from the Bayes theorem. It has several advantages, including a fast computing process, easy to implement with a simple, effective structure,and suitable for determining the polarity of documents. This method is simple but has goodperformance in sentiment classification in determining the polarity of text documents from various data sources and Indonesian text documents. By determining the research probability value, the accuracy of the classification of hate speech texts is 83% (Alwiah Musdar et al., 2021). This methodis proven to have high accuracy and also has fast computation time if implemented to solve a numberof on existing problem (Ivan, 2019)

Hate speech as intended aims to incite and incite hatred against individuals or groups of people in various communities (Fauzi et al., 2018). A public statement as we know “Hate Speech” is made consciously and intentionally with the intent to embarrass a group of people.

There is much to gain from American and British studies of HS and communication science, with 71 of the 123 studies selected from the United States (48) and the United Kingdom (23) combined (Paz et al., 2020). Withthe relevant data will be analyzed in each document which is considered a unit of analysis that determines the characteristics of each field of knowledge that discusses Hate Speech quantified.

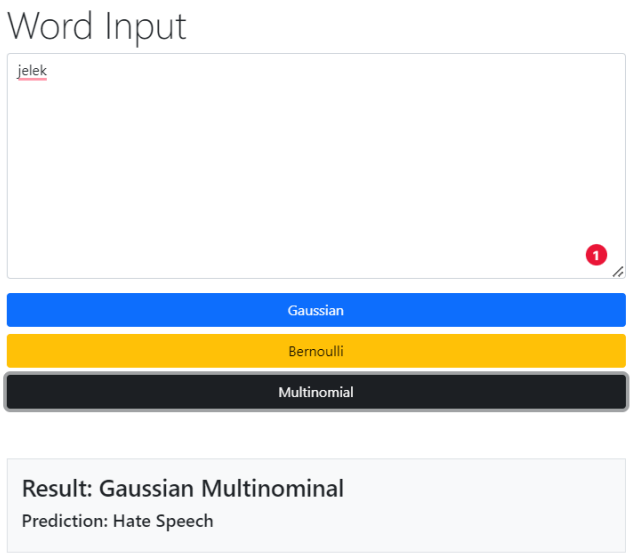
Hate speech is very easy to find in written communication on social media, such as Twitter, Facebook,and Instagram. Anyone can spread hate speech or abusive language to other people they don't like, one of which is in Indonesian (Ihsan et al., 2021). Hate speech is often used to express hateful ideology, and also as communication that disparages a person or a group based on several characteristics such as ethnicity, race, religion, sexual orientation, skin color, gender, nationality, or other characteristics. Several acts of hate speech are included in criminal acts, including insulting, committing blasphemy, doing something unpleasant, provoking, inciting, and spreading fake news (Ivan, 2019).

Several studies have been carried out, it has succeeded in achieving the goals and objectives of the research. The author carries the title of this research on Indonesian hate speech, which is carried outwith the aim of detecting and classifying words into a certain vocabulary. Therefore, this study aimsto find the level of accuracy by using the Naïve Bayes classification method. In this study, it can be implemented in a web-based user interface, which uses the Python programming language. The system is implemented using Flask API which is a framework for creating web applications in Python and using libraries in Python such as Scikit Learn, Pandas, etc (S Narayanan et al., 2022). By using flask as a framework that functions as a web application. Flask is a microframework based on the Python language which does not have many tools and libraries. (Ngantung & Pakereng, 2021).



Picture 3. Design Web Page

That it can be proven that the algorithm can already be implemented through a web application based on the user interface.



Picture 4. Output Process

4. CLOSSING

This hate speech detector research has been able to handle the classification process of the naive Bayes method which can be implemented into a web-based system. The results of testing the system with the Gaussian, Bernoulli, and Multinomial classifications show that the system can carry out all its functions correctly. The test results with the Multinomial classification type reach a percentage of 80% pre-processing and 79% by going without the pre-processing process. Where this is included in the category "Clear / Good" which means the system is acceptable. There are deficiencies in the implemented system that can only handle pre-processing accuracy. In the future, the system will be developed to handle more detailed accuracy without using pre-processing when needed. In this research, there are two data accuracy studies using pre-processing processes and not using pre-processing processes. This is because in the previous research only used the pre-processing process, where there is no other comparison to find the highest accuracy in the system in order to get optimal results.

BIBLIOGRAPHY

Alwiah Musdar, I., Angriani, H., Studi Informatika, P., & KHARISMA Makassar JlBaji, S. (2021). *Implementasi Teori Naive Bayes dalam Klasifikasi Ujaran Kebencian di Facebook*. *6*(4), 2622–4615. https://doi.org/10.32493/informatika.v6i4.12593

Arboleda, E. R. (2019). Comparing performances of data mining algorithms for classification of green coffee beans. *Int. J. Eng. Adv. Technol*, *8*(5), 1563-1567.

Ashari, H., Arifianto, D., & Al Faruq, H. A. (2020). Perbandingan Kinerja Algoritma Multinomial Naïve Bayes (MNB), Multivariate Bernoulli Dan Rocchio Algorithm Dalam Klasifikasi Konten Berita

Hoax Berbahasa Indonesia Pada Media Sosial. *Fakultas Teknik, Jurusan Teknik Informatika, Universitas Muhammadiyah Jember*.

Athira, W., Gholissodin, I., & Perdana, rizal setya. (2018). Analisis Sentimen Cyberbullying Pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *2*(11), 4704–4713. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/3051

Dwitama, A. P. J., & Hidayat, S. (2021). Identifikasi Ujaran Kebencian Multilabel Pada Teks Twitter Berbahasa Indonesia Menggunakan Convolution Neural Network. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *3*(2), 117. https://doi.org/10.30865/json.v3i2.3610

Fauzi, A., Rais, A. N., Faittullah Akbar, M., & Gata, W. (2018). *46 Seminar Nasional Teknologi Informasi Universitas Ibn Khaldun*.'

Gravita, A. (2022). Mengenal Algoritma Naive Bayes dan Kegunaannya. *PT Semua Mahir Teknologi (SMART)*.

Gunawan, D. (2016). Evaluasi Performa Pemecahan Database dengan Metode Klasifikasi Pada Data Preprocessing Data mining. *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika*, *2*(1), 10-13.

Guterres, A., & Santoso, J. (2019). Stemming Bahasa Tetun Menggunakan Pendekatan Rule Based. *Teknika*, *8*(2), 142-147.

Ihsan, F., Iskandar, I., Harahap, N. S., & Agustian, S. (2021). Decision tree algorithm for multi-labelhate speech and abusive language detection in Indonesian Twitter. *Jurnal Teknologi Dan Sistem Komputer*, *9*(4), 199–204. https://doi.org/10.14710/jtsiskom.2021.13907

Ivan, Y. A. S., & Adikara, P. P. Klasifikasi Hate Speech Berbahasa Indonesia di Twitter Menggunakan Naive Bayes dan Seleksi Fitur Information Gain dengan Normalisasi Kata. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN*, *2548*, 964X.Luthfi, H. W., & Riasti, B. K. (2013). Sistem Informasi Perawatan dan Inventaris Laboratorium pada SMK Negeri 1 Rembang Berbasis Web. Speed-Sentra Penelitian Engineering dan Edukasi, 3(3), 69-77.

N. T. Romadloni, I. S., & S. Budilaksono. (2019). Perbandingan metode naive bayes, knn dan decision tree terhadap analisis sentimen transportasi commuter line. *Jurnal Komputer Dan Informatika*, *3*, 1–9.

Narayanan, S., Balamurugan, N. M., Maithili, K., & Palas, P. B. (2022, May). Leveraging Machine Learning Methods for Multiple Disease Prediction using Python ML Libraries and Flask API. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 694-701). IEEE.

Ngantung, R. K., & Pakereng, M. I. (2021). Model Pengembangan Sistem Informasi Akademik Berbasis User Centered Design Menerapkan Framework Flask Python. *Jurnal Media Informatika Budidarma*, *5*(3), 1052-1062.

Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate Speech: A Systematized Review.In *SAGE Open* (Vol. 10, Issue 4). SAGE Publications Inc. https://doi.org/10.1177/2158244020973022

Rahmadhany, A., Aldila Safitri, A., & Irwansyah, I. (2021). Fenomena Penyebaran Hoax dan Hate Speech pada Media Sosial. *Jurnal Teknologi Dan Sistem Informasi Bisnis*, *3*(1), 30–43. https://doi.org/10.47233/jteksis.v3i1.

Sari, I. M., Wijaya, D. R., Hidayat, W., & Kannan, R. (2021, June). An approach to classify rice quality using electronic nose dataset-based Naïve bayes classifier. In *2021 International Symposium on Electronics and Smart Devices (ISESD)* (pp. 1-5). IEEE.

Thamrin, H. (2015). Efektivitas Algoritma Semantik dengan Keterkaitan Kata dalam Mengukur Kemiripan Teks Bahasa Indonesia. *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika*, *1*(1), 7-11.

Xu, K., Li, Y., Liu, C., Liu, X., Hao, X., Gao, J., & Maropoulos, P. G. (2020). Advanced data collection and analysis in data-driven manufacturing process. *Chinese Journal of Mechanical Engineering*, *33*(1), 1-21.