Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 1, 2024, 1-9
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Naive Bayes Classification of Bullying and Non-Bullying Comments in Instagram Social Media Posts

Naflah Faulina

Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Lampung
corresponding author: naflahfaulinaa@gmail.com

Received March 09, 2024; Received in revised form April 25, 2024; Accepted April 27, 2024

Abstract. *Machine learning is a type of artificial intelligence that provides computers with the ability to learn from data. There are three main branches of machine learning, namely supervised machine learning, unsupervised learning, and reinforcement learning. One of the categories in supervised machine learning is classification. Classification is the process of assessing a data object where the object is put into a certain class from the number of classes available. An example of a classification algorithm is Naive Bayes with classification using probability and statistical methods. This algorithm is used to classify Bullying and Non-bullying with a division of training data and testing data, namely 60:40, 70:30, and 80:20 resulting in the best accuracy value for training data and testing data 60:40 of 66%.*
Keywords: *classification; naive bayes algorithm; supervised learning*

INTRODUCTION

Machine learning is a method that can be used to carry out classification and predictions. The algorithm works by studying existing historical data sets so that patterns can be found to predict new data. There are 3 main branches in machine learning, namely supervised machine learning, unsupervised learning, and reinforcement learning. One of the categories in supervised machine learning is classification (Budiharto, 2016; Divyashree Nandini, 2022; Möller, 2023).

Classification is the process of assessing a data object where the object is put into a certain class from the number of classes available. This classification forms a model based on existing training data, then classifies new data using this model. An example of a machine learning classification algorithm is Naive Bayes (Utomo & Mesran, 2020). Naive Bayes is a classification using probability and statistical methods and predicting future opportunities based on the results of previous experience (Peling et al., 2017; Ali et al., 2020). Cases in the real world that are a problem are the large number of abusers of Instagram social media who do not understand the ethics of socializing in cyberspace so that teasing, insults, insults and bullying often occur through comments on Instagram and if this continues it will become an act of bullying (Wu et al., 2008).

Bullying is a negative action carried out by humans continuously or repeatedly (Liu et al., 2022; Alomar & Alabady, 2023). This action often leaves the victim helpless, physically and mentally injured (Mutiara, 2023; Karthikeyan, 2022).). Bullying has an impact that can

Volume 2, No. 1, 2024, 1-9
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

threaten every party involved, including the party being bullied, the party who bullies and the party who witnesses the bullying. Decreased physical and mental health is a bad influence of bullying. In severe cases, bullying can be the cause of fatal actions, such as suicide and others (Chandra, 2020). If bullying detection is carried out conventionally, it will take a long time because it has to detect thousands of comments. So we need an application that can help the public detect bullying acts on a mass scale, namely by classifying comments with existing comment data repositories using the Naive Bayes classification technique. This research aims to obtain the results of comment classification analysis to differentiate comments containing bullying and non-bullying.

METHOD

Data will be collected from Instagram social media posts that are marked as containing bullying and non-bullying content. Data will include comment text as well as class labels (bullying or non-bullying). Data preprocessing steps will be performed to clean and prepare the data for analysis, including punctuation removal, text normalization, and stopword removal. Features will be extracted from comment text using text representation approaches such as TF-IDF (Term Frequency-Inverse Document Frequency) or word frequency counting.

The data will be divided into two subsets: training data and testing data with certain proportions, for example 80% for training and 20% for testing. The Naive Bayes classification model will be trained using the training data. Naive Bayes is a classification method that utilizes Bayes' theorem with the assumption of independence between features.

Here is the general form of Bayes' theorem (Rrmoku et al., 2022):

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \tag{1}$$

with,

$P(A|B)$: the probability of A occurring based on condition B (posteriori prob).
$P(A)$: the probability that A will occur (prior prob).
$P(B|A)$: the probability of B occurring based on the conditions in hypothesis A.
$P(B)$: probability of B occurring.

In the Naive Bayes classification process, several guidelines are needed that can be used to determine the appropriate class for the sample being analyzed so that the Naive Bayes method can be adjusted as follows:

$$(Y|X_1, X_2, \dots, X_n) = \frac{P(Y).P(X_1, X_2, \dots, X_n|Y)}{P(X_1, X_2, \dots, X_n)}$$

$$Posterior = \frac{Prior \times Likelihood}{Evidence}$$

Confusion matrix, also known as error matrix, is a specific table layout and allows visualization of algorithm performance, which is usually typical of supervised learning (in unsupervised learning it is usually called a matching matrix) (Saputro & Sari, 2020).

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 1, 2024, 1-9
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Table 1. *Confusion matrix*

Actual Class	Prediction Class	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

with,

1. True Positive (TP) is data that is predicted to be positive and the actual data is positive.
2. True Negative (TN) is data that is predicted to be negative and the actual data is negative.
3. False Positive (FP) is data that is predicted to be positive and the actual data is negative.
4. False Negative (FN) is data that is predicted to be negative and the actual data is positive.

Performance classification can be evaluated by paying attention to the following measures (Coronado et al., 2022):

with

accuration

$$\text{accuration} = \frac{TP + TN}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%$$

precision

$$\text{precision} = \frac{TP}{TP + FP} \times 100\%$$

recall

$$\text{recall} = \frac{TP}{TP + FN} \times 100\%$$

F1-Score

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 1, 2024, 1-9
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Research Stages

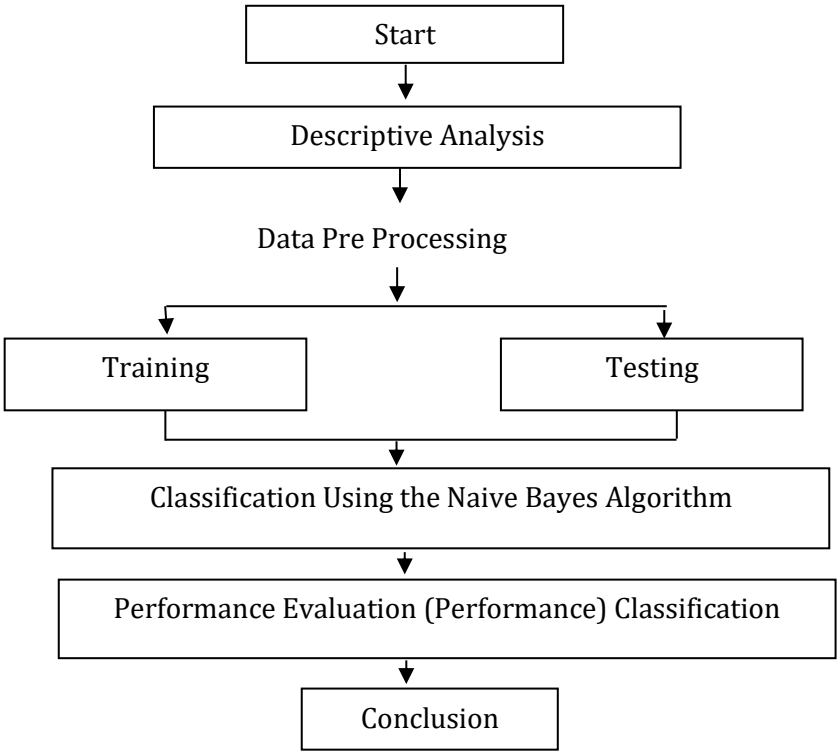

Figure 1. Research Flow Diagram

RESULT AND DISCUSSION

Descriptive Analysis

An overview of the research data used in classifying bullying and non-bullying comments using the Naive Bayes algorithm is data from 620 Instagram comments taken https://www.kaggle.com/code/dhealarasati/analysis-sentimen. This dataset includes comment text with a label indicating whether the comment is undesirable. Undesirable comments are labeled bullying, while desirable comments are labeled non-bullying. The following is a histogram diagram of bullying and non-bullying comments.

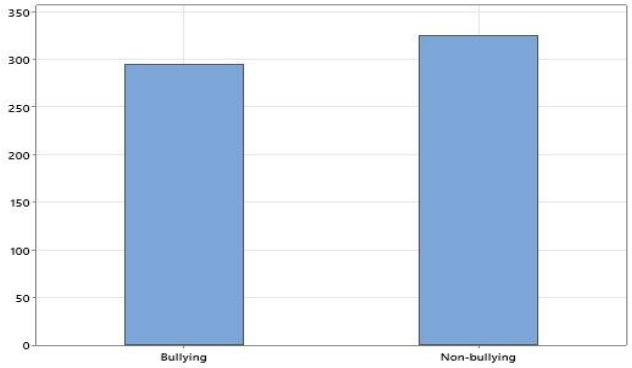

Figure 2. Histogram diagram of bullying and non-bullying comments

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 1, 2024, 1-9
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Figure 2 shows that there were 295 bullying comments and 325 non-bullying comments. From the research data, a classification analysis of bullying and non-bullying comments was carried out using the Naive Bayes algorithm with the help of R software. The stages carried out were as follows. The first step to build a classifier requires data transformation, namely changing the label variable from a categorical variable to a factor so that value 1 is non-bullying and value 2 is bullying. Data as follows:

Table 2. Data Transformation

Label	Text	Data Transformation
Non-bullying	"Hai kak Isyana aku ngefans banget sama kak Isyana.aku paling suka lagu kak Isyana itu lagu tetap didalam jiwa"	1
Non-bullying	"Manusia apa bidadari sih HERANN deh cantik terus.."	1
Bullying	"Sebelum aku Unfoll, aku mau bilang NAJISSSS"	2
Bullying	"Mba bisa gk sih, kalau GK merem melek gitu matanya kaya orang cacingan"	2
:	:	:
Non-bullying	"Cantik banget kalo meldi muka natural kek gini"	1

The next stage of data pre processing aims to prepare data that will later be used in the classification process. There are steps to be taken, namely: Changing all text comments to lowercase, removing numbers, removing punctuation marks and spaces. The results of pre processing data are as follows:

Table 3. Hasil data *pre processing*

Label	Text	Data Transformation
Non-bullying	hai kak isyana aku ngefans banget sama kak isyana.aku paling suka lagu kak isyana itu lagu tetap di dalam jiwa	1
Non-bullying	manusia apa bidadari sih heran cantik terus	1
Bullying	sebelum aku unfoll aku mau bilang najis	2
Bullying	mba bisa gk sih kalau gk merem melek gitu matanya kaya orang cacingan	2
:	:	:
Non-bullying	cantik banget kalo meldi muka natural kek gini	1

From Table 3, the results of the pre-processing data contain no more capital letters, numbers and punctuation marks. All words are also separate and form basic words. This

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 1, 2024, 1-9
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

can see words that often appear in bullying and non-bullying comments. It can be seen in the following image.

(a) Bullying

(b) Non-bullying

Figure 3. Visualization of words that often appear

Data Sharing

The next stage is dividing training and testing data in a ratio of 60:40 and 70:30 which is shown in Table 4.

Table 4. Split Dat.

NO	Data	Amount
1	Training 60 %	372
	Testing 40%	248
2	Training 70 %	434
	Testing 30%	186
3	Training 80 %	496
	Testing 20%	124

Classification Using the Naive Bayes Algorithm

In building the model there are 3 stages, namely calculating the prior opportunity or probability, then calculating the opportunity or probability of each variable for each word and finally calculating the product of the prior probability and variable probability for each word.

Table 5. Previous odds in each category

Comment	Total	Probability
Bullying	227	0,6102151
Non-bullying	145	0,3897849

Calculate the likelihood probability of each term from all documents. Total number of words. The test data classification process is to multiply all the probability values. Higher values constitute a new class of data.

```
Pada data uji
semoga lancar terus usaha nya meldi

semoga
data_raw_train$type       NO       Yes
     Bullying    1.00000000 0.00000000
     Non-bullying 0.95172414 0.04827586
lancar
```

Volume 2, No. 1, 2024, 1-9
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

```
data_raw_train$type       NO       Yes
     Bullying     0.97356828 0.02643172
     Non-bullying 0.98620690 0.01379310
terus
data_raw_train$type       NO       Yes
     Bullying     1.00000000 0.00000000
     Non-bullying 0.95172414 0.04827586
usaha
data_raw_train$type       NO       Yes
     Bullying     0.969162996 0.030837004
     Non-bullying 0.993103448 0.0068965
       52
nya
data_raw_train$type       NO       Yes
     Bullying     0.91189427 0.08810573
     Non-bullying 0.95172414 0.04827586
meldi
data_raw_train$type       NO       Yes
     Bullying     0.98237885 0.01762115
     Non-bullying 0.89655172 0.10344828
```

Figure 4. R Naïve Bayes output

So these comments are classified into the non-bullying class.

Performance Evaluation (Performance) Classification

The results of the classification performance evaluation on the comparison of training and testing data with 60:40, 70:30, and 80:20 are explained in the following table 6.

Table 6. *Confusion matrix* 60:40

Actual Class	Prediction Class	
	Bullying	*Non-bullying*
Bullying	60	77
Non-bullying	8	103

From table 6, the Naive Bayes method produces a value of TP or comments that are correctly predicted as bullying comments of 60, FP or comments that are actually considered bullying but predicted to be non-bullying is 77, FN or comments that are non-bullying but predicted to be bullying are 8, and TN or comments that were correctly predicted as non-bullying comments were 103.

Table 7. *Confusion matrix* 70:30

Actual Class	Prediction Class	
	Bullying	*Non-bullying*
Bullying	16	79
Non-bullying	4	87

From table 8, the Naive Bayes method produces a value of TP or comments that are correctly predicted as bullying comments of 16, FP or comments that actually have the value of bullying but are predicted to be non-bullying of 79, FN or comments that are considered non-bullying but are predicted to be bullying of 4, and TN or comments that were correctly predicted as non-bullying comments were 87. So the precision, recall, f1-score and accuracy values can be seen as follows.

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 1, 2024, 1-9
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Table 8. *Confusion matrix* 80:20

Actual Class	Prediction Class	
	Bullying	*Non-bullying*
Bullying	2	51
Non-bullying	1	70

From table 8, the Naive Bayes method produces a TP value or comments that are correctly predicted as bullying comments of 2, FP or comments that are actually considered bullying but predicted to be non-bullying are 51, FN or comments that are non-bullying but predicted to be bullying are 1, and TN or correct comments predicted as non-bullying comments were 70.

Table 9. Comparison of accuracy, precision, recall and f1-score values

Data Comparison	Acuration	*Precision,*	*Recall*	*F1-Score*
60:40	**66%**	**44%**	**88%**	**59%**
70:30	56%	16%	80%	26,7%
80:20	58%	3,7%	66,7%	7%

Based on table 9, it shows that the results of the 60:40 data comparison have greater accuracy, precision, recall and f1-score values than other data.

CONCLUSION

Based on the results of research on the classification of bullying and non-bullying comments in Instagram social media posts using the Naive Bayes method, with a 60:40 distribution of training and testing data, good results were obtained and a high level of accuracy was seen in the confusion matrix test showing that the accuracy value was 66%, 44% precision value, 88% recall value, and 59% F1-Score value.

REFERENCE

Ali, Z. M., Hassoon, N. H., Ahmed, W. S., & Abed, H. N. (2020). The application of data mining for predicting academic performance using k-means clustering and naïve bayes classification. *International Journal of Psychosocial Rehabilitation*, *24*(03), 2143-2151.

Alomar, A. M., & Alabady, H. S. (2023). The Phenomenon of Cyber Bullying: Interpretation, Confrontation, and the Position of Islamic Law. *Journal of Namibian Studies: History Politics Culture*, *34*, 746-768.

Budiharto, W. (2016). Machine learning & computational intelligence. ANDI, Yogyakarta.

Candra, R. M., & Rozana, A. N. (2020). Klasifikasi Komentar Bullying pada Instagram Menggunakan Metode K-Nearest Neighbor. *IT Journal Research and Development*, *5*(1), 45-52.

Coronado, E., Kiyokawa, T., Ricardez, G. A. G., Ramirez-Alpizar, I. G., Venture, G., & Yamanobe, N. (2022). Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an industry 5.0. *Journal of Manufacturing Systems*, *63*, 392-410.

Divyashree, N., & Nandini Prasad, K. S. (2022). Algorithms: Supervised machine learning types and their application domains. In *Proceedings of Second International Conference on Sustainable Expert Systems: ICSES 2021* (pp. 787-807). Singapore: Springer Nature Singapore.

Karthikeyan, C. (2022). Conceptualizing Causative Factors of Workplace Cyberbullying on Working Women. In *Research Anthology on Changing Dynamics of Diversity and Safety in the Workforce* (pp. 164-184). IGI Global.

Volume 2, No. 1, 2024, 1-9
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Möller, D. P. (2023). Machine Learning and Deep Learning. In *Guide to Cybersecurity in Digital Transformation: Trends, Methods, Technologies, Applications and Best Practices* (pp. 347-384). Cham: Springer Nature Switzerland.

Mutiara, Y. (2023). The Importance of Character Education in School Bullying Cases. *Journal of Childhood Development*, *3*(2), 130-139.

Liu, W., Qiu, G., & Zhang, S. (2022). Situational action theory and school bullying: Rethinking the moral filter. *Crime & Delinquency*, *68*(4), 572-593.

Peling, I. B. A., Arnawan, I. N., Arthawan, I. P. A., & Janardana, I. G. N. (2017). Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm. *Int. J. Eng. Emerg. Technol*, *2*(1), 53.

Rrmoku, K., Selimi, B., & Ahmedi, L. (2022). Application of trust in recommender systems—utilizing naive bayes classifier. *Computation*, *10*(1), 6.

Saputro, I. W., & Sari, B. W. (2020). Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Creative Information Technology Journal*, *6*(1), 1-11.

Utomo, D. P., & Mesran, M. (2020). Analisis komparasi metode klasifikasi data mining dan reduksi atribut pada data set penyakit jantung. Jurnal Media Informatika Budidarma, 4(2), 437-444.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14, 1-37.