SEXISM SPEECH DETECTION ON THE INSTAGRAM PLATFORM WITH NAÏVE BAYES ALGORITHM

Kamila Narendragharini[1], Endang Wahyu Pamungkas[2]

Informatics Engineering, Faculty Of Communication And Informatics, Muhammadiyah University Of Surakarta

Abstrak

Seksisme di media sosial, khususnya di Instagram, telah menjadi masalah yang merajalela dalam beberapa tahun terakhir. Meskipun kebijakan dan pedoman Instagram melawan ujaran kebencian dan diskriminasi, pengguna terus menghadapi diskriminasi berbasis gender, pelecehan, dan objektifikasi di platform Instagram. Hal ini dapat termanifestasi dalam berbagai aspek, termasuk seksualisasi tubuh perempuan, komentar merendahkan tentang perempuan, dan pemeliharaan stereotip gender. Perempuan, khususnya, sering menjadi sasaran perilaku seksis ini, yang berdampak negatif pada kesehatan mental dan harga diri mereka. Studi ini menggunakan Instagram karena ketersediaan komentar publik untuk pendapat individu, dengan tujuan mengembangkan program yang mampu mendeteksi bahasa seksis dalam bagian komentar Instagram. Para peneliti menggunakan algoritma Naive Bayes dan berbagai jenis klasifikasi, seperti Gaussian, Bernoulli, dan Multinomial, untuk mencapai studi ini. Metode ini secara akurat menentukan apakah kata dalam sebuah kalimat bersifat seksis atau non-seksis. Proses klasifikasi ini dilakukan menggunakan bahasa pemrograman Python. Pada tahap klasifikasi, data yang digunakan berasal dari komentar Instagram dengan total 814, terdiri dari 177 komentar yang dilabeli sebagai seksisme dan 637 komentar yang dilabeli sebagai non-seksisme. Dataset kemudian dibagi menjadi 80% untuk pelatihan dan 20% untuk tujuan pengujian, di mana 651 data sebagai data pelatihan dan 163 sebagai data pengujian. Hasil dari proses pengujian dengan data tanpa preprocessing memiliki akurasi yang lebih baik dan metode multinomial memiliki akurasi tertinggi sebesar 88% menggunakan data tanpa preprocessing.

Kata kunci: seksisme, perempuan, Instagram, naïve bayes, media sosial.

Abstract

Sexism on social media, particularly on Instagram, has become a widespread problem in recent years. Despite Instagram's policies and guidelines against hate speech and discrimination, users continue to face gender-based discrimination, harassment, and objectification on the Instagram platform. This can manifest in a variety of aspects, including the sexualization of women's bodies, derogatory remarks about women, and the perpetuation of gender stereotypes. Women, in particular,are frequently become the targets of these sexist behaviors, which have a negative impact on their mental health and self-esteem. The study used Instagram because of the availability of public comments for individual opinions, with the goal of developing a program capable of detecting sexistlanguage in Instagram comments section. The researchers used the Naive Bayes algorithm and multiple classification types, such as Gaussian, Bernoulli, and Multinomial, to achieve this study. This method accurately determines whether a word in a sentence is sexist or non-sexist. This classification process was carried out using the Python programming language. at classification stage, The data used originates from Instagram comments with total of 814, consisting of 177 comments labeled as sexism and 637 comments labeled as non-sexism. The dataset is then split into 80% for training and 20% for testing purposes, where 651 data as training and 163 as testing data. The results of the testing process with data without preprocessing have better accuracy and the

multinomial method has the highest accuracy of 88% using data without preprocessing.

Keywords: sexism, women, Instagram, naïve bayes, social media.

1. INTRODUCTION

Indonesia Ministry of Communication and Informatics reported from kominfo.go.id, Internet users in Indonesia currently reach 63 million people, 95 percent use the internet to access socialmedia. The high number of social media use by the Indonesian people, makes the risk of spreading negative content as well as provocative messages and hate speech that can cause conflict. The reason for how easily hate speech spreads in the media social is because of the ease of using social media. This is because social media is not one-way but two-way, so it provides opportunities for all people to join the conversation (Fermina et al., 2020).

One of the social media that is commonly used to communicate is Instagram. Since its presence, Instagram has attracted the attention of users from various circles (Syahreza & Tanjung, 2018). The choice of Instagram as a medium for communication of course inseparable from the facilities available on Instagram. By using a cellphone that is connected to the internet,users can download the Instagram application onto their cellphone and are ready to share information with fellow Instagram users.

Anyone with a social media account can express their thoughts and feelings. Despite the fact that they are aware of the consequences, which will be read and commented on by others. However, many account holders ignore other risks, such as hurting others by insulting, spewing hate speech, or stating sexism (Suryani et al., 2021).

Sexism speech on social media platforms like Instagram has become a pervasive issue, leading to online harassment, cyberbullying, and discrimination. The world appears to be compelled to view an event through the eyes of each gender. According to Sara Mills, sexism occurs as a result of gender discrimination. Meanwhile, a woman may face the following forms of sexism: 1) Body anatomy of a catcalling woman, 2) Discrimination against women in the workplace, 3) Female sexual objects, 4) Less respectful of women's ideas. Meanwhile, men were subjected to sexism, such as: 1) men should not be whiny and soft; and 2) men should not use women's goods. 3) No man should be weak (Umami & Zahroh, 2018).

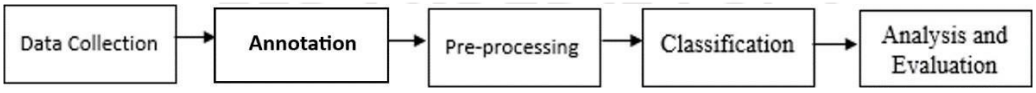
Detecting sexism speech is a challenging problem due to the large volume of data and the complex language used in online conversations. In recent years, natural language processing techniques have been used to address this problem. The classification methods used are quite popular including Decision Tree, Naive Bayes Classifier (NBC), and K-nearest Neighbor (k- NN) (Willianto et al., 2022) also Support Vector Machine (SVM). which is a probabilistic algorithm that can be used for text classification tasks. Based on the results of previous sentiment analysis research

conducted by Athira Luqyana, Cholissodin and Perdana, in this research, sentiment analysis was carried out on Instagram objects by implementing the Maximum Entropy and SVM methods and producing results accuracy value of 90% (Athira Luqyana et al., 2018). In this study, the text of the comment will be analyzed to determine whether it contains or does not contain sexism speech. The text documents used are obtained from Instagram comments which will then be classified using the Naive Bayes method.

The goal of the study "Sexism Speech Detection on Instagram with Naive Bayes Algorithm" is to use the Naive Bayes method to classify sexism speech in Instagram comments. In addition, the performance of the Naive Bayes method in classifying sexism speech text on Instagram social media was evaluated. This study is divided into sections that explain the introduction, background, and research objectives. The second section explains how the author gathered data from several posts on Instagram, which was used as reference material by the author while conducting research. The third section describes the method, which includes an explanation of the system and how it works. The fourth section discusses the results obtained after using the determined methods and systems. The final section describes the conclusion, which includes the expected outcomes and whether they are appropriate and have met the author's goals.

2. METHOD

In this study, the sexism speech classification system on Instagram platform and topic #menoverwomen, #slutshaming, #sexism and so on uses Naive Bayes with a workflow that can be seen in Picture 1.



Picture 1. Method flow

2.1. Data collection

The data collection process involved manually selecting a number of posts from several Instagram accounts, as well as posts with sexism-related hashtags such as #menoverwomen, #slutshaming, #sexism and so on. These posts were then crawled using website crawler called phantom buster to extract the comments, which were then saved in CSV format for further storage and analysis.

2.2. Annotation

The data annotation process is carried out manually by two people with different backgrounds, it is hoped that discussions and collaborations will occur which can improve the quality of the data annotation results. As a result, more accurate and consistent annotation results will be produced when gathering data for further analysis. Furthermore, this manual annotation method helps to reduce

human errors that may occur during the annotation process, resulting in more reliable data. The background of the people annotating the data originates from IT and psychology, so that the annotated data can precisely yield results between sentences of sexism or non-sexism. This illustrates multidisciplinary collaboration and enriches understanding of the complex aspects within the annotated data. Consequently, interpretations of cultural and social contexts can be better understood, enhancing accuracy in identifying relevant sentences.

2.3. Pre-processing

The pre-processing process is a stage for the process of eliminating several problems that can interfere with data processing. It refers to the cleaning and transformation of raw data before it can be used for analysis. Pre-processing involves several techniques to convert the raw data intoa format suitable for analysis because a lot of data has an inconsistent format. However, in certain cases, unprocessed data can yield better results than preprocessed data. This occurs because preprocessing often eliminates many aspects and cleans up the sentence structures, potentially obscuring or altering the original meanings of the sentences in the dataset. For example, important information or subtle nuances in the original data may be lost during the preprocessing process, causing the model to miss capturing the true context or characteristics of the data. Therefore, sometimes using raw data without preprocessing can retain a greater wealth of information and produce more accurate results in specific cases. In this pre- processing process, they were carried out with the following:

2.3.1. Case Folding

Case folding is a pre-processing technique that eliminates capitalizationdifferences by changing all the characters in a text data set to lowercase, uppercase,or a blend of both. This method is frequently used in text analytics and natural language processing (NLP) to standardize the text data and make it more dependable and similar.

2.3.2. Tokenization

Tokenization is a natural language processing (NLP) pre-processing method that divides text data into discrete units, or tokens, such as words, phrases, or symbols.Tokenization is a method used to convert unstructured text data into a structured format that can be used by machine learning algorithms for analysis and processing.

2.3.3. Filtering

Filtering is a data analysis pre-processing method that identifies and removes or modifies data that is irrelevant, noisy, or includes errors in order to improve the quality and accuracy of the data set.

2.3.4. Stemming

Stemming is a natural language processing (NLP) pre-processing method that involves reducing words to their base or root form by removing suffixes and prefixes. The aim of stemming is to reduce the number of words used in text data analysis by treating different spellings of the same word as a

single term.

2.3.5. TF-IDF Weighting

TF-IDF (Term Frequency-Inverse Document Frequency) weighting is a text mining and information retrieval method that measures how frequently a term appears in a document and how uncommon it is in the entire corpus to determine the significance of a term in a document or corpus.

2.4. Classification

This research employs Naive Bayes classification because it performs well in text classification. The Naive Bayes classification is founded on the Bayes theorem for conditionalprobabilities (Salmi & Rustam, n.d.), which is written as follows:

P(A|B) = P(B|A) * P(A) / P(B)

Description:

P(A|B) : is the conditional probability of A given B (i.e., the probability of A occurring given that B has occurred).

P(B|A) : is the conditional probability of B given A (i.e., the probability of B occurring given that A has occurred).

P(A) : is the prior probability of A (i.e., the probability of A occurring beforetaking into account any evidence or data).

P(B) : is the prior probability of B (i.e., the probability of B occurring before taking into account any evidence or data).

This study employs three Naive Bayes algorithms: the Naive Bayes Gaussian, theNaive Bayes Bernoulli, and the Naive Bayes Multinomial (Ismail et al., 2020).

2.4.1 Gaussian Naive Bayes

This algorithm is used for continuous numerical data and implies that the featuresare distributed Gaussian (normally). The algorithm calculates the probability of each feature given the class by estimating the mean and standard deviation of eachfeature for each class.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

2.4.2 Bernoulli Naive Bayes

This method is used for binary data. (i.e., features that take on only two values, such as 0 or 1). As a Bernoulli distribution, it represents the likelihood of each feature given the class. (i.e., a probability distribution over binary events). The multinomial performs even better at larger vocabulary sizes than the multivariateBernoulli model, giving a 27% reduction in error on average over the multivariate Bernoulli model at any vocabulary size (Ismail et al., 2020)

$$P(x_i \mid y) = P(i \mid y)x_i + \big(1 - P(i \mid y)\big)(1 - x_i)$$

2.4.3 Multinomial Naïve Bayes

For discrete count data, this method is used. (i.e., features that take on integer values). It uses a multinomial distribution to describe the likelihood of each feature given the class. (i.e., a probability distribution over multiple discrete events).

The preprocessed text data is used to train the Nave Bayes classification model using the Python

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

programming language. The presence of labeled data is critical because cleaning the data without the labels is difficult (Arboleda, 2019). The classification procedure is divided into several stages, which are outlined below:

2.4.3.1 Presenting

The first step in this method is to create the vector object that will beresponsible for converting the unprocessed text document set into a matrix. The transform function is then used to adapt the vector to the text data in the data frame data set's 'Comments' column, resulting in a numerical representation of the text data. In this process, the fit method is also used to train models on specific datasets. This method trains the model to improve its predictions by taking training data as input and learning from it.

	Unnamed: 0	texts	scores
0	0	Follow for more pics and quotes . Keep followi...	0
1	1	Life with long hair... they want your hair lon...	0
2	2	Bloody men are like buses- You wait for about ...	1
3	3	Maybe it was a book Maybe a musician or band N...	1
4	4	(with .for.insta) ... Today is the second anni...	1

Picture 2. Data Matrix

2.4.3.2 Divide data into training and testing data

This procedure entails splitting the available data into two groups: one for training the model and another for testing its performance. The training set isused to train the model to predict, whereas the test set is used to assess the model's accuracy (Thamrin, 2015).

2.4.3.3 Scanning process

The initial step in processing the data involves scanning all the available datafor further analysis. Additionally, these data transactions can be sorted into multiple categories based on the similarities between the items present in eachtransaction (Gunawan, 2016). This grouping helps to identify

patterns and relationships among the items, which can be used for various data analysis tasks, such as clustering, classification, and association rule mining.

2.5. Analysis and Evaluation

The data used originates from Instagram comments with total of 814, consisting of 177 comments labeled as sexism and 637 comments labeled as non-sexism. The dataset is then split into 80% for training and 20% for testing purposes, where 651 data as training and 163 as testing data. The objective of the test is to compare the accuracy, recall, precision, and F1-score values of three Naive Bayes methods using preprocessed and non-preprocessed data. Table 1 below is a comparison of the data before and after preprocessing.

Table 1. Text Before and After Preprocessing

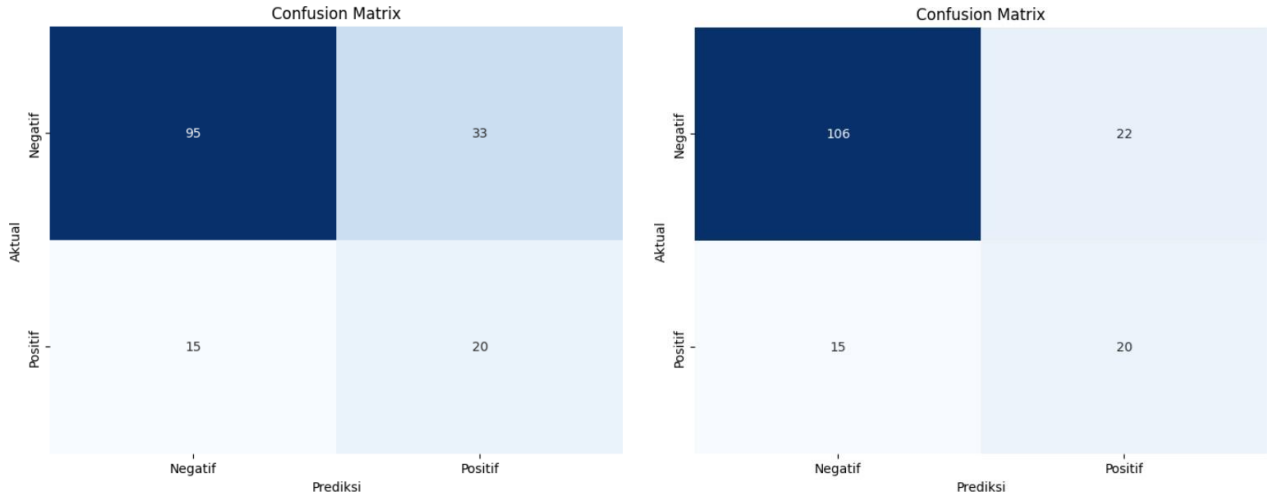
No	Label	Before	After
1	Not Sexism	Some people stride toward a better future. Others have chauffeurs.	people stride toward better future others chauffeur
2	Sexism	Women are too nurturing to be in competitive fields.	woman nurturing competitive field

After the data has gone through the preprocessing stage, the data will then go through the TF-IDF weighting stage. Data that has gone through TF-IDF weighting is then ready to be used for the training process. Table 2 below shows the performance of the gaussian method inclassifying data without preprocessing and data with preprocessing.

Table 2. The Performance of The Gaussian Method

Metrics	***Preprocessed***	***Non-Preprocessed***
Precision	76%	79%
Recall	71%	77%
F1-score	72%	78%
Accuracy	71%	77%

Based on Table 2, it can be seen that the gaussian method has better performance using data without preprocessing. To find out more about how the performance of the gaussian method can be seen in Picture 3 below.



With Preprocessing Without Preprocessing

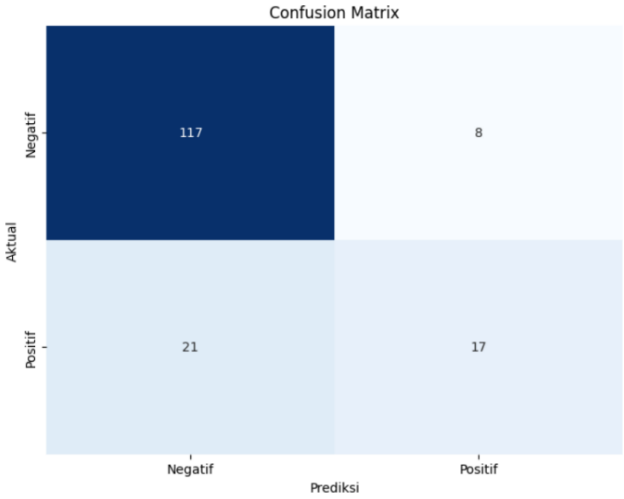
Picture 2. Confusion Matrix generated by Gaussian Method

As can be seen from Picture 2 above, the gaussian method can perform better classification on negative labels using data without preprocessing, while on positive labels, both data using preprocessing or not produce the same performance. For the performance of the bernoulli method on test data can be seen in table 3 below.

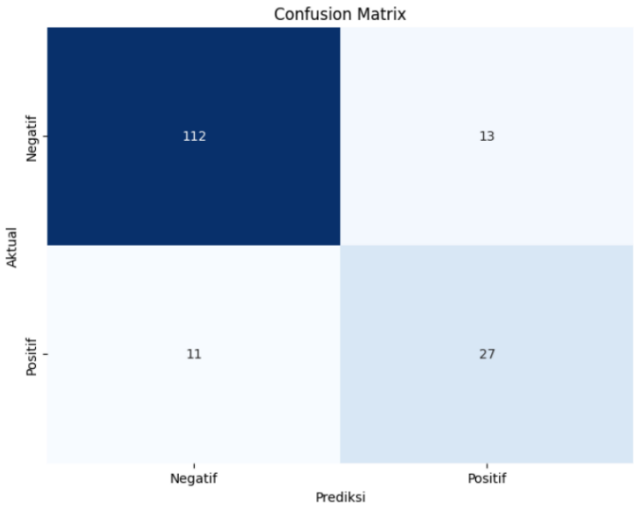
Table 3. The Performance of The Bernoulli Method

Metrics	***Preprocessed***	***Non-Preprocessed***
Precision	81%	86%
Recall	82%	85%
F1-score	81%	85%
Accuracy	82%	85%

Based on Table 3, the bernoulli method has better performance than the gaussian method. However, just like the gaussian method, data without preprocessing produces better performance. To find out more about how the performance of the bernoulli method can be seen in Figure 4 below.



With Preprocessing



Without Preprocessing

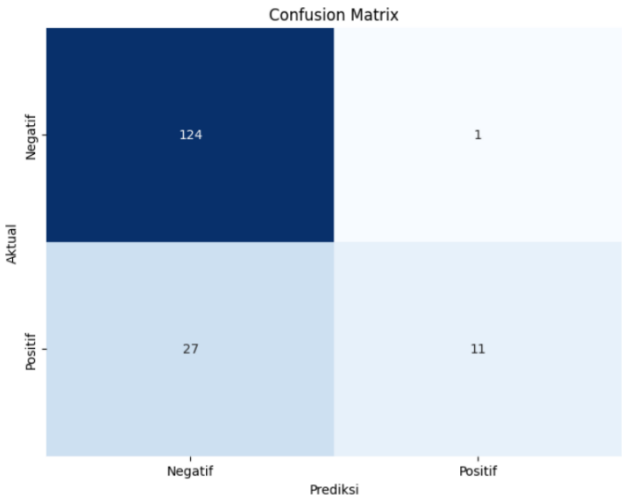
Picture 3. Confusion Matrix generated by Bernoulli Method

As can be seen from Picture 3 above, the performance of the bernoulli method is better on data without preprocessing. On data without preprocessing, the bernoulli method can classify data with positive labels better. For the performance of the multinomial method on test data can be seen in Table 4 below.

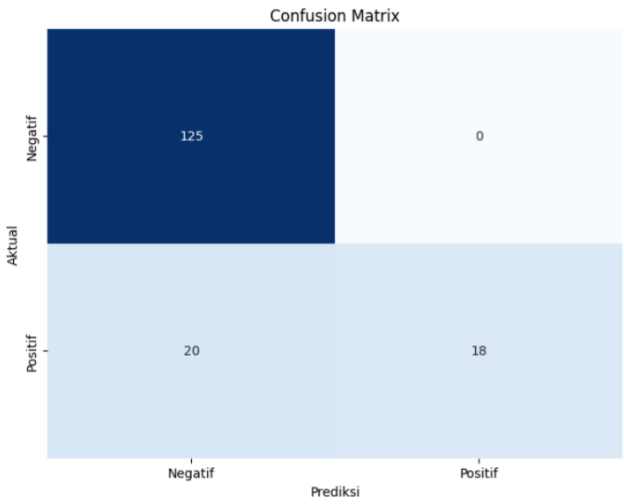
Table 4. The Performance of The Multinomial Method

Metrics	***Preprocessed***	***Non-Preprocessed***
Precision	84%	89%
Recall	83%	88%
F1-score	79%	86%
Accuracy	83%	88%

Table 4 above shows that the multinomial method with data without preprocessingproduces the best accuracy compared to other methods with an accuracy of 89%. The confusion matrix generated by the multinomial method can be seen in Figure 5 below.



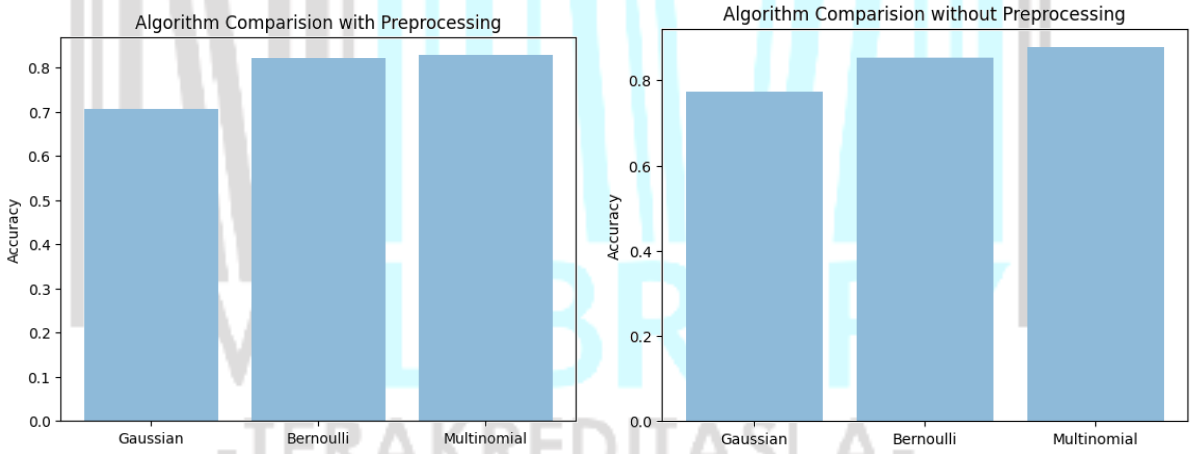
With Preprocessing



Without Preprocessing

Picture 4. Confusion Matrix generated by Multinomial Method

As can be seen in Picture 4 above that the multinomial method produces very good performance for negative labels on data without preprocessing. Multinomial method can classify all negative data correctly. While on positive labels, the multinomial method has notbeen able to classify well.



Picture 5. Algorithm comparison of Gaussian, Bernoulli, Multinomial

As seen in the comparison graph of Gaussian, Bernoulli, and Multinomial Naive Bayes above, with or without preprocessing, the Multinomial method yields the best results. The graph indicates that, although the Gaussian and Bernoulli Naive Bayes methods also show improved performance without preprocessing, the highest accuracy is still observed with the Multinomial Naive Bayes method, both with and without preprocessing. This indicates that the Multinomial Naive Bayes method has better capabilities in handling text data, even without preprocessing steps. This is due to the small data sample for positive or sexism labels.

3. FINDING AND DISCUSSION

Sexism is a discriminatory comment based on gender. Sexism can affect anyone, especially women. Sexism is related to gender streotypes, but is often also found in the form of sexual harassment (Jahan & Oussalah, 2023). Sexism often found on social media platforms like Instagram has become a widespread issue, leading to online harassment, cyberbullying, and discrimination (Plieger et al., 2021). One method for detecting sexist speech is through the use ofmachine learning. Machine learning algorithms are trained to classify sexist speech by utilizing atraining dataset consisting of text and labeled data indicating whether the text contains sexism or not. The trained algorithm is subsequently evaluated for its performance using a test dataset. The Naïve Bayes method is one of the frequently employed approaches in text classification tasks (Okky Ibrohim et al., 2019; Putri et al., 2020; Subramanian et al., 2023).

Naïve Bayes is derived from the principles of Bayes' theorem, commonly used to solve classification problems. It has been noted that Naïve Bayes performs accurately in determining the true polarity of a given sentence, even in imbalanced datasets. Moreover, it is a high-bias and low-variance classifier that performs well even in small datasets. Naïve Bayes comes from two words. Naïve comes from this method which assumes that one occurrence of a particular feature is independent of the occurrence of other features. Thus, each feature contributes individually to he classification without any dependency on the other features (Villavicencio et al., 2021). Research by (Putri et al., 2020) using naïve bayes resulted in the highest accuracy compared to other methods, namely 76% in text classification tasks.

Based on several previous studies that use naïve bayes and produce good performance. This research will apply three naïve bayes methods to detect sexism speech. In this study, the data used is data on Instagram comments that have sexist hashtags. The data that has been collected is labeled as sexist and non-sexist. The data then goes through preprocessing and TF- IDF weighting stages. The data is divided into training data and test data with a ratio of 80% for training data and 20% for test data. This research uses three naïve bayes methods, namely gaussian, bernoulli, and multinomial. The three methods were trained using data with preprocessing and without preprocessing. The multinomial method trained with data without preprocessing showed the highest accuracy with 88% accuracy. In addition, accuracy using the gaussian and bernoulli methods is also better when trained with data without preprocessing. This is because the preprocessing stage carried out automatically using the python programming language is still not perfect. The gaussian, bernoulli, and multinomial models that have been trained with data without preprocessing are then implemented into a website using the python programming language and the flask framework to perform text classification.

4. CONCLUSION

This research successfully implements the naive bayes method to detect sexism speech on social media instagram. This research uses data on Instagram post comments that have sexismhashtags. The data that has been collected is then labeled as sexist and non-sexist. The labeled data is then divided with a ratio of 80% for training and 20% for testing. Furthermore, the data goes through preprocessing and TF-IDF weighting stages. The gaussian, bernoulli, and multinomial methods which are naive bayes algorithms are then trained using data that has been processed and data without preprocessing. The results of the training process with data without preprocessing have better accuracy and the multinomial method has the highest accuracy of 88% using data without preprocessing. This is because the automatic preprocessing stage using python is still not perfect and preprocessing often eliminates many aspects and cleans up the sentence structures, potentially obscuring or altering the original meanings of the sentences in the dataset. Suggestions for further research can check data that has been preprocessed automatically and improve data quality. In this study, the gaussian, bernoulli, and multinomial models were trained with data without preprocessing and then implemented into a web using the python programming language and the flask framework

REFERENCES

Arboleda, E. R. (2019). Comparing Performances of Data Mining Algorithms for Classification of Green Coffee Beans Hybrid Cryptosystem View project Fuzzy Logic and Image Processing View project. *International Journal of Engineering and Advanced Technology (IJEAT)*, *5*, 2249–8958. Https://www.researchgate.net/publication/335104787

Athira Luqyana, W., Cholissodin, I., & Perdana, R. S. (2018). Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. 2(11), 4704–4713. Http://j-ptiik.ub.ac.id

Fermina, D., Waruwu, M., & Vera, N. (2020). Ujaran Kebencian di Media Sosial (Studi Netnografi di Akun Instagram @prof.tjokhowie). AGUNA: Jurnal Ilmu Komunikasi, 1(1), 55–

69. Https://ejournal.amikompurwokerto.ac.id/index.php/AGUNA/article/view/1034

Gunawan, D. (2016). Evaluasi Performa Pemecahan Database dengan Metode Klasifikasi Pada Data Preprocessing Data mining. Khazanah Informatika : Jurnal Ilmu Komputer Dan Informatika, 2(1). Https://doi.org/10.23917/khif.v2i1.1749

Ismail, M., Hassan, N., & Saleh Bafjaish, S. (2020). Comparative Analysis of Naive Bayesian Techniques in Health-Related For Classification Task. Journal of Soft Computing and Data Mining, 1(2), 1–10. Https://doi.org/10.30880/jscdm.2020.01.02.001

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. Neurocomputing, 546, 126232. https://doi.org/10.1016/j.neucom.2023.126232

Okky Ibrohim, M., Sazany, E., & Budi, I. (2019). Identify abusive and offensive language in indonesian twitter using deep learning approach. Journal of Physics: Conference Series, 1196(1). https://doi.org/10.1088/1742-6596/1196/1/012041

Plieger, T., Groote, O., Hensky, R., Hurtenbach, L., Sahler, S., Thönes, L., & Reuter, M. (2021). The Association Between Sexism, Self-Sexualization, and the Evaluation of Sexy Photos on Instagram. Frontiers in Psychology, 12(August). https://doi.org/10.3389/fpsyg.2021.716417

Putri, T. T. A., Sriadhi, S., Sari, R. D., Rahmadani, R., & Hutahaean, H. D. (2020). A comparison of

classification algorithms for hate speech detection. IOP Conference Series: Materials Science and Engineering, 830(3). https://doi.org/10.1088/1757-899X/830/3/032006

Salmi, N., & Rustam, Z. (n.d.). Naïve Bayes Classifier Models for Predicting the Colon Cancer.

Https://doi.org/10.1088/1757-899X/546/5/052068

Subramanian, M., Easwaramoorthy Sathiskumar, V., Deepalakshmi, G., Cho, J., & Manikandan, G. (2023). A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. Alexandria Engineering Journal, 80(May), 110–121. https://doi.org/10.1016/j.aej.2023.08.038

Suryani, Y., Istianingrum, R., Hanik, S. U., Bahasa, P., Indonesia, S., Pgri, U., Tuban, R., & Balikpapan, U. (2021). Linguistik Forensik Ujaran Kebencian terhadap Artis Aurel Hermansyah di Media Sosial Instagram. BELAJAR BAHASA: Jurnal Ilmiah Program Studi Pendidikan Bahasa Dan Sastra Indonesia, 6(1), 107–118. Https://doi.org/10.32528/bb.v6i1.4167

Syahreza, M. F., & Tanjung, I. S. (2018). MOTIF DAN POLA PENGGUNAAN MEDIA SOSIAL INSTAGRAM DI KALANGAN MAHASISWA PROGRAM STUDI PENDIDIKAN EKONOMI

UNIMED. Jurnal Interaksi: Jurnal Ilmu Komunikasi, 2(1), 61–84. Https://jurnal.umsu.ac.id/index.php/interaksi/article/view/1788

Thamrin, H. (2015). Efektivitas Algoritma Semantik dengan Keterkaitan Kata dalam Mengukur Kemiripan Teks Bahasa Indonesia. Khazanah Informatika : Jurnal Ilmu Komputer Dan Informatika, 1(1). Https://doi.org/10.23917/khif.v1i1.1174

Umami, I. M., & Zahroh, L. F. (2018). SEXISM IN DIGITAL COMIC “TAHI LALATS” INSTAGRAM

VERSION, A COMIC BY NUR FADLI MURSYID. Https://www.researchgate.net/publication/336767670

Villavicencio, C., Macrohon, J. J., Inbaraj, X. A., Jeng, J. H., & Hsieh, J. G. (2021). Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes. Information (Switzerland), 12(5). https://doi.org/10.3390/info12050204

Willianto, W., Musdar, I. A., Junaedy, J., & Angriani, H. (2022). Implementasi Teori Naive Bayes dalam Klasifikasi Ujaran Kebencian di Facebook. Jurnal Informatika Universitas Pamulang, 6(4), 666–671. Https://doi.org/10.32493/INFORMATIKA.V6I4.1259