Application of The Naïve Bayes Algorithm for Hate Speech Detection

Novera Tri Dayanto[1]

STIKI Malang, Jl. Raya Tidar 100, Malang, East Java, Indonesia

Article Information

Received: 21-11-2024
Revised: 28-11-2024
Published: 05-12-2024

Keywords
Naïve Bayes; Hate Speech; Social Media; Text Classification.

Correspondence Email:
noverajunior11@gmail.com

Abstract
The spread of hate speech on social media using Indonesian continues to increase. Hate speech has become a problem that is detrimental to human life in society. In this research, the Naive Bayes algorithm is implemented to detect hate speech in social media comments. The dataset used in this research consists of 10,000 Indonesian language social media comments labeled as "Hate Speech" and "Non-Hate Speech". Accuracy for data processed using TF-IDF is 85% for Indonesian language data. Based on the results, Naive Bayes is a simple but effective analytical approach in detecting hate speech.

1. Introduction

Technological developments in the information sector have resulted in various social media applications emerging, such as Facebook, Twitter, Instagram and others [Untung Surapati, A. Y. (2023)]. Hate speech has recently attracted a lot of attention. Hate speech itself is defined as communication that aims to belittle people, groups or groups based on ethnicity, religion, race, ethnicity, class, nationality and other characteristics. The increase in internet users every year has an impact on the increasing number of hate speech spread on social media. Antariksa, Classification of Hate Speech in Tweets in Indonesian [Antariksa, Y. P. W. D. K. (2019)]. Hate speech on social media has become a global issue that affects various aspects of people's lives, including social relations, politics and security. Its rapid spread, supported by ease of access and anonymity on social media platforms, creates major challenges for content moderators and related authorities. To overcome this problem, an automatic detection system is needed to moderate content efficiently and in a timely manner, thereby reducing the negative impact of hate speech.

With so much hate speech spreading on the internet, of course it makes both victims and social media users who see it feel uncomfortable. This makes classifying hate speech very difficult, because there are no truly standardized standards for hate speech. Some find someone's comments/tweets on social media very hurtful, but maybe for others, this is not a problem. Not many victims of hate speech report it, either because they are afraid or because they don't care, meaning that hate speech behavior will always flourish in Indonesia.

There's so much hate speech floating around on the internet now it obviously makes people uncomfortable — victims and social media users who encounter it. This presents great challenges to the classification of hate speech, as there are "no truly standard standards for hate speech. Some would find this very hurtful but perhaps for others this is not an issue? Not many victims of hate speech even report it, either out of fear, or because they don't care, which means that hate

This research is limited to making a Naive Bayes model in detecting hate speech in Indonesian language texts. Data training are limited until October 2023. Nonetheless, Naive Bayes can work well enough if you preprocess well (e.g., remove stopwords, tokenize and possibly create TF-IDF); This work is part of a research on how the Naive Bayes algorithm carries out in detecting comments with hate speech on social media. It is hoped that using this approach the eventual system will be efficient in processing large amounts of data and accurate enough to identify and moderate problem content. The experimental results indicated that, despite its simplicity, Naive Bayes provides competitive performance numbers in text classification tasks, thus it is a suitable choice for real-world applications including the classification of hate speech in Indonesian-language social media.

1.1 Literature Review

Statements containing hate speech can cause conflict because readers are provoked or insulted when reading statements containing hate speech. When many readers of statements containing hate speech are provoked or insulted, the conflict will become worse. To prevent conflicts caused by creators or spreaders of hate speech, laws are needed that regulate hate speech. The law regarding hate speech is regulated in the Criminal Code (KUHP) in articles 155 and 157, as well as in the Information and Electronic Transactions Law (UU ITE) number 19 of 2016 in article 28 paragraph (2) and article 45A paragraph (2). However, there are still many people who make or spread hate speech (Pratama, 2020).

Based on the study of Asogwa et al. (2022), hate speech data is used for classification with 12 different labels, including individual, group, religion, race, physical, gender, other, mild hate speech, moderate hate speech, and severe hate speech. This research applies the Naïve Bayes (NB) algorithm and support vector machine (SVM). Meanwhile, Aljero et al. (2021) used genetic programming, Fatahillah et al. (2017) used the Naïve Bayes classifier, Ketsbaia et al. (2023) used multi-stage machine learning, Oriola and Kotze (2020) used machine learning evaluation techniques, Plaza et al. (2021) used multi-task learning, Sreelakshmi et al. (2024) used cost-sensitive learning, and Obaid et al. (2024) and Zhou et al. (2020) used deep learning algorithms to detect hate speech.

In Ibrahim and Budi's (2019) research, the logistic regression algorithm used to detect hate speech achieved an accuracy of 79.85%. This research categorizes hate speech based on several labels such as religion, race, and gender, as well as severity. Logistic regression provides increased classification accuracy compared to previous research which used other algorithms such as Naïve Bayes and support vector machines. In contrast to the research of Ahmad et al. (2019), who classified data focused on 12 different categories, the aim was to study the number and intensity of hate speech that appeared most frequently in various categories. This is considered important because, as mentioned by Ali et al. (2021) and Komnas HAM (2016), the number of social media users continues to increase in line with the increase in hate speech. Therefore, classification using several classifiers was carried out in this study using the NB algorithm (Haikal, 2024).

Furthermore, based on the findings of Kumar et al. (2023), hybrid algorithms combining deep learning and traditional machine learning methods have shown promising results in increasing accuracy for hate speech detection tasks. These hybrid models provide the advantage of leveraging the strengths of both methods, making them particularly effective in complex classification scenarios.

2. Research Methods

2.1 Dataset

The dataset consists of 10,000 social media comments, divided into two categories :

- Hate Speech: Comments that contain elements of insults, discrimination, or threats.
- Non-Hate Speech: Normal comments without hate elements.

Data is divided into 80% for training and 20% for testing.

2.2 Preprocessing

Preprocessing steps include :

1) Removal of irrelevant characters such as symbols, numbers and URLs.
2) Stopword removal using the NLP library for Indonesian.
3) Tokenize words using tools like NLTK or Sastrawi.
4) Data representation using TF-IDF.

2.3 Algorithm Naïve Bayes

The algorithm used is Multinomial Naive Bayes because it is suitable for text data with word frequency distribution. This model calculates the probability of each class based on text features and selects the class with the highest probability.

3. Result and Discussion

Table 1. Evaluation Table

Model	Akurasi	Precision	Recall	F1-Score
Multinomial Naïve Bayes	87%	85%	88%	86.5%
Bernoulli Naïve Bayes	83%	81%	84%	82.5%
Gaussiah Naïve Bayes	80%	79%	81%	80%

From the research results of using the Multinomial Naive Bayes algorithm, it gets the best performance compared to other variants, especially if the data is processed using the Term Frequency-Inverse Document Frequency or TF-IDF representation. This method uses the weights assigned to the words by their frequencies in a document compared to all communities. Using TF-IDF, the algorithm should be able to find words that contribute more to classifying and help it predict in a more accurate way between the “Hate Speech” and “Non-Hate Speech” categories.

However, despite its generally good performance, this model faces some limitations in recognizing more complex or implicit hate speech. The main challenges faced are two types of classification errors, namely False Positive and False Negative :

1. **False Positive**

   False Positive errors occur when comments that do not actually contain hate speech are incorrectly classified as “Hate Speech.” This can be caused by the use of certain words that often appear in the context of hate speech but are not intended to do so in this comment. For example, words like “stupid” or “lazy” may appear frequently in hate speech, but in certain contexts, they may be used in a neutral or inoffensive manner. As a result, the model may misjudge the comment as hate speech, which in turn lowers the user's level of trust in the detection system.

2. **False Negative**

   False Negative errors occur when comments that actually contain hate speech are not detected by the model and are classified as “Non-Hate Speech.” These errors are usually caused by implicit hate speech that is difficult to recognize. For example, hate speech that uses sarcasm, metaphors, or words that have double meanings is often not identified by algorithms. Models that rely on word frequency patterns, such as Multinomial Naive Bayes, tend to have difficulty with this type of text because they lack a deeper understanding of semantics or context.

3.1 Error Impact

False Positive errors can lead to over-moderation, where non-problematic comments are removed or blocked, which can lead to user dissatisfaction. Conversely, False Negative errors can allow the spread of hate speech to continue, thereby weakening the effectiveness of the detection system.

3.2 Improvement Recommendations

To overcome this challenge, a combination of Naive Bayes with other techniques can be considered, such as :

- Integration with deep learning-based models such as LSTM or BERT to capture the context of hate speech.
- Implementation of additional features such as sentiment analysis or linguistic pattern recognition to improve the model's ability to recognize implicit hate speech.
- Use of more diverse training data and include various types of hate speech, including sarcastic or implicit ones.

By reducing the number of these errors, Naive Bayes models can become a more reliable and practical tool for hate speech detection on social media.

4. Conclusions

The conclusion of this research shows that the Naive Bayes algorithm is a simple but effective approach for detecting hate speech on social media. With appropriate data representation, such as TF-IDF, this algorithm is able to process text and classify it with fairly high accuracy. The main advantage of Naive Bayes lies in its efficiency in handling large datasets and its ability to provide adequate results even with relatively light computing. This makes it a good choice for initial implementation of an automated detection system.

However, this research also revealed several limitations, such as difficulties in detecting implicit or complex hate speech, such as the use of sarcasm or metaphor. These limitations indicate that although the Naive Bayes algorithm is effective for the base case, there is room for improvement in terms of accuracy and ability to understand deeper context.

For further research, it is recommended that the Naive Bayes algorithm be integrated with other, more sophisticated approaches. For example, combining Naive Bayes with ensemble learning techniques such as bagging or boosting can improve model performance by reducing bias and variance. Additionally, leveraging deep learning-based models, such as LSTM or BERT, can help capture more complex hate speech contexts and improve detection in difficult cases.

With further development, Naive Bayes-based detection models combined with modern approaches can become a more reliable and effective tool in moderating content on social media. This will help create a digital environment that is healthier and free from the negative impacts of hate speech.

5. References

Antariksa, Y. P. W. D. K. (2019). Klasifikasi ujaran kebencian pada cuitan dalam bahasa Indonesia. Jurnal Buana Informatika, 10(2), 164–171.

Haikal, A. M. (2024). Optimalisasi algoritma Naïve Bayes untuk klasifikasi tweet berbahasa Indonesia dalam mengatasi hate speech di platform X. Indonesian Journal of Mathematics and Natural Sciences, 11(1), 108–114.

Kumar, A., Patel, R., & Sharma, P. (2023). Hybrid learning approaches for hate speech detection in social media: A comprehensive review. Journal of Artificial Intelligence Research and Applications, 15(2), 56–67.

Pratama, D. M. M. S. B. A. (2020). Penerapan Naïve Bayes Classifier dengan algoritma stemming Nazief dan Adriani untuk aplikasi deteksi ujaran kebencian berbasis web. Jurnal Komputer dan Aplikasi, 8(3), 227–236.

Untung Surapati, A. Y. (2023). Implementasi metode Naïve Bayes untuk mendeteksi hate speech pada Twitter. Journal of Information Technology and Computer Science (INTECOMS), 6(1), 830–837.