Journal of Computing and Data Science (JODA)
Vol. 1, No. 1, 2023, pp. 47~53

Classification of Hate Speech Against Cak Nun on Twitter Multinomial Naive Bayes

Irfan Aufa Fadilla*
Jurusan Teknik Informatika, Fakultas Sains & Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim, Malang
200605110086@student.uin-malang.ac.id

*Corresponding Author

Abstract

This research aims to identify hate speech against Cak Nun on Twitter social media using the Multinomial Naive Bayes method focusing on text pre-processing. Pre-processing involves case folding, tokenization, stopword removal, and stemming to improve classification accuracy. The tweet data taken from Cak Nun's Twitter account was analyzed to measure the level of hatred using the Multinomial Naive Bayes classification model. The case folding process is used to convert all text into lowercase letters, tokenization is performed to break the text into tokens that can be processed, stopword removal aims to remove common words that do not contribute significantly to sentiment analysis, and stemming is implemented to convert words into their basic form. The results show that pre-processing improves classification performance, achieving an accuracy of 85.135%. The findings contribute to creating a more positive and safe social media environment.

Keywords: Cak Nun, Hate speech, Multinomial Naive Bayes, Twitter

INTRODUCTION

During the digital era, numerous facets of human existence have transformed, most notably how individuals retrieve information. The digital age has given rise to a profound metamorphosis in information. Specifically, hate speech has become effortlessly disseminated and embraced by numerous individuals via social media. In the bygone era, information dissemination was constrained by various factors, including temporal limitations, financial considerations, and accessibility constraints. Nevertheless, owing to the ubiquity of social media platforms, information can now be readily accessed at any time and from any location. It undoubtedly confers advantages such as heightened information accessibility and greater citizen participation in the democratic processes. Conversely, the information age also facilitates the propagation of hate speech [1].

Any mode of communication to assail an individual or a collective based on their race, religion, tribe, ethnicity, gender, sexual orientation, or disability is regarded as hate speech. Hate speech comprises a form of prejudice executed through language and possesses a multitude of adverse consequences, both on an individual and societal level. Victims of hate speech may experience feelings of seclusion, intimidation, and even psychological trauma, while the broader impacts can result in societal divisions, violence, and even conflict.

Cik Nun, also known by his full name Emha Ainun Najib, is a prominent Indonesian public figure who frequently becomes the object of hate speech. It is primarily attributed to Cak Nun's tendency to criticize various social and political matters in Indonesia strongly. Cak Nun's perspective often attracts hostility from various circles because he is known for his brave and critical nature. The hate speech directed towards Cak Nun can be viewed as an attempt to conceal or undermine his critical views.

One endeavor to counteract the dissemination of hateful rhetoric directed towards Cak Nun encompasses the construction of a system endowed with the capability to identify instances of hate speech effectively. This system can be developed by

implementing machine learning techniques to discern recurring patterns associated with hate speech. An approach to uncovering these patterns entails examining the lexical units, idiomatic expressions, and syntactic structures that are recurrently employed to convey sentiments of hatred.

The initial investigation of this study was conducted under the title 'Classification of Hate Speech in Chinese in Bahasa Indonesia.' This investigation employed three techniques: Support Vector Machine, Logistic Regression, and Naïve Bayes [2]. This investigation aimed to assess the relative efficacy of the three abovementioned techniques in classifying hate speech in Indonesian.

The inaugural investigation associated with this particular study encompassed an examination titled 'Categorization of Hate Speech in Chinese Conceived in Bahasa Indonesia.' This examination employed three techniques: Support Vector Machine, Logistic Regression, and Naïve Bayes [2]. This investigation's primary objective revolved around juxtaposing three more productive techniques in categorizing hate speech in the Indonesian language.

The subsequent associated investigation emerges from a scholarly inquiry titled "Classification of Hate Speech on Social Media Twitter Using the Support Vector Machine." This examination employed the social media platform Twitter as the focal point of scrutiny. Additionally, the research utilized the Support Vector Machine (SVM) approach to classify hate speech [3].

METHODS

In this study, a series of steps are taken. The research process begins with acquiring data and manually categorizing three classes: positive, negative, and neutral. Following the allocation of labels, the data proceeds to the pre-processing stage. This stage entails various procedures, including transforming all characters in the dataset to lowercase letters (Case Folding), segmenting the text into word fragments (Tokenization), eliminating insignificant words (Stop Word Removal), and reducing the word to its base form (Stemming) [4]. Once the pre-processing process is finalized, the data is transmitted to the Classifier for classification, determining whether the text is classified as open or hate speech. The research method is illustrated in Figure 1.

Data was gathered from 122 tweets categorized as positive, 122 as negative, and 122 as neutral. These tweets were sourced from the social media platform Twitter. Following the collection process, a manual labeling procedure was employed to assign appropriate labels to the tweets. The IT Team of PT facilitated data collection for this study—Arion Indonesia, which utilized a web crawler. The resulting data was subsequently stored in an xlsx file.

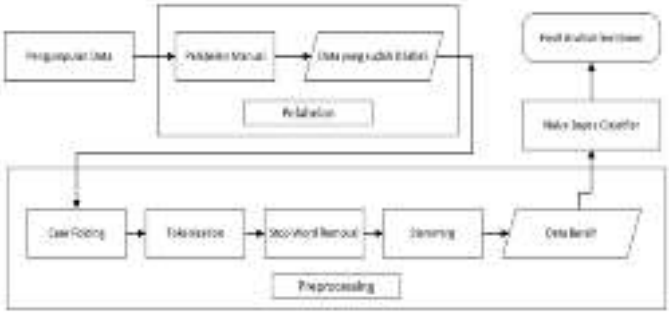

Figure 1. Hate Speech Classification Methods

The data that was previously analyzed undergoes several preliminary phases. The process of pre-processing commences with the conversion of all characters to lowercase. Additionally, each text document is divided into multiple words during the tokenization stage. Subsequently, any word lacking significance or meaning is eliminated after tokenization. Next, each word is transformed into its raw form during

the Stemming stage. The pre-processing procedure uses the Python Programming Language, the NLTK Library for tokenization and stops words and the Sastrawi Library for stemming. The data set consists of two parts: one for training the Classifier (Training Data) and another for evaluating the Classifier (Test Data).

The utilization of pre-processed text data, along with the determination of labels, is employed in the instruction of the employed classification techniques. By utilizing the Scikit-learn library in conjunction with the Python programming language, the application of the Naïve Bayes Multinomial classification method becomes feasible. The Multinomial Naïve Bayes model is a valuable tool for constructing Bayesian algorithms that can effectively classify texts or documents. Furthermore, in addition to including observed words, the Multinomial Naive Bayes Classifier formula specifies the class of the document [5].

$$CMAP = \underset{c\epsilon\{cl,cs\}}{\operatorname{argmax}} P(c|d)$$

$$= \underset{cs\{cl,cs\}}{\operatorname{argmax}} P(c|d)\, P(c)\prod P(tk|c) \qquad \text{.....(1)}$$

The likelihood probability parameter P(tk|c) in equation 1 is estimated by counting the occurrence of tk in all training documents in c, using a Laplacian prior, as shown in equation 2.

$$P(tk\ |c) = \frac{1+Nk}{V|+N} \qquad \text{.....(2)}$$

The variable Nk represents the count of instances of " kindergarten " in the training document labeled as c, while N signifies the overall count of word instances within c.

RESULTS AND DISCUSSION

DATA ANALYSIS

In this study, information gathered from the social media platform Twitter, which includes private direct messages, comments, and the textual content of posts, was employed to categorize hate speech directed towards Cak Nun. Subsequently, this information underwent a process of purification to eliminate any extraneous elements, such as punctuation marks and errors in spelling. Following this purification process, the information was then divided into two distinct categories: instances of hate speech and instances that did not qualify as hate speech. The process of categorization was conducted manually by individuals.

The classified data is subsequently partitioned into two sets: the training and testing data. The training data is employed to establish classification models, whereas the testing data is utilized to assess the performance of said models. The testing data was comprised of a total of 74 data points. In comparison, the training data encompassed a more extensive set of 292 data points, consisting of 97 positive tweet data, 97 negative tweet data, and 98 neutral tweet data.

Data sharing was carried out randomly to guarantee the representativeness of the train and test data for the entire data population [6]. The expectation is that the classification model generated through this approach will yield precise outcomes.

PRE-PROCESSING

The initial step in the pre-processing procedure involves case folding. The outcome of this case folding operation is presented in Figure 2. All characters have been transformed into lowercase.

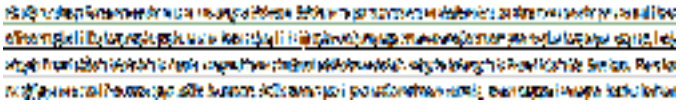

Figure 2. Case folding Results

The subsequent pre-processing procedure is known as tokenization. Tokenization, as a process, converts text or documents into smaller components known as "tokens." These tokens can be words, phrases, sentences, or other elements of language. The primary objective of tokenization is to facilitate the analysis or processing of text [7]. When case folding text data in the format of sentences, it is divided into individual words. The outcome of tokenization is illustrated in Figure 3. It is worth noting that instead of sentences, each line in a text document contains a collection of words from the text document.

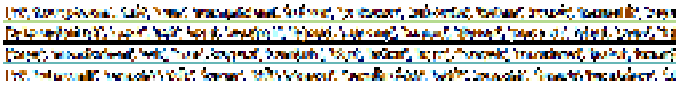

Figure 3. Tokenization Results

The outcome of tokenization subsequently undergoes the process of Stopwords Removal. Within this procedure, words that are deemed neutral and devoid of any polarization are excluded [8]. These words, known as stopwords, include examples such as "and," "or," "of," and so forth. The process of Stopwords Removal is executed by utilizing the NLTK library, which already incorporates a compilation of stopwords for Bahasa Indonesia. Words present in this compilation of stopwords are omitted from the textual data. The outcomes of the Stopwords Removal process can be observed in Figure 4.

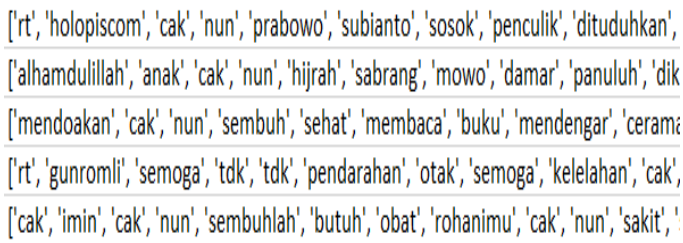

Figure 4. Stopwords Removal Results

The final step in the pre-processing phase involves stemming. Stemming entails the elimination of prefixes and suffixes from a word, thereby treating words with a joint base as identical [9]. In this research, the Sastrawi library was utilized for stemming purposes, offering a dictionary for Indonesian text documents. The outcomes of the stemming procedure are depicted in Figure 5.

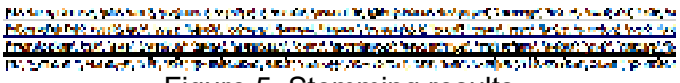

Figure 5. Stemming results

Once the Pre-processing stage has been completed, the text data is partitioned into two distinct segments. The initial segment pertains to the text data designated for training the Classifier, whereas the remaining segment assumes the role of test data.

TESTING

A grand sum of 74 test data was subjected to testing using the Classifier that was constructed. The outcomes of the classification are juxtaposed with the manually assigned labels. The results exhibit the level of precision at which the Classifier is utilized. The ensuing equation presents the calculation employed to derive the percentage of accuracy. Accuracy is procured through the summation of the accurately predicted test data, divided by the summation of the entire dataset, and subsequently multiplied by 100 [10].

$$\text{Akurasi} = \frac{\text{Jumlah data uji dengan polaritas yang sesuai}}{\text{Jumlah seluruh data uji}} \; x\; 100\% \quad \text{.....(3)}$$

The classification results of the test data using the Naïve Bayes Multinomial technique are presented in Table 1, where they are compared to the manual labels. Additionally, Table 2 displays the classification results for text documents that do not align with the manual labels. The label column in both tables represents the polarity manually assigned to each text document, while the classification result column indicates the label assigned by the Classifier.

Table 1. Corresponding Classification Results

No	Text Data	Label	Result
1	Semoga lekas pulih Cak Nun dan bisa berbagi ilmu kepada para generasi muda Indonesia lagi 🙏	positive	Positive
2	kamu plintir jadi menyerang dgn penuh kebencian. Asu Kowe.	Negative	negative
3	Pemulihan Cak Nun Usai Pendarahan Otak Berjalan Baik, Kesadaran... Emha Ainun Nadjib atau Cak Nun sedang menjalani pemulihan (recovery) usai mengalami pendarahan otak dan dirawat di RSUP	Neutral	neutral
4	Disaat seluruh bangsa mendoakan kesembuhan, pendukung ganjar justru kompak menyerang Cak Nun.. Hati mereka dipenuhi kebencian!!	positive	positive

Table 2. Results of Inappropriate Classification

No	Text Data	Label	Result
1	Sama, saya dukung Ganjar, tp mendoakan Cak Nun sembuh. Trus Om Eko bilang semua pendukung Ganjar? Lebih bahaya Fitnahmu @ekobuy2 sama pemerintah dan Ganjar drpd rasa benci mereka pada Cak Nun.	positive	Negative
2	saya..saya pendukung ganjar....apakah sy menyerang kyai Cak Nun...sungguh terlalu beringas dan kejam dan fitnah pada kami..sekaligus mengadu domba...hai penyebar hoax yg nga ada kapok2nya..mau masuk lg pean..!	positive	Negative
3	Cak Nun menyatakan bahwa Prabowo Subianto bukan sosok penculik seperti yang dituduhkan banyak kalangan. Semoga lekas sembuh Cak Nun. Jemaah dan bangsa Indonesia masih butuh panjenengan.	positive	negative
4	Cak Nun Dirawat di ICU RS Sardjito Yogyakarta KNews.id - Cendekiawan Emha Ainun Nadjib atau Cak Nun dikabarkan sakit hingga tak sadarkan diri. Cak Nun kini sedang dirawat Cak Nun Dirawat di ICU RS Sardjito Yogyakarta Share on Facebook Share on	neutral	negative

The proportion of text documents that were not successfully categorized amounted to 11 out of 74. By evaluating the outcomes of the classification process on

the 74 testing data, it was observed that the Naïve Bayes Multinomial Classifier achieved accurate classification for 63 instances out of the entire set of 74 testing data. Consequently, the Naïve Bayes Classifier demonstrated an accuracy rate of 85,135% in classifying hate speech within text documents.

$$\text{Accuration} = \frac{63}{74} \text{ X } 100\% = 85.135\%$$

After performing the necessary calculations to determine the model's accuracy, an evaluation is conducted to determine the feasibility of applying the model to real-world scenarios. This assessment involves the calculation of the F1-score, which is a metric of classification performance that takes into account both accuracy and precision [11].

CONCLUSION

The study found that the precision of categorizing hate speech against Cak Nun through employing the Multinomial Naive Bayes approach achieved 85,135%. The outcomes of this level of precision suggest that the performance of the Naïve Bayes Multinomial Classifier is considerably commendable in classifying hate speech. The findings demonstrated that the Naive Bayes Multinomial approach could accurately categorize hate speech to a substantial extent. It can be observed from the F1-score, which attains a value of 0.851.

ACKNOWLEDGMENT *(optional)*

With utmost gratitude and appreciation to the Almighty Allah for His countless blessings and invaluable gifts, the researcher accomplished the study entitled "Enhancing Public Access to Information on the Ministry of Religious Affairs of Malang Regency's Website through the Implementation of a Design Thinking Approach." Furthermore, the researcher extends heartfelt thanks to PT. Arion Indonesia, thank you for allowing me to conduct this research and for your invaluable assistance. The researcher also expresses profound gratitude to the Department of Informatics Engineering at the esteemed Islamic State University Maulana Malik Ibrahim Malang for their unwavering support and invaluable contribution to this research. Last, the researcher sincerely appreciates the esteemed mentors and friends, whose names cannot be mentioned individually, for their invaluable guidance and support throughout this endeavor.

REFERENCES

[1] Y. Rohmiyati, "Analisis Penyebaran Informasi Pada Sosial Media", *anuva jurnal kajian budaya perpustakaan dan informasi*, vol. 2, no. 1, p. 29, 2018, https://doi.org/10.14710/anuva.2.1.29-42.

[2] K. Antariksa, Y. Wp, & E. Ernawati, "Klasifikasi Ujaran Kebencian Pada Cuitan Dalam Bahasa Indonesia", *jurnal buana informatika*, vol. 10, no. 2, p. 164, 2019, https://doi.org/10.24002/jbi.v10i2.2451.

[3] O. Rahman, G. Abdillah, & A. Komarudin, "Klasifikasi Ujaran Kebencian Pada Media Sosial Twitter Menggunakan Support Vector Machine", *jurnal resti (rekayasa sistem dan teknologi informasi)*, vol. 5, no. 1, p. 17-23, 2021, https://doi.org/10.29207/resti.v5i1.2700.

[4] I. Huda, "Implementasi Natural Language Processing (NLP) untuk Aplikasi Pencarian Lokasi", *Sekolah Vokasi Universitas Gadjah Mada*, Vol. 3, No. 2, 2019.

[5] E. Ningrum and A. Widodo, "Implementasi Metode Multinomial Naïve Bayes Classifier Untuk Analisis Sentimen", *journal of fundamental mathematics and applications (jfma)*, vol. 1, no. 2, p. 128, 2018, https://doi.org/10.14710/jfma.v1i2.18.

[6] A. Rahman, Wiranto, & A. Doewes, "Online News Classification Using Multinomial Naive Bayes." *Jurnal Ilmiah Teknologi dan Informasi*, Vol. 6, No. 1, June 2017.

[7] B. Hakim, "Analisa Sentimen Data Text Pre-processing Pada Data Mining Dengan Menggunakan Machine Learning", *Jbase - Journal of business and audit information systems*, vol. 4, no. 2, 2021, https://doi.org/10.30813/jbase.v4i2.3000.

[8] A. Rinandyaswara, Y. A. Sari, & M. T. Furqon, "Pembentukan Daftar Stopword Menggunakan Term Based Random Sampling pada Analisis Sentimen dengan Metode Naïve Bayes (Studi Kasus: Kuliah Daring di Masa Pandemi)." *Jurnal Teknologi Informasi dan Ilmu Komputer*, Vol. 9, No. 4, Agustus 2022.

[9] L. F. Narulita, "Analisa Sentimen Pada Tinjauan Buku Dengan Algoritma K-Nearest Neighbour", *Konvergensi*, vol. 13, no. 2, 2019, https://doi.org/10.30996/konv.v13i2.2758.

[10] N. A. Susanti, M. Walid, & Hoiriyah, "Klasifikasi Data Tweet Ujaran Kebencian di Media Sosial Menggunakan Naive Bayes Classifier." *Jurnal Mahasiswa Teknik Informatika*, Vol. 6, No. 2, September 2022.

[11] W. Hidayat, M. Ardiansyah, & A. Setyanto, "Pengaruh Algoritma Adasyn Dan Smote Terhadap Performa Support Vector Machine Pada Ketidakseimbangan Dataset Airbnb", *Edumatic: Jurnal Pendidikan Informatika*, vol. 5, no. 1, p. 11-20, 2021, https://doi.org/10.29408/edumatic.v5i1.3125.