MATICS : Jurnal Ilmu Komputer dan Teknologi Informasi ISSN : 1978-161X(p); 2477-2550(e)
Volume 17, No. 2(2017), pp 52-62 DOI : 10.18860/mat.v17i2.29866
Received : August 07th 2025; **Accepted** : August 28th 2025 ; **Avalaible Online** : September 29th 2025

Classification Cyber Harassment on Twitter using Multinomial Naïve Bayes

Kiki Oktavia, Ria Dhea L.N. Karisma, and Heni Widayani

***Abstract*— The study uses the Multinomial Naïve Bayes (MNB) method to classify four types of cyber harassment on Twitter, namely Physical Threats, Purposeful Embarrassment, Racist, and Sexual Harassment. A total of 2,000 Indonesian language tweets were used as samples and have been manually labeled using training data and testing data. The classification results show that the model achieves an accuracy of 77%, with a consistent accuracy value of 76.21% based on the K-fold cross-validation test. This study shows that MNB is effective in multiclass text classification to detect cyber harassment and provides a computationally efficient solution to the real-time content moderation system.**

***Index Terms*— Cyber Harassment, Multinomial Naïve Bayes, Twitter, Purposeful Embarrassment, Physical Threats, Racist, Sexual Harassment**

I. Introduction

MULTINOMIAL Naïve Bayes is a method that uses probability theory in statistics that can also solve multiclass classification problems on big data in machine learning.

Machine learning is a discipline in computer science that uses algorithms and statistical models to analyze and draw conclusions from patterns in data. Multiclass classification belongs to supervised learning, which will be learned using training data or labels as a set of examples to be used to classify the data [1]. Multinomial Naïve Bayes is a method that works optimally in the case of multiclass classification but has a constraining factor in the classification process due to the "zero frequency" in the probability calculation. Furthermore, it calculates the probability of occurrence of each word by multiplying the class prior probability by the likelihood value of the occurrence of each word in each class in the text data classification problem [2].

The freedom of using social media that cannot be controlled becomes a gap used to commit abusive behavior or cybercrime, especially in Indonesia. Cyber Harassment is a type of cybercrime defined as intentional and repeated behavior utilizing technology such as mobile phones, e-mail, websites, and others. It aims to harm or humiliate people. The Cyber Crime Directorate of the Indonesia National Police, through Patrolisiber, revealed that there were 15,152 reports regarding cybercrime in Indonesia throughout 2021, of which 3,101 were reported on harassment and threats. The Southeast Asia Freedom of Expression Network (SAFEnet) stated a tenfold increase in Cyber Harassment cases in 2020 compared to 2019. Pew Research Center stated that Cyber Harassment is classified into four types of behavior: Physical Threats, Purposeful Embarrassment, Racist, and Sexual Harassment. National Commission on Violence against Women (KOMNAS Perempuan) reported data based on a survey by the Coalition for Safe Public Space (KRPA) in 2021 that Cyber Harassment most commonly happens in social media (42%) and chatting apps (33%). Indonesia ranks fifth with the most people using Twitter worldwide, with 18.54 million users. Prieto [3] stated that one of the social media with the highest number of cybercrime cases is Twitter, in which at least five tweets every minute contain criminal acts.

Akhter [4] classified 1,000 cyberbullying comments on Facebook into three classes: shaming, harassment, and racism. It shows that the Multinomial Naïve Bayes method can classify data with an accuracy value of 88.89%, where the model uses 80% training data and 20% testing data. Pardede [5] conducted similar research, classifying 2.736 cyberbullying comments data from Kaggle into two classes: bullying and non-bullying. The study used a data partition of 80% training and 20% testing data and produced an accuracy value of 80%. El Akbar [6] classified 2,271 data tweets into three categories: non-political hate speech, motivated political speech, and non-hate speech, using an 80% partition of the training data and 20% testing data. This research shows excellent results where the Naïve Bayes method succeeded in classifying data with an accuracy of 94%.

Although some other classification methods are slightly superior in terms of precision, such as SVM and Random Forest in the case of Sexual Harassment, MNB provides a more balanced Performance in the four categories of harassment. In addition, Multinomial

Manuscript received August 07, 2025. This work was supported in part by Informatics Engineering Department of Maulana Malik Ibrahim Islamic State University.

R.D.L.N.K. Author is with the Mathematics Departement of Maulana Malik Ibrahim Islamic State University , Malang, Indonesia (corresponding author provide phone 0812-16002783; email riadhea@uin-malang.ac..id)

K.O. Author is is the Mathematics Departement of Maulana Malik Ibrahim Islamic State University, Malang, Indonesia (corresponding author provide phone 0341-000000; email author@uin-malang.ac..id

Naïve Bayes is much more computationally efficient and less sensitive to parameters than SVM and RVM. Then, Multinomial Naïve Bayes is more suitable for large-scale and real-time text classification. Unlike deep learning methods such as LSTM or BERT, which require large computing resources and extensive labeled datasets. Multinomial Naïve Bayes can provide a simpler and faster solution with competitive accuracy. The study by Akhter et al [4]; Veziroğlu et al [7]; and Shkurti et al [8] showed that Multinomial Naïve Bayes could outperform more complex models, such as the classification of short texts full of noise, such as data from social media Twitter.

The study aims to obtain accurate classification results regarding the types of Cyber Harassment on Twitter using the Multinomial Naïve Bayes method, in which the types of Cyber Harassment will be classified into four classes, namely Physical Threats, Purposeful Embarrassment, Racist, and Sexual Harassment. Data used in this study are Indonesian tweets that represent acts of Cyber Harassment on Twitter.

II. LITERATURE REVIEW

A. Cyber Harassment

Cyber harassment is using any form of digital media, including chatrooms, email, websites, cell phones, and others. It can be used anonymously to humiliate, harass, and threaten victims, which can cause physical and psychological damage. Actions or behaviors that are categorized as cyber harassment are sending threatening or harassing messages, disseminating the victim's data or using it for criminal purposes, using the internet to humiliate people, and uploading inappropriate and non-consensual content on social media [9].

Pew Research Center is a research center in the United States that focuses on social issues, public opinion, and demographic trends. Pew Research Center its research called "The State of Online Harassment" (2021) explained that 41% of Americans experience cyber harassment, which the types of cyber harassment are divided into four classes, namely Physical Threats, Purposeful Embarrassment, Racist, and Sexual Harassment. The results showed that the factors behind cyber harassment were political views (50%), gender (33%), race or ethnicity (29%), religion (19%), and sexual orientation (16%).

B. Physical Threats

Physical threats are actions that cause or impact both physical and mental. Physical threats are acts that cause someone to injure, hurt, or damage objects. Suyanto [10] divides several types of physical threats into categories: hitting, pushing, slapping, threatening, kicking, and so on. Physical threats on social media usually contain words that are a form of physical threats, such as "bunuh" (kill), "pukul" (hit), "tampar" (slap), "jambak" (snatch), and "tendang" (kick) [11].

C. Purposeful Embarrassment

A purposeful embarrassment is an act that is intentionally committed and offends. Purposeful embarrassment on social media is an act of insult or defamation. It can also be referred to as verbal bullying, defined as bullying or humiliation by using inappropriate words to ridicule, mock, and speak harshly [12]. Inappropriate words that are most often used to embarrass someone on social media are "*bodoh*" (stupid), "*asu*" (dog), "*sampah*" (shit), "*pecundang*" (coward), "*jelek*" (ugly), and "idiot" [13].

D. Racist

Racism is an act or attitude that intends to create hatred, violence, and discrimination against individuals or groups based on skin color, nationality, ethnicity, and religion. Racist acts on social media are one of the hate speeches that aim to disrespect each other. Words that are often found for racist acts on social media are "*cina*" (Chinese), "*kafir*" (infidel), "*Islam*," "*muslim*" (Muslim), "*masjid*" (mosque), "*yahudi"* (Jewish), "*hitam*" (black), and "*putih*" (white) [14].

E. Sexual Harassment

Sexual harassment is behavior that gives unusual and unwanted sexual attention. In the legal realm, "sexual harassment" is "the act of forcing a sexual desire or sexually assaulting" [18]. Sexual harassment on social media is found in words or images targeted at victims of various ages. The most common words found in sexual harassment cases are "*perkosa*" (rape), "*sentuh*" (touch), "*raba*" (grope), "cat calling", "*pelecehan seksual*" (sexual harassment), and "*lonte*" (bitch) [15].

F. Multinomial Naïve Bayes (MNB)

Multinomial Naïve Bayes (MNB) is a classification method in Naïve Bayes Classifier (NBC) based on Bayes Theorem and multinomial distribution. Bayes's Theorem explains whether a hypothesis, possibility, or probability can change when new evidence or information is obtained. Bayes' Theorem uses the premise that knowledge of previous probabilities (prior) can influence subsequent probabilities (posterior) [16]. Meanwhile, the Naïve Bayes Classifier (NBC) is a classification method that uses the formula in the Bayes Theorem but adds a high assumption of independence to its calculations.

MNB is a method that calculates the conditional probability of an experiment by multiplying the prior probability by the likelihood of each experiment. In text classification problems, MNB models the distribution of words in the document based on a multinomial distribution. It treats the document as a sequence of words and assumes that each position for each word is independent or does not influence each other, or is also known as the "bag of words assumption" method [17].

In classifying text data, a document D that has been labeled manually is presented as $D = \{W_1, W_2, \ldots, W_n, C\}$, in which the variable or vector W represents a word in document D, and C is the class label of document D. Multinomial Naïve Bayes (MNB) algorithm is known as follows [18]:

$$P(C|D) = P(C)\prod_{i=1}^{n} P(W_i|C) \quad (1)$$

where

$P(C|D)$: Posterior probability of class C given the document D

$P(C)$: Prior probability of class C

$P(W_i|C)$: Probability of the i^{th} word in the document given class C (likelihood)

In Equation (1), variable C represents the class, while variable D represents the document. Therefore, it can be interpreted that $P(W_i|C)$ measures how much the evidence, which is the occurrence of a word W in a document that can determine that C is the right class. The prior probability value $P(C)$ can be calculated using Equation (2) as follows:

$$P(C = c_i) = \frac{n_{c_i}}{N_d} \quad (2)$$

where

c_i : The i^{th} of sub-class C

n_{c_i} : Number of the i^{th} c class

N_d : Total of all classes

To calculate the likelihood value, a multinomial distribution approach is used. Suppose $x_c = \{x_{c_1}, x_{c_2}, \dots, x_{c_n}\}$ is a discrete random variable which is a vector or count the number or frequency of occurrence of the word W that appears with a total of N in the document, then to maximize the likelihood value as follows [19]:

$$P(W_i = x_{c_i}|C = c_i) = \frac{N!}{x_{c_1}!\,x_{c_2}!\dots x_{c_n}!}\theta_{c_1}^{x_{c_1}}\theta_{c_2}^{x_{c_2}}\dots\theta_{c_n}^{x_{c_n}}$$

$$= \frac{N!}{\prod_{i=1}^{n} x_{c_i}!}\prod_{i=1}^{n}\theta_{c_i}^{x_{c_i}} \quad (3)$$

where

W_i : The i^{th} word

x_{c_i} : Number of occurrences of the i^{th} word in class c

C : Class

c_i : The i^{th} of sub-class C

N : Total number of words appearing

θ_{c_i} : Probability of occurrence of the i^{th} word in class c

Given $\sum_{i=1}^{n}\theta_{c_i} = 1$ and $\sum_{i=1}^{n} x_{c_i} = N$.

Parameter estimation in MNB can use the Maximum Likelihood Estimate (MLE) method. The MLE method can be used if it is based on a certain distribution, whereas in MNB, it is based on a multinomial distribution. In order to find the parameter estimation value using MLE, the first step is to define the likelihood function of the probability mass function of the multinomial distribution as follows:

$$L(\theta_{c_i}) = L(\theta_{c_1}, \dots, \theta_{c_n}) =$$
$$C(x)\prod_{i=1}^{n}\theta_{c_i}^{x_{c_i}}, \textit{where } C(x) = \frac{N!}{x_{c_1}!\,x_{c_2}!\dots x_{c_n}!} \quad (4)$$

Then determine the log-likelihood function $\ln L(\theta_{c_i})$ which is as follows:

$$\ell(\theta_{c_1}, \dots, \theta_{c_n}) = \log L(\theta_{c_1}, \dots, \theta_{c_n})$$
$$= \log C(x) + \log\prod_{i=1}^{n}\theta_{c_i}^{x_{c_i}}$$
$$= \ln C(x) + \sum_{i=1}^{n} x_{c_i}\ln\theta_{c_i} \quad (5)$$

Given $\sum_{i=1}^{n}\theta_{c_i} = 1$ and $\sum_{i=1}^{n} x_{c_i} = N$, then $\theta_{c_n} = 1 - \sum_{i=1}^{n-1}\theta_{c_i}$ and $x_{c_n} = N - \sum_{i=1}^{n-1} x_{c_i}$ then:

$$\ell(\theta_{c_1}, \dots, \theta_{c_{n-1}}) = \ln L(\theta_{c_1}, \dots, \theta_{c_{n-1}})$$
$$= x_{c_1}\ln\theta_{c_1} + \dots + x_{c_i}\ln\theta_{c_i} + \dots + x_{c_{n-1}}\ln\theta_{c_{n-1}} + x_{c_n}\ln\left(1 - \sum_{i=1}^{n-1}\theta_{c_i}\right) \quad (6)$$

To maximize the likelihood function, then the differential equation of $\ln L(\theta_{c_i})$ is as follow:

$$\frac{\partial\ell}{\partial\theta_c} = \frac{\partial\ell(\theta_{c_1}, \dots, \theta_{c_{n-1}})}{\partial\theta_{c_i}} = \frac{x_{c_i}}{\theta_{c_i}} - \frac{x_{cn}}{\theta_{cn}} = 0 \quad (7)$$

Where, $i = 1, \dots, n-1$

Then,

$$\frac{x_{c_i}}{\theta_{c_i}} = \frac{x_{cn}}{\theta_{cn}} \quad (8)$$

Therefore, equation (8) satisfies the equation:

$$\frac{x_{c1}}{\theta_{c1}} = \frac{x_{cn}}{\theta_{cn}}, \frac{x_{c2}}{\theta_{c2}} = \frac{x_{cn}}{\theta_{cn}}, \dots, \frac{x_{cn-1}}{\theta_{cn-1}} = \frac{x_{cn}}{\theta_{cn}} \quad (9)$$

In which equation (9) is equivalent to:

$$\frac{x_{c1}}{x_{cn}} = \frac{\theta_{c1}}{\theta_{cn}}, \frac{x_{c2}}{x_{cn}} = \frac{\theta_{c2}}{\theta_{cn}}, \dots, \frac{x_{cn-1}}{x_{cn}} = \frac{\theta_{cn-1}}{\theta_{cn}} \quad (10)$$

Using the information that

$\sum_{i=1}^{n}\theta_{c_i} = 1 = \sum_{i=1}^{n}\frac{x_{c_i}}{x_{cn}}\theta_{cn} = \frac{\theta_{cn}}{x_{cn}}\sum_{i=1}^{n} x_{c_i} = N\frac{\theta_{cn}}{x_{cn}}$, *then* the MLE values for $\theta_{c_1}, \dots, \theta_{c_n}$ are as follows:

$$\hat{\theta}_{c_i} = \frac{x_{c_1}}{N}, \hat{\theta}_{c_2} = \frac{x_{c_2}}{N}, \dots, \hat{\theta}_{c_n} = \frac{x_{cn}}{N} \quad (11)$$

Then, we obtain the parameter estimation value of θ_{c_i} using MLE method is:

$$\hat{\theta}_{c_i} = \frac{x_{c_i}}{N} \quad (12)$$

Furthermore, to overcome the zero probability, a simple way is performed by using the additive smoothing technique. Additive smoothing (or Laplace smoothing) is used for smoothing or overcoming zero probability values. Additive smoothing in Naïve Bayes generally adds value $\alpha = 1$ to the numerator and α_d which represents the number of total words (vocabulary) in the text data to the denominator [18]. Them to calculate the likelihood value $P(W_i|C)$ in MNB using MLE and smoothing techniques, we can use Equation (13)

$$\hat{\theta}_{c_i} = \frac{x_{c_i} + \alpha}{N + \alpha_d} \quad (13)$$

where

x_{c_i} : Number of occurrences of the i^{th} word in class c

N : Total number of words appearing

α : Add-one smoothing for each word in class c

α_d : Total of unique words (vocabulary)

G. *Calculate Accuracy and Error Values*

The study calculates the accuracy and error values using the confusion matrix table. Confusion matrix is a table with a combination of values: predicted and actual values. These values provide the ability to calculate and determine the accuracy of the classification results. The confusion matrix for multiclass classification is shown in Table 1.

Table 1. Confusion Matrix Multiclass Classification

	Actual class				
	Class	PT	PE	R	SH
Predicted class	PT	T_{11}	F_{12}	F_{13}	F_{14}
	PE	F_{21}	T_{22}	F_{23}	F_{24}
	R	F_{31}	F_{32}	T_{33}	F_{34}
	SH	F_{41}	F_{42}	F_{43}	T_{44}

then,
PT is Physical Threats
PE is Purposeful Embarrassment
R is Racist
SH is Sexual Harassment

Table 1 shows the confusion matrix for multiclass classification with T_{mn} = T_{11} representing True Class in m = 1st row and n = 1st column and F_{mn} = F_{12} representing True Class in m = 1st row and n = 2nd column. To calculate the accuracy value, which is the ratio of the classes predicted correctly by the model out of all classes, we can use the following equation [20]:

$$Accuracy = \frac{\text{The number of correct predictions}}{\text{The total number of predictions}} \tag{14}$$

Meanwhile, the error value or Apparent Error Rate (APER) which is the ratio of the comparison of the classes predicted incorrectly by the model of the entire class, which can be formulated as follows [35]:

$$APER = \frac{\text{The number of false predictions}}{\text{The total number of predictions}} \tag{15}$$

III. RESEARCH METHODOLOGY

The data used in this study is collected using API (Application Programming Interface) provided by Twitter. Using data crawling techniques, the Twitter API allows users to download tweets based on predefined keywords. The sample in this study is 2,000 Indonesian tweets collected from January to June 2023. The types of cyber harassment, which consist of Physical Threats, Purposeful Embarrassment, Racist, and Sexual Harassment, are the dependent variables. At the same time, the words collected in the tweet data are the independent variables.

Twitter social media does not provide demographic attributes such as gender, age, or user location in metadata accessed through the Twitter API. Therefore, this study does not explicitly include user demographic information. Potential bias can still occur because data collection is based on keywords and typical user behaviour on a particular platform. The selection of keywords can indirectly direct to certain demographic groups, such as young people and urban languages. Therefore, the limitations in this case can be acknowledged, and for further research, it is recommended to use metadata enrichment techniques or conduct survey-based labeling to evaluate and reduce bias.

In this study used training and testing 95% and 5% respectively. To evaluate the model's performance in this study using K-fold cross-validation and calculating the accuracy and error values using APER (Apparent Error Rate). Classifying tweet data into four types of Cyber Harassment has seven stages.

A. Data Collection Using Data Crawling

Data crawling collects data on websites or social media by downloading data from the database. It automatically collects data based on keywords specified by the user. While collecting data, researchers conducted a literature review to obtain the keywords representing Cyber Harassment. The keywords that have been obtained are adjusted by including synonyms of each keyword because people who post or tweet on Twitter more commonly use words in their daily conversations. The keywords used for each class or type of Cyber Harassment are different, as shown in Table 2. The process is repeated to obtain relevant tweets, which also influences the selection of classification algorithms. Since the data collected is limited, Multinomial Naive Bayes is used because it is more effective in conditions with limited data. Thus, the method is more efficient in computation and has been proven reliable for text classification tasks [21].

Twitter API collects the tweets. It used data crawling techniques by accessing the Twitter Developer page, which is required to download data sourced from Twitter. Data crawling is executed using R programming by entering predefined keywords and access codes obtained after registering on the Twitter Developer page. The number of tweets successfully downloaded in the data collection process is 8,500, with general tweets including news, advertisements, and neutral opinions. Tweet data related to cyber harassment cases on Twitter social media were collected from January to June 2023. Table 2 shows the number of tweet data with manual labeling results.

Table 2. Number of Tweets Cyber Harassment

Number	Class	Number of tweets
1	*Physical Threats*	500
2	*Purposeful Embarrassment*	500
3	*Racist*	500
4	*Sexual Harassment*	500
	Total	2.000

The dataset used in this study consists of 2000 Indonesian tweets and has been categorized in a balanced manner in the type of cyber harassment. Consequently, there was no need to apply data balancing techniques such as oversampling or under sampling. Balanced data conditions allow model performance evaluation to be carried out proportionally to all categories without the dominance of one particular class during the training and validation process.

The attributes obtained in the raw data include date, username, and tweet:

Table 1. Keywords for Data Collection

Number	Class	Keywords

1	Physical Threats	“bunuh”, “pukul”, “mukul”, “tampar”, “nampar”, “jambak”, “tenandg”, and “nenadang”. ("kill", "hit", "punch", "slap", " smack", "snatch", "kick", and "kicking")
2	Purpuseful Embarrassment	“bodoh”, “bego”, “tolol”, “goblok”, “asu”, “anjing”, “sampah”, “pecundang”, “pengecut”, “jelek”, and “idiot”. ("stupid", "dumb", "moron", "dog", "scum", "coward", "loser", "ugly", and "idiot")
3	Racist	“islam”, “cina”, “muslim”, “masjid”, “yahudi”, “hitam”, “putih”, and “kafir”. ("islam", "chinese", "moeslim", "mosque", "jewish", "black", "white", and "kafir")
4	Sexual Harassment	“perkosa”, “sentuh”, “raba”, “cat calling”, “pelecehan”, “leceh”, “jalang”, “lonte”, and “pelacur”. ("rape", "touch", "grope", "cat calling", "harassment", "harass", "slut", "bitch", and "whore")

B. Filtering and Labelling Data

The repetition in the data crawling process is due to the limited number of tweets that can be downloaded. The tweets downloaded are unfiltered, containing many tweets of advertisements, news, and neutral opinions that cannot be categorized as Cyber Harassment. Data filtering and labeling are done manually to select and categorize tweets identified as Cyber Harassment. By filtering and labeling the data, we finally obtained 2,000 tweets, consisting of 500 tweets for each class of type of Cyber Harassment. Tweets that are Cyber Harassment will be labeled based on the types of Cyber Harassment, while tweets that do not qualify as Cyber Harassment will be labeled as "neutral" and excluded from the research data.

C. Data Pre-processing

Data pre-processing is one of the important steps in machine learning that aims to filter or clean data to avoid inconsistent or imperfect data-. Then, it improves data classification accuracy. Data pre-processing is used to remove irrelevant and redundant features that might cause confusion or affect the accuracy of a method. Data pre-processing has six stages, which are case folding (homogenizing letters into lowercase), data cleansing (removing links, emoticons, usernames, and other unnecessary attributes), tokenization (converting sentences into separate words), normalization (homogenizing variations of words that have the same meaning), stop-words removal (remove unnecessary conjunction, affixes), and stemming (convert each word into a standardized word by Indonesian dictionary).

D. Feature Extraction

Feature extraction is a process to convert unstructured data (in the form of text) into structured data (numerical data). Bag of Words (BOW) is one of the techniques that can be used to extract features. BOW is defined as a model that can learn vocabulary in documents, which will then count the number of occurrences of each word. BOW represents text or a collection of words in a bag without regard to word order.

E. Classification using Multinomial Naïve Bayes

The following are the steps in the Multinomial Naïve Bayes algorithm:

1) Calculate the prior probability, which is the probability of each class by equation (2);
2) Calculating the likelihood, which is the probability of each feature (word) in each class by equation (13);
3) Calculate the posterior probability by multiplying the prior probability with the likelihood by equation (1);
4) Find the highest posterior probability value to classify the correct class; and

Calculate the accuracy and error (APER) values of the classification results using the confusion matrix.

IV. RESULT AND DISCUSSION

Multinomial Naïve Bayes analyzed the tweets data after the data screening process. Data screening consists of data crawling, which aims to collect data based on the keywords in Table 2; filtering and labeling data, which aims to select tweets that are considered Cyber Harassment; and data pre-processing, which aims to clean data from noise, duplication, and to eliminate unnecessary attributes in order to improve features and make it easier for the model to classify data in the right class. Furthermore, the 2,000 tweets collected have 5,728 unique words shown in Table 3.

Table 2. Unique Words in 2,000 Tweets

Number	Unique Words
1	abadi
2	abai
3	abal
4	abang
5	abdi
6	abuse
7	abused
8	abusive
9	acara
10	acting
⋮	⋮
5728	zuckerberg

Table 3 contains 5,728 words with the word that has the highest frequency of occurrence is "*lonte*" (bitch) appeared 415 times and the word that has the lowest frequency of occurrence is "*muslim*" (moeslim) appeared 96 times. It shows in Fig.1.

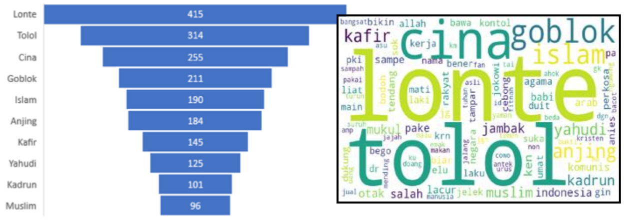

Fig. 1. Ten Words with the Highest Frequency of Occurrence in Cyber Harassment and Wordcloud of Cyber Harassment

Furthermore, in the Physical Threats class, the word that has the highest frequency of occurrence is "*jambak*" (snatch) appears 79 times, and the word that has the lowest frequency of occurrence is "*kesal*" (pissed off) appears 12 times. It is shown in Fig. 2.

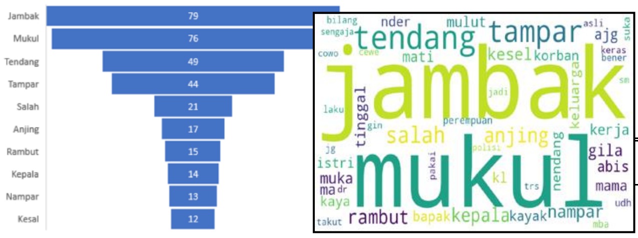

Fig. 2. Ten Words and Worldcloud with the Highest Frequency of Occurrence in Physical Threats Class

In the Purposeful Embarrassment class, the word that has the highest frequency of occurrence is "*tolol*" (stupid) appears 372 times, and the word that has the lowest frequency of occurrence is "*babi*" (pig) appears 19 times. It is shown in Fig. 3.

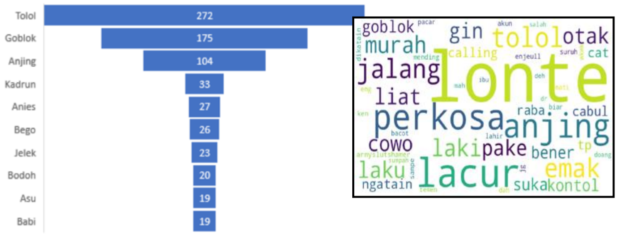

Fig. 3. Ten Words and Worldcloud with the Highest Frequency of Occurrence in Purposeful Embarrassment Class

In the Racist class, the word that has the highest frequency of occurrence is "*cina*" (chinese) appears 249 times and the word that has the lowest frequency of occurrence is "*agama*" (religion) appears 42 times. It shows in Fig. 4.

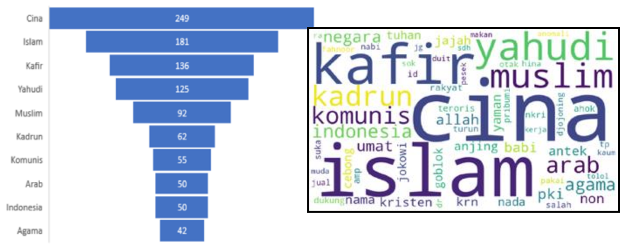

Fig. 4. Ten Words and Worldcloud with the Highest Frequency of Occurrence in Racist Class

In the Sexual Harassment class, the word that has the highest frequency of occurrence is "*lonte*" (bitch) appears 414 times and the word that has the lowest frequency of occurrence is "*laku*" (sold) appears 17 times. It shows in Fig. 5.

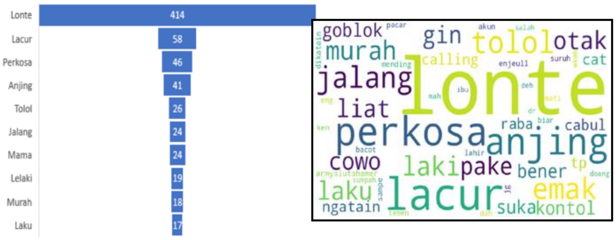

Fig. 5. Ten Words and Worldcloud with the Highest Frequency of Occurrence in Sexual Harassment Class

The next stage is feature extraction using the Bag of Words (BOW) model, which aims to convert text data into numerical data by counting the occurrence of each word. Table 4 shows the feature extraction results.

Table 3. Results of Feature Extraction

	anak	…	cina	…	muslim	…	perkosa	…	rame	…	zina
	3	…	0	…	0	…	0	…	0	…	0
2	0	…	0	…	0	…	0	…	0	…	0
3	0	…	1	…	0	…	0	…	0	…	0
4	0	…	0	…	0	…	0	…	2	…	0
5	0	…	1	…	0	…	0	…	0	…	0
…	…	…	…	…	…	…	…	…	…	…	…
2000	0	…	0	…	0	…	1	…	0	…	0

A. Classification using Multinomial Naïve Bayes

The combination of training data and testing data is done first. Furthermore, the data that has been divided can be used to classify with the Multinomial Naïve Bayes method. Table 5 displays five documents to be used as examples in the classification process with the Multinomial Naïve Bayes model, where the model will classify the right class for the 5th document in the testing data.

Table 5. Sample Dataset for Classification

	Doc.	Word	Class	Class combination
Data Trainin g	1.	nikah kali main tangan *verbally abusive* doyan nampar anak doyan pentok kepala kaca suka mukul anak ampe ungu suka nyambit pakai *belt* kulit	*Physic al Threats*	*Physical Threats*
	2.	*bye* muka burik bopeng wahai	*Purpos eful Embar rassme*	*Purposeful Embarrassm ent*

		anjing tolol biar bisa dapat kerja mapan	*nt*	
	3.	mana saya tanya bukti lu sebut kadrun budak wni kalau cina sipit banyak bukti	*Racist*	*Racist*
	4.	ibu lama gak pake kali kayak anjing keroyok bagus perkosa rame biar kapok	*Sexual Harass ment*	*Sexual Harassment*
Data Testing	5.	lu adalah keturunan lonte perkosa anjing buduk	*Sexual Harass ment*	?

Determining Prior Probability of Cyber Harassment Classes

Table 5 shows that sample of 5 documents, there is 1 document with the Physical Threats class, 1 document with the Purposeful Embarrassment class, 1 document with the Racist class, and two with the Sexual Harassment class. The prior probability is calculated using equation (2) for each class.

Table 5. Prior Probability of Sample Dataset for Classification

Probability			
Physical Threats	*Purposeful Embarrassment*	*Racist*	*Sexual Harass ment*
	$\frac{1}{5} = 0{,}2$	$\frac{1}{5} = 0{,}2$	$\frac{2}{5} = 0{,}4$

Table 6 shows the prior probability values for each class in the types of Cyber Harassment based on Table 4. It presents the results of the prior probability calculations for each class. Subsequently, the likelihood is calculated for the independent variable (x), representing each word in the document for each class (y), using Equation (13)

Table 6. Dataset Sample for Classification Based on Likelihood in each Cyber Harassment Class

	Doc.	Word	Number of word (N)
Data Training	1.	nikah kali main tangan *verbally abusive* doyan nampar anak doyan pentok kepala kaca suka mukul anak ampe ungu suka nyambit pakai *belt* kulit	23
	2.	*bye* muka burik bopeng wahai anjing tolol biar bisa dapat kerja mapan	12
	3.	mana saya tanya bukti lu sebut kadrun budak wni kalau cina sipit banyak bukti	14
	4.	ibu lama gak pake kali kayak anjing keroyok bagus perkosa rame biar kapok	13
Data Testing	5.	lu adalah keturunan lonte perkosa anjing buduk	7

Table 6 illustrates the number of word present in each document (N). Document 1 contains 23 words, Document 2 contains 12 words, Document 3 contains 14 words, Document 4 contains 13 words, and Document 5 contains 7 words. Based on the information, the total number of unique words or vocabulary (α_d) across all documents can be determined as shown in Table 7.

Table 7. Number of unique words (vocabulary)

List of word						number of unique words or vocabulary (α_d)
abusive	adalah	ampe	anak	anjing	bagus	59
banyak	*belt*	biar	bisa	bopeng	budak	
buduk	bukti	burik	*bye*	cina	dapat	
doyan	gak	ibu	kaca	kadrun	kalau	
kali	kapok	kayak	kepala	kerja	keroyok	
keturunan	kulit	lama	lonte	lu	main	
mana	mapan	muka	mukul	nampar	nikah	
nyambit	pakai	pake	pentok	perkosa	rame	
saya	sebut	sipit	suka	tangan	tanya	
tolol	ungu	*verbally*	wahai	wni	tolol	

Table 7 shows a list of words or vocabulary, representing the occurrence of distinct words across all document (Document 1-5), totalling 59 words. Having determined the number of word occurrences in each document (N) and the total number of distinct words across all document (α_d), then the feature extraction value for the 5th document calculated. Table 8 shows the feature extraction result of the 5th document.

Table 8. Feature Extraction in 5th Document

Word in 5th Document

Doc.	*lu*	*adalah*	*keturunan*	*lonte*	*perkosa*	*anjing*	*buduk*
1.	0	0	0	0	0	0	0
2.	0	0	0	0	0	1	0
3.	1	0	0	0	0	0	0
4.	0	0	0	0	1	1	0
5.	0	0	0	0	0	0	0

Table 8 shows that the word "*lu*" (*you*) appears once in the 3rd document, the word "*perkosa*" (rapped) appears once in the 4th document, and the word "*anjing*" (dog) appears once in both the 2nd and 4th documents. Based on this information, the likelihood value for each word in Document 5 can be calculated for each class using Equation (13). Table 9 shows the calculations for the likelihood value of each word in Physical Threats and Purposeful Embarrassment classes. Similarly, the calculations are performed for the Racist and Sexual Harassment classes, respectively.

Table 9. Likelihood computation $P(W_i|C)$

Physical Threats	*Purposeful Embarrassment*
$\boldsymbol{P(\text{"lu"}\|PT)}$	$P(\text{"lu"}\|PE)$
$= \frac{\mathbf{0+1}}{\mathbf{23+59}} = \mathbf{0{,}0121}$	$= \frac{0+1}{12+59} = 0{,}0140$
$\boldsymbol{P(\text{"adalah"}\|PT)}$	$P(\text{"adalah"}\|PE)$
$= \frac{\mathbf{0+1}}{\mathbf{23+59}} = \mathbf{0{,}0121}$	$= \frac{0+1}{12+59} = 0{,}0140$
$\boldsymbol{P(\text{"keturunan"}\|PT)}$	$P(\text{"keturunan"}\|PE)$
$= \frac{\mathbf{0+1}}{\mathbf{23+59}} = \mathbf{0{,}0121}$	$= \frac{0+1}{12+59} = 0{,}0140$
$\boldsymbol{P(\text{"lonte"}\|PT)}$	$P(\text{"lonte"}\|PE)$
$= \frac{\mathbf{0+1}}{\mathbf{23+59}} = \mathbf{0{,}0121}$	$= \frac{0+1}{12+59} = 0{,}0140$
$\boldsymbol{P(\text{"perkosa"}\|PT)}$	$P(\text{"perkosa"}\|PE)$
$= \frac{\mathbf{0+1}}{\mathbf{23+59}} = \mathbf{0{,}0121}$	$= \frac{0+1}{12+59} = 0{,}0140$
$\boldsymbol{P(\text{"anjing"}\|PT)}$	$P(\text{"anjing"}\|PE)$
$= \frac{\mathbf{0+1}}{\mathbf{23+59}} = \mathbf{0{,}0121}$	$= \frac{1+1}{12+59} = 0{,}0281$
$\boldsymbol{P(\text{"buduk"}\|PT)}$	$P(\text{"buduk"}\|PE)$
$= \frac{\mathbf{0+1}}{\mathbf{23+59}} = \mathbf{0{,}0121}$	$= \frac{0+1}{12+59} = 0{,}0140$

Table 9 presents that words like "*lu,*" "*adalah,*" "*keturunan,*" "*lonte,*" "*perkosa,*" and "*buduk*" have zero occurences in both Physical Threats and Purposeful Embarrassment. However, due to Laplace smoothing, they are assigned by non-zero probability as,

$$P(W|PT) = \frac{\mathbf{0+1}}{\mathbf{23+59}} = \mathbf{0,0121} \tag{16}$$

$$P(W|PE) = \frac{0+1}{12+59} = 0{,}0140 \tag{17}$$

Equation (16) and (17) indicate that the likelihood of these words appearing is slightly higher in Purposeful Embarrassment or PE compared to Physical Threats or PT. Since, the denominator which is *total words+distinct words* is smaller for Purposeful Embarrassment. The likelihood value for words are generally higher in the Purposeful Embarrassment than Physical Threats class due to smaller total word that count in Purposeful Embarrassment (12 words versus 23 words in Physical Threats). It highlights impact of class-specific word distribution on likelihood calculation. The role of Laplace smoothing is the addition of 1 in the numerator prevents likelihood values from becoming zero for words that do not appear in the respective class. Thus, approach is particularly useful for addressing unseen word during the classification process. The higher likelihood value in each class suggest that these words are marginally more others class due to its smaller vocabulary size and word count.

The next step is calculated the conditional or posterior probability in each independent variable (*x*) for each class (*y*) by multiplying the prior and likelihood values in each class (Table 10). It is determined by Equation (3).

Table10. Conditional Probability or The Highest Posterior

Conditional Probability			
Physical Threats	*Purposeful Embarrassment*	*Racist*	***Sexual Harassment***
$7{,}5949 \times 10^{-15}$	$4{,}3787 \times 10^{-14}$	$3{,}4548 \times 10^{-14}$	$\mathbf{1{,}5360 \times 10^{-11}}$

Based on Table 10, the conditional probability for Sexual Harassment exactly for fifth document is $1{,}5360 \times 10^{-11}$. It has the highest probability than others. However, the fifth document classified as Sexual Harassment class.

B. *Classification Result*

Multinomial Naïve Bayes determines the appropriate class for each document following the steps outlined in the previous section. In this study, the classification of Cyber Harassment on social media platform Twitter conducted using MNB method with the assistance of Phyton. Table 11 shows the results of the classification for identify the appropriate in Cyber Harassment based on Twitter data.

Cyber Harassment

Physical Threats	Purposeful Embarrassment	Racist	Sexual Harassment
27	19	31	23

Table 11 presents that the classification results of Cyber Harassment tweet using 5% (100 instance) of testing data. Based on the results, 27 documents were classified as Physical Treats class, 19 documents to the Purposeful Embarrassment class, 31 documents to the Racist class, and 23 documents to the Sexual Harassment class. It indicates that the Racist class has the highest number of classified document (31), while the Purposeful Embarrassment class has the lowest (19).

C. Classification Accuracy

The classification accuracy can be known by calculating the classification results on the confusion matrix, which consists of True Class and False Class, shown in Table 12.

Table 12. Confusion Matrix

		Predicted clas			
	Class	PT	PE	R	SH
Actual class	PT	20	4	1	0
	PE	5	10	5	0
	R	1	3	25	1
	SH	1	2	0	22

Table 12 shows that there are 77 out of 100 tweets classified correctly, which consist of 20 tweets classified correctly as Physical Threats class, 10 tweets as Purposeful Embarrassment class, 25 tweets classified correctly as Racist class, and 22 tweets classified correctly as Sexual Harassment class. The other 23 data were classified in the wrong class, which consisted of 7 tweets incorrectly predicted as Physical Threats class, 9 tweets incorrectly classified as Purposeful Embarrassment class, six tweets classified incorrectly as Racist class, and one tweet classified incorrectly as Sexual Harassment class. Based on the results, we know that the accuracy value is 77% and the error value is 23%. The calculation is shown as follows:

$$\text{Accuracy} = \frac{20+10+25+22}{20+4+1+0+5+10+5+0+1+3+25+1+1+2+0+22} \times 100\% = 77\%$$

$$\text{APER} = \frac{4+1+0+5+5+0+1+3+1+1+2+0}{20+4+1+0+5+10+5+0+1+3+25+1+1+2+0+22} \times 100\% = 23\%$$

Furthermore, Table 12 illustrates that the model performs relatively well in classifying the Racist and Sexual Harassment categories the Racist (25 tweets) and Sexual Harassment (22 tweets) categories. The evident are high number of true positives and low false positive and false negative classes respectively. In contrast, the model performance on the Purposeful Embarrassment class is relatively low because the model the model often fails to recognize tweets from the correct class and misclassifies them into that category. The Physical Threats class, consisting of 7 tweets that were misclassified as that class. It shows imperfections in classification, although not as severe as in Purposeful Embarrassment (consisting of 9 tweets classified as that class). Thus, although the model accuracy is quite good overall, there is an imbalance in performance between classes that needs to be considered for further model improvement, especially in distinguishing classes that have similar characteristics, such as Purposeful Embarrassment and other classes.

Some types of harassment, such as Purposeful Embarrassment, are more difficult to classify due to the tendency of general, contextual language, and lack of explicit keywords such as Sexual Harassment or Racist. Words such as "*bodoh*" (stupid) or "*sampah*" (shit) can be used in various contexts, often confusing the model and leading to high false positives and false negatives. [13] and [4] found that text-based classification models are more accurate in recognizing explicit hate speech but have difficulty in capturing emotional context or implicit sarcasm.

D. The Performance Model Test using K-Fold Cross Validation

In this study, use K-Fold Cross Validation to test performance of the model. The method allows all data used in the study to be selected randomly and utilized as either training or testing data. K-Fold in this study is $k = 10$, meaning the data is divided by 10 folds.

Table 13. K-Fold Cross Validation Result

Fold	Accuracy (%)
1	75,09
2	74,82
3	75,62
4	75,94
5	75,94
6	75,62
7	**76,21***
8	75,89
9	76,16
10	75,94

* the highest accuracy performance

Table 13 shows that the accuracy value based on each fold. It represents that variation across fold ranging from 75,09% to 76,21%. Thus, variation arises because K-Fold Cross Validation randomly selects by data for use as training and testing sets. Fi=or instance, when fold 1 use as testing data, the remaining $k - 1$ folds serve as training data. The process repeats for each fold. Based on Table 13, the highest accuracy value is 7th fold, with an accuracy of 76,21%. Additionally, the table presents that the accuracy values are consistent and exhibit minimal variation across fold. It suggests that the model and data that used in the study are reliable.

E. Model Implementation in Social Media Moderation System

In addition to producing relatively high accuracy, the classification model based on Multinomial Naive Bayes used in this study has great potential to be integrated into an automatic moderation system in social media, especially Twitter. The accuracy in model is 77%, the model can act as an initial step in detecting and filtering content that is in real-time, before being further reviewed by human moderators. This integration is relevant given the high volume of user uploads that cannot be selected manually in their entirety. Thus, this model is expected to automatically identify and mark tweets containing Physical Threats, Purposeful Embarrassment, Racist, or Sexual Harassment. It is expected to increase the efficiency and speed of responses to reporting problematic content. In addition, this model is computationally lightweight and does not require large amounts of training data so that it will be appropriate to be applied in systems with limited resources, including social media, e-commerce and text-based community applications.

V. CONCLUSSION

The Multinomial Naïve Bayes method in classifying types of Cyber Harassment using text data in tweets on Twitter had the best classification results in the partition of 95% training data and 5% testing data. The classification results show that from the 100 tweets in testing data, there are 77 correctly classified data, consisting of 20 tweets correctly classified as Physical Threat class, ten tweets correctly classified as Purposeful Embarrassment class, 25 tweets correctly classified as Racist class, and 22 tweets correctly classified as Sexual Harassment class. The classification accuracy of Cyber Harassment types with the Multinomial Naïve Bayes method is 77%, and the model performance test using K-fold cross-validation obtained an accuracy score of 76.21%, which shows that the Multinomial Naïve Bayes model is a model with good effective performance.

Data collection by crawling, especially using keywords, has the potential for bias. It can show demographics that appear because Twitter users do not reflect the population, especially regarding age, socioeconomics, and geography. In addition, keyword selection causes bias because only certain tweets containing specific terms are collected. Therefore, some forms of Cyber Harassment with synonyms or slang can be missed, and bias in language variations and dialects can also affect classification because models cannot always understand informal language, sarcasm, or expressions that vary between communities. As a result, bias in manual classification is a challenge because labeling can be done subjectively, which causes inconsistency in the Cyber Harassment category. Further research is expected to expand the keyword's scope using data augmentation techniques and apply labeling validation with multiple annotators.

In addition, the model still has difficulty in classifying the Purposeful Embarrassment class, which is characterized by a high number of false positives and false negatives. Therefore, it shows that the model has difficulty in understanding the context of ambiguous, sarcastic, or implicit speech. Thus, in further research, an exploration of approaches such as Long Short-Term Memory (LSTM) can be carried out, which are superior in understanding word order and contextual meaning in text. It is expected to improve accuracy, especially in classes that are more linguistically subtle, such as Indonesian language data with a broader range of meanings.

ACKNOWLEDGMENT

The author would like to thank the financial support for the research of the Lembaga Penelitian dan Pengabdian Kepada Masyarakat (LP2M), UIN Maulana Malik Ibrahim Malang and the Twitter API contributors

REFERENCES

[1] S. Garcia, J. Luengo and F. Herrera, Data Preprocessing in Data Mining, Spain: Springer International Publishing Switzerland, 2015.

[2] J. D. M. Rennie, L. S, J. Teevan and D. R. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifier," *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003),* p. 8, 2003.

[3] R. Prieto, S. Cresci, C. L. M. and B. S. R, "Crime and its fear in social media," *Palgrave Commun,* vol. 6, p. 57, 2020.

[4] A. Akhter, U. K. Acharjee and M. M. A. Polash, "Cyber Bullying Detection and Classification using Multinomial Naïve Bayes and Fuzzy Logic," *I.J. Mathematical Sciences and Computing,* vol. 4, pp. 1-12, 2019.

[5] J. Pardede, Y. Miftahuddin and W. Kahar, "Deteksi Komentar Cyberbullying Pada Media Sosial Berbahasa Inggris Menggunakan Naive Bayes Classification," *Jurnal Informatika,* vol. 7, pp. 46-54, 2020.

[6] R. R. E. Akbar, R. N. Shofa and M. I. Paripurna, "The Implementation of Naïve Bayes Algorithm for Classifying Tweets Containing Hate Speech with Political Motive," pp. 144-148, 2019.

[7] E. V. a. İ. Ö. B. M. Veziroğlu, "Performance Comparison between Naive Bayes and Machine Learning Algorithms for News Classification," *Bayesian Inference - Recent Trends,* vol. 15, no. 1, pp. 1-20 doi: 10.5772/intechopen.1002778., 17 January 2024.

[8] F. K. V. S. a. A. S. L. Shkurti, "Performance comparison of machine learning algorithms for Albanian news articles," *IFAC-PapersOnLine,* vol. 55, no. 39, pp. 292-295 https://doi.org/10.1016/j.ifacol.2022.12.037, 2022.

[9] C. Wolford-Clevenger, H. Zapor, H. Brasfield, J. Febres, J. Elmquist, M. Brem, R. C. Shorey and G. L. Stuart, "An Examination of the Partner Cyber Abuse Questionnaire in a College Student Sample," *American Psychological Association,* pp. 1-23, 2015.

[10] U. Kango, "Bentuk-Bentuk Kekerasan yang Dialami Perempuan," *Jurnal Legalitas,* vol. 2, pp. 13-20, 2009.

[11] U. Karimah and T. Suyanto, "Model Pelayanan Sosial Anak Korban Kekerasan Fisik di Lembaga Perlindungan Anak Jawa Timur," *Kajian Moral dan Kewarganegaraan,* vol. 7, pp. 1053-1067, 2019.

[12] S. R. A. Putri, "Fenomena Verbal Bullying di

Masyarakat Pedawang," 2020.

[13] G. Gorrell, M. E. Bakir, I. Roberts, M. A. Greenwood and K. Bontcheva, "Which politicians receive abuse? Four factors illuminated in the UK general election 2019," *EPJ Data Science,* pp. 1-20, 2020.

[14] V. Ibrahim, J. A. Bakar, N. H. Harun and A. F. Abdulateef, "A Word Cloud based on Hate Speech in an Online Social Media Environtment," *Baghdad Science Journal,* vol. 18, pp. 937-946, 2021.

[15] T. A. M. Putri, U. Enri and B. N. Sari, "Analisis Algoritma Naive Bayes Classifier untuk Klasifikasi Tweet Pelecehan Seksual dengan #MeToo," *Indonesian Journal on Computer and Technology,* vol. 5, pp. 126-135, 2020.

[16] C. Zhai and S. Massung, Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining, 2016.

[17] T. E. G. Reiten, Classification with Multiple Classes using Naive Bayes and Text Generation with a Small Data Set using a Recurrent Neutral Network, Norway: University of Agder, 2017.

[18] C. D. Manning, P. Raghavan and H. Schutze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[19] J. Kamel, Graphical Models for Classification and Time Series, France: Aix-Marseille University, 2017.

[20] A. N. Roifa, Text Mining dengan Metode Naive Bayes Classifier untuk Mengklasifikan Berita Berdasarkan Konten, Surabaya: Institur Teknologi Sepuluh November, 2018.

[21] C. P. G. D. a. X. W. A. McCallum, "Multi-Conditional Learning: Generative/Discriminative Training for Clustering and Classification," *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI),* vol. 1, pp. 1-7 DOI: 10.5555/1597538.1597608, 16 July 2006.

Kiki Oktavia is a graduate with a Bachelor's Degree in Mathematics from the Department of Mathematics at UIN Maulana Malik Ibrahim Malang. She specializes in technology and creative industries, focusing on social media management and data analysis. Her expertise includes text data processing, data crawling, and analytical techniques such as clustering and classification using R, Python, Minitab, and SPSS. She combines technical and creative skills to uncover data insights and enhance digital engagement.

Ria Dhea L.N Karisma is a researcher with expertise in Biostatistics and Machine Learning. She completed a master's in Statistics at Sepuluh Nopember Institute of Technology Surabaya (ITS Surabaya). She is an awardee of Erasmus+ in Tallinn University of Technology, Tallinn, Estonia, 2014 in the Healthcare faculty. Her research interests include biostatistics, social computer science, and application multivariate analysis.

Heni Widayani is a faculty member in the Mathematics Department at Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang. She earned her doctoral degree from the Faculty of Mathematics and Natural Sciences at Institut Teknologi Bandung. Her research interests include applied mathematics and dynamic systems. While Widayani primarily focuses on applied mathematics and dynamic systems, her work intersects with areas relevant to machine learning, particularly in modeling complex systems and analyzing data patterns.