Sentiment Analysis using Naive Bayes in Rude Word Replacement on Digital Platforms

Ilham Aufal Hadad [1], Endah Sudarmilah [2]

[1]Informatics Engineering Department, Pabelan, Surakarta, Indonesia
[2]Informatics Engineering Department, Pabelan, Surakarta, Indonesia

Email korespondensi: l200214071@student.ums.ac.id, Endah.Sudarmilah@ums.ac.id

Abstract. Indonesia has entered the era of revolution 4.0, which in fact uses a lot of digital access in its daily life. One of the most popular digital accesses is social media, a place where everyone can connect with each other online, which certainly brings good and bad impacts. An example of a bad impact that we often encounter is the use of harsh words to fellow social media users. This research aims to minimize the existence of the use of harsh words by filtering every harsh word that appears. The research uses Machine Learning with the Naïve Bayes method by training a dataset of harsh words and a special dictionary for subtle words as a translator in a more subtle language. The results of the research are expected to provide insight into the accuracy of using the Naïve Bayes method and provide output of word filtering features that have an impact on the character building of digital platform users anywhere and anytime.

Keywords: *Rude Word; Analysis Sentiment; Naïve Bayes; Translation*

INTRODUCTION

The life of the modern era will certainly not be far from digital life through the internet which can be accessed anytime and anywhere via various mobile devices. Social media is one of the tools to communicate to seek and receive the latest national and international information (Yanti et al., 2021).

Excessive internet use, especially on social media, can lead to mental health issues such as anxiety, depression and loneliness (Mohammad & Ryca Maulidiyah, 2023). Based on (Gunasekaran, 2019) research, today's mobile crowdsourcing applications generally consist of various Internet media, including online news, microblogs, blogs, and forums. Many people abuse the ethics of socializing in cyberspace by uploading or commenting on posts using harsh words called verbal violence (Susetyo et al., 2020). A large number of social network users often leads to uncontrolled communication and many netizens communicate with abusive language (Tjahyanti, 2020).

UNESCO in 2015 conducted a study entitled "Countering Online Hate Speech" which stated that the growing phenomenon of hate speech online has caused various problems both inside and outside Europe (Astuti, 2019). Rude words can be expressed by mentioning certain types of animals, such as dogs, monkeys, and so on (Hidayatullah et al., 2019). According to (Adelia & Mayong, 2022), the impact of social media also affects the formation of student's language ethics who tend to imitate new or trending language on social media. One way technology is utilized in preserving language is using automatic language translation machines (Purwaningsih & Wahyu, 2023).

According to (Fikri et al., 2020), Sentiment analysis belongs to one of the fields of Natural Language Processing (NLP) and is a process used to help identify the content of a dataset in the form of opinions or views referring to positive, negative, or neutral sentiments. Naïve Bayes Method is used because it has a high probability of up to 73.81% and has a strong mathematical basis using a probability model (Yudhanata & Sudarmilah, 2024).

The previous research has conducted research in the form of sentiment analysis that implements the Naïve Bayes Classifier method with data from opinions on the twitter social media platform. The analysis research uses NLP in understanding everyday language. Previous research has also shown that the accuracy obtained is fairly high with the level obtained is 90% with details of the 92% precision value, 90% recall and 90% f - measure. Another research, according to the research of (Sari et al., 2023), Naïve Bayes classification test results obtained an accuracy of 93.33% as well as for power efficiency with low power mode.

The novelty of research is the stage when after finding the labelling data in the rude or subtle classification, there will be another stage that replaces certain words with words that have been provided in a special dataset.

METHOD

Natural Language Processing (NLP) can be used to convert unstructured text into structured form, perform syntactic processing such as tokenization, capture meaning by assigning ideas to words or groups, and discover relationships between concepts (Khanbhai et al., 2021). Complete with algorithms in research, Naïve Bayes is the choice in the execution of sentiment analysis datasets because it has certain advantages like classification by Naïve Bayes can handle datasets with correlated patterns with the aim of increasing the rarity of solutions. (Blanquero et al., 2021). Not only that, according to (Tanggraeni & Sitokdana, 2022) The Naïve Bayes algorithm has another advantage, which is that it is efficient because it is able to make the process short with accuracy that tends to

be high. Sentiment analysis research using the Naïve Bayes algorithm will be tested through the workflow as shown in Figure 1.

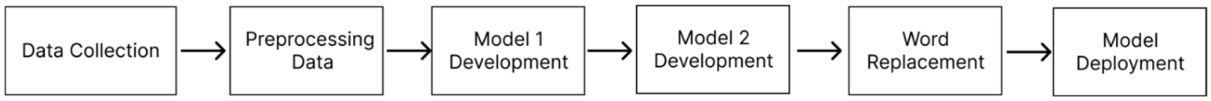

Figure 1. Workflow of Method

Data Collection

The data is collected in the form of text taken from Kaggle site. The dictionary is taken any link of GitHub community, github.com/drizki/indonesian-badwords and github.com/dikako/list_badword. The dataset contains blasphemy against a youtuber named Ericko Lim in https://www.kaggle.com/code/alvinf/eda-data-komentar-youtube-toxic.

Preprocessing Data

Data preprocessing involves the system in matching words that are still raw due to human writing is unreadable. This stage involves methods as in figure 2.

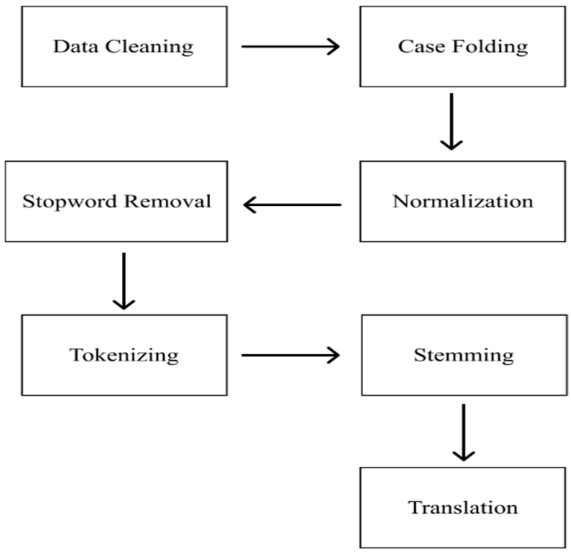

Figure 2. Workflow of preprocessing data

The Data Cleaning stage clean punctuation, symbols, emojis, and so on. The Social Media community has its own slag language for posting messages where the message contains symbols, misspelled words and sarcastic sentences (Singh et al., 2020). The next step is Folding stage, change all of letters become lowercase for consistency. Normalization stage fix words from any wrong spell and correct it. The dataset will be

tokenized so that the data becomes separated per word, make it easier to process and detect rude words. Stop word Removal stage will removes words that are not important in detection. Stemming remove words that have the same meaning or synonyms and leave one alone in order to make the data more efficient. Translation takes over the words from English language to Indonesia language, with the purpose of align the data. A brief overview of the process is provided in Table 1.

Table 1. Example data before and after of Preprocessing Data

Step	Before	After
Data Cleaning	Hahaha dnger tuh heters t o l o l 🤣	Hahaha dnger tuh heters tolol
Case Folding	Hahaha dnger tuh heters tolol	hahaha dnger tuh heters tolol
Normalization	hahaha dnger tuh heters tolol	hahaha denger tuh heters tolol
Tokenizing	hahaha denger tuh heters tolol	["hahaha", "denger", "tuh", "heters", "tolol"]
Stop word Removal	["hahaha", "denger", "tuh, "heters", "tolol"]	["dnger", "heters", "tolol"]
Stemming	["denger", "heters", "tolol"]	dengar haters tolol
Translation	dengar haters tolol	dengar pembenci tolol

Model 1 Development

Model 1 plays a role in sentiment analysis by labeling either positive, neutral, or negative. There are five stages in model 1 processing from labeling to model evaluation.

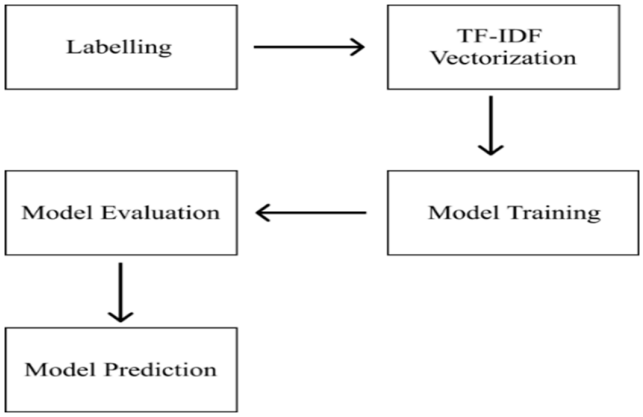

Figure 3. Model 1 Workflow

Labelling Change the data that was in the form of string into numeric with the aim that the system can process the data because the system detects integers. The TF-IDF Vectorization stage divides the training data and test data, then converts them into floats so that the system can read the data. The model will be trained using Multinomial Naive Bayes with data that has been divided between positive/neutral/negative. The Prediction model plays a role for the classification of new data that will get the appropriate sentiment. Prediction will be done through analyzing the data that has been processed by the model. A brief overview of the process is provided in Table 2.

Table 2. Example data before and after of model 1 development

Step	Before	After
Labelling	['positif','netral','negatif']	['1','0','-1']
TF-IDF Vectorization	dengar pembenci tolol	0.7675

After run the training the whole stage, the model will be calculated such as precision, recall, f1- score, and accuracy using test data and compared with the results of the training data. The following is the formula for calculation:

$$P(C|X) = \frac{P(X \mid C) \cdot P(C)}{P(X)} \quad (1)$$

Description:

P(c|x) : the posterior probability of class (c, target) given predictor (x, attributes).

P(c) : the prior probability of class.

P(x|c) : the likelihood which is the probability of the predictor given class.

P(x) : the prior probability of the predictor

The model evaluation score will be a measure of whether the model is suitable for making sentiment predictions. Prediction model plays a role for the classification of new data that will get the appropriate sentiment. Prediction will be done through analyzing the data that has been processed by the model.

Table 3. Final results of model 1 development

Step	Before	After

Model Prediction	dengar pembenci tolol	negatif

Model 2 Development

Model 2 plays a role in sentiment analysis that labels either rude or subtle words. There are six stages in model 2 processing from labeling to TF-IDF frequency scoring that shown in the figure 4 below.

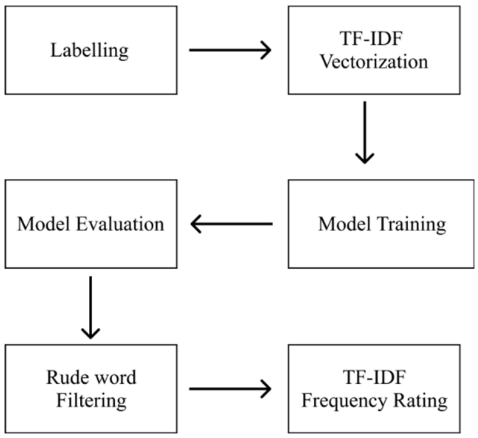

Figure 4. Model 2 Workflow

The labelling stage will be classifying the data between rude and subtle words. The classification stage using the Naïve Bayes algorithm involves counting the number of occurrences of words in the dataset for each class (Pamungkas & Fattah, 2024). The TF-IDF Vectorization stage divides the training data and test data, then converts them into floats so that the system can read the data. The model will be trained with data that has been divided to fit the researcher's expectations and will be labeled as rude/subtle.

Table 4. Example data before and after of model 2 development

Step	Before	After
Labelling	dengar pembenci tolol	kasar
TF-IDF Vectorization	dengar pembenci tolol	0.7675

The model will be calculated with various aspects such as precision, recall, f1-score, and accuracy using test data and compared with the results of the training data. The model's accuracy is determined by its number of accurate predictions, while precision measures its ability to make correct predictions among positive ones. Its sensitivity is measured by its F1 Score, which is evaluated for viability. The Rude Word Filtering stage capture rude words present in the data with the help of a rude word dictionary as a detection lever. The Data Frame that contains the rude word will be recalculated for its frequency in the dataset. Any rude word that has a rating above 0.5 will be dealt with as a threshold for using model 1.

Word Replacement

The model will able to target words that are considered rude word, the system automatically replaces the rude word with subtle word.

Table 5. Results before and after model 1

Step	Before	After
Rude Word Filtering	dengar pembenci tolol	tolol
TF-IDF Frequency Rating	tolol	0.8
Word Replacement	tolol	tidak pengertian

Model Deployment

The Model deployment stage will implement models with the applied method in this design is classification based on text classification. Text Mining is the process stage of analyzing data in the form of text where the data source is obtained from a document such as a word data sentence, The concept of text mining is usually used in the classification of textual documents where these documents will be classified according to the topic of the document (Darwis et al., 2021).

The Naïve Bayes algorithm predicts future sentiment based on previous experience so that it is known as the Bayes Theorem obtained from a document such as a word data sentence, the concept of text mining is usually used in the classification of textual documents where these documents will be classified according to the topic of the

document (Duei Putri et al., 2022). The framework used for model development is Streamlit which is an open-source python coding framework for building web applications or "web-apps" and is now used by researchers to share large datasets of published studies and other resources (Nápoles-Duarte et al., 2022).

RESULT

The research was characterized by various planning changes in order to get the highest potential of the research objectives.

Changes Occured During The Research

Some data added manually to complete the special case, it needed because the dataset have no correlation with special case so that the program can also detect relatively rude words that actually have meanings like positive sentiments in general. The example of data needed is, "I have two dogs in home". The word dog can be positive or negative regarding to its context and will be replaced if the sentiment is not positive.

The researcher utilized the rude word dictionary for negative sentiment detection, provided that the dataset used was hate speech so that the context was also equally negative. Researchers also initially used the K-Means Clustering method for dataset maturation outside of the research method, but the accuracy obtained was only around 20%, this happened because K-Means Clustering uses polarity that does not have an Indonesian library.

Method Implementation

Preprocessing Data

The dataset used is is still raw, there are still many symbols, random data patterns, and so on. Data cleaning section detects various unstructured data, such as URLs, tags, emojis, etc. The purpose of case folding make data becomes equal so that data checks such as duplicates can be more easily detected. The normalization stage replace words with incorrect spelling and it can't detect all words due to limited content in dictionary. The function is using sastrawi package with additional words to maximize the removal. Tokenizing stage separated each words to facilitate machine analysis into smaller units in the form of words. The stemming function is using sastrawi package to run.

Not all data replaced to its original due to the dataset was unstructured. Additional backup needed cause of long time process in stemming. The t ranslat ion of dataset contain mixture of english and indonesia language so it needs to be equalized to match the installed library. Additional backup needed cause of a long time process in translate section.

Model 1 Development

Each sentiment label will be converted to an numeric using map labelling. The dataset will also be determined by the feature variables and targets for training the Naive Bayes model. The TF-IDF system ensures to clean the data from stopwords so that the package stop_words_id will be used for further removal. After that, the data will converted into numeric. The next stage creates a model using Multinomial Naive Bayes and processing the training data. The TF-IDF feature trains data that has been converted into a vector presentation.

The purpose of model evaluation to extent the model can make predictions in accordance with the expectations of the researcher. The results of the model training data will compared against the results of the test data to measure the percentage difference. The prediction result is a numeric between 1,0,-1 as a representation of sentiment. The output will show what is the sentiment of text.

Model 2 Development

The Labelling function works by tokenizing the data and checking if there is a rude word in the sentence. The rude word detection will be matched with the key from when there is a match, unless if it is subtle. Then dataset is converted into vector representation through TF-IDF by the package installed in project. Training stage creates a model using Multinomial Naive Bayes and processing the training data only about 20% of whole data. Evaluation stage shows the score of the model in aspects of accuracy, precition, recall, and f1-score. The purpose of model evaluation is to see the extent to which the model can make predictions in accordance with the expectations of the researcher. The results compared against the test data result to measure the percentage difference.

After the result of comparison appear, function is going to tokenizing and checking each words, if a rude word is detected in kamus_kasar.json key, then the rude word is highlighted. TF-IDF Frequency Rating take the filtered rude word only and rating the frequency, so we can check how many the rude word detected. Take the score has value above 0.5 and use it as rude word treshold of positive sentiment. Last stage of model work is check the rude word by using TF-IDF score, then system will replace only if the sentiment is not positive.

Model Deployment

This stage create the function to predict the sentiment by new text so the system convert it to sav file for use it in streamlit. Create new streamlit file with python, the first

column has the main role that store the input results. There is also apply button that preprocess the data, followed by sentiment prediction which will be adjusted to the model.

The second column plays a role in displaying various variables that are executed in the first column. After doing customize in code, run the streamlit_app.py and check the web page. Commit all files and data in the github repository for synchronization with Streamlit. Sync github with streamlit and choose a domain name for the App URL.

Analysis and Evaluation

Accuracy testing acts as an indicator of the success of the development model which will later be implemented on the Streamlit Website. The following are the results of accuracy, precision, recall, and f -1 score for each model:

Model 1 (Positive, Negative, and Neutral)

Table 7. Model 1 Accuration on Data Evaluation

Metric	Score (%)
Accuracy	81
Precision	82
Recall	81
F1-Score	79

Model 2 (Rude and Subtle)

Table 8. Model 2 Accuration on Data Evaluation

Metric	Score (%)
Accuracy	91
Precision	92
Recall	92
F1-Score	91

This research uses supervised learning with Naive Bayes that is proved to have a simple fast and precise performance. The changing accuracy of the first

Figure 5. Processing of Rude Word with negative sentiment

Figure 6. Processing of Rude Word with positive sentiment

model from 75% to almost 80% after changing the threshold. The data containing rude words above a frequency value of 0.5 will be followed up. Here is the implementation via streamlit:

Figure 5 shows that when we input a sentence with negative sentiment, the model will replace it with a more subtle basic sentence. Figure 6 shows that when we input a positive sentence, the model will not replace any words.

CLOSSING

The researchers managed to create a machine learning model that not only detects but also replaces with more subtle vocabulary. The system provides a different flavor when dealing with special cases so that it does not seem to only replace rude word. The system performed quite well at high percentage, around 80% of model 1 and 90% of model 2. Hopefully the model can be utilized for every chat in the future. Many aspects of the research are still lacking such as library limitation and unique labeled data limitation, so in the future, research can still be improved for system improvement.

CONCLUSION

The author states that this study demonstrates how the accuracy of Machine Learning using the Naïve Bayes method is capable of identifying and transforming speech that may be harsh or non-harsh into a more polite lexical form in the context of digital communication. Through the two models developed, the system demonstrates adequate performance; Model 1 achieves 81% accuracy in classifying positive, negative, and neutral sentiments, while Model 2 achieves 91% accuracy in detecting coarse and non-coarse language. Both models employ a semi-supervised learning approach, where the database is not entirely original, though it has met the author's expectations thus far.

The construction of this model implies that the Naïve Bayes algorithm has the capability to analyze unstructured text data with high efficiency, provided it is supported by systematic and comprehensive preprocessing steps. This finding opens up opportunities for the use of similar systems in supporting language ethics in the digital realm, although its optimization is still constrained by data corpus limitations and dependence on external dictionary sources. In the near future, it is anticipated that expanding training data and integrating with more adaptive lexicon sources will be the primary recommendations for further research, which will not only utilize the Naive Bayes approach but also collaborate with other approaches.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Allah SWT for His blessings and grace in completing this thesis. I am very grateful to my supervisor, Dr. Endah Sudarmilah, S.T, M.Eng for her invaluable guidance, patience, and valuable input during this research process. I would also like to extend my heartfelt thanks to all the faculty members and staff of the Computer Science Program for their invaluable guidance and support. I sincerely thank my family and dear friends for their unwavering support, constant encouragement, and continuous motivation.

REFERENCES

[1] Adelia, D. P. N., & Mayong. (2022). Krisis Kesantunan Berbahasa Dalam Kolom Komentar Media Sosial Tiktok. Sastra, Dan Pengajaran, 1(1). https://vt.tiktok.com/ZGJS3a999cS/

[2] [12] Astuti, F. (2019). Perilaku Hate Speech Pada Remaja di Media Sosial Instagram.

[3] Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., & Sillero-Denamiel, M. R. (2021). Variable selection for Naïve Bayes classification. Computers & Operations Research, 135, 105456. https://doi.org/10.1016/J.COR.2021.105456

[4] [13] Duei Putri, D., Nama, G. F., & Sulistiono, W. E. (2022). Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier. Jurnal Informatika Dan Teknik Elektro Terapan, 10(1).

[5] Fikri, M. I., Sabrila, T. S., Azhar, Y., & Malang, U. M. (2020). Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter. 10.

[6] Gunasekaran, A. (2019). Behavioural Based Online Comment Spammers in Social Media. International Journal of Innovative Technology and Exploring Engineering, 9(1S), 175–179. https://doi.org/10.35940/ijitee.a1037.1191s19

[7] Hidayatullah, A. F., Aulia, A., Yusuf, F., Juwairi, K. P., Abida, R., & Nayoan, N. (2019). Identifikasi Konten Kasar pada Tweet Bahasa Indonesia. In JLK (Vol. 2, Issue 1). https://t.co/YQCC0CM4gG

[8] Khanbhai, M., Anyadi, P., Symons, J., Flott, K., Darzi, A., & Mayer, E. (2021). Applying natural language processing and machine learning techniques to patien experience feedback: A systematic review. In BMJ Health and Care Informatics (Vol. 28, Issue 1). BMJ Publishing Group. https://doi.org/10.1136/bmjhci-2020-100262

[9] Mohammad, W., & Ryca Maulidiyah, N. (2023). Triwikrama: Jurnal Multidisiplin Ilmu Sosial Pengaruh Akses Internet Terhadap Aspek Kualitas Kehidupan Masyarakat Indonesia. 01, 30–45.

[10] Pamungkas, E. W., & AlFattah, H. (2024). Pendeteksian Ujaran Seksisme Pada Platform X Dengan Algoritma Machine Learning Tradisional.

[11] Purwaningsih, L., & Wahyu, E. (2023). Perbandingan Kinerja Sistem Neural Machine Translation Opennmt Dan Thumt Dalam Eksperimen Menerjemahkan Bahasa Jawa Ngoko-Krama.

[12] Singh, N. K., Tomar, D. S., & Sangaiah, A. K. (2020). Sentiment analysis: a review and co mparative analysis over social media. Journal of Ambient Intelligence and Humanized Computing, 11(1), 97-117. https://doi.org/10.1007/s12652-018-0862-8

[13] Susetyo, Ariesta, R., & Utoro, D. Y. S. (2020). Kekerasan Verbal Dalam Media Sosial Facebook. 3(2). https://doi.org/10.31540/silamparibisa.v3i2

[14] Tanggraeni, A. I., & Sitokdana, M. N. N. (2022). Analisis Sentimen Aplikasi E-Government Pada Google Play Menggunakan Algoritma Naïve Bayes. 9(2), 785–795.

[15] Tjahyanti, L. P. A. S. (2020). Pendeteksian Bahasa Kasar (Abusive Language) dan Ujaran.

[16] Yanti, L. P. F., Suandi, I. N., & Sudiana, I. N. (2021). Analisis Kesantunan Berbahasa Warganet Pada Kolom Komentar Berita Di Media Sosial Facebook. In Jurnal Pendidikan dan Pembelajaran Bahasa Indonesia (Vol. 10, Issue 1).

[17] Yudhanata, D. I., & Sudarmilah, E. (2024). Analisis Sentimen Terhadap Isu Islamofobia Pada Platform Twitter Menggunakan Metode Klasifikasi Naïve Bayes.

[18] Astuti, D. P., & Sari, C. A. (2023). Rancang Bangun Sistem Deteksi Hipoksia Menggunakan Metode Naïve Bayes Berbasis Mikrokontroler. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 7(1), 123-130.

[19] Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan Algoritma Naive Bayes untuk Analisis Sentimen ReviewData Twitter BMKG Nasional. 15, 131–145.

[20] Nápoles-Duarte, J. M., Biswas, A., Parker, M. I., Palomares-Baez, J. P., Chávez-Rojo, M. A., & Rodríguez-Valdez, L. M. (2022). Stmol: A component for building interactive molecular visualizations within streamlit web-applications. Frontiers in Molecular Biosciences, 9.https://doi.org/10.3389/fmolb.2022.9908