Journal Computer and Technology
Vol. 2, No. 1, Juli 2024, hlm. 01-09
ISSN : 3048-1880 (Media Online)
DOI : https://doi.org/10.69916/comtechno.v2i1.144

CYBER BULLYING SENTIMENT ANALYSIS BASED ON SOCIAL CATEGORIES USING THE CHI-SQUARE TEST

Zulpan Hadi[1*], Emi Suryadi[2], Ardiyallah Akbar[3], Zaenudin[4] , Rudi Muslim[5]

[1,2]Rekayasa Sistem Komputer, Fakultas Teknologi Informasi dan Komunikasi, Universitas Teknologi Mataram, Indonesia
[3]Teknik Komputer, Fakultas Vokasi, Universitas Teknologi Mataram, Indonesia
[4]Komputerisasi Akuntansi, Fakultas Vokasi, Universitas Teknologi Mataram, Indonesia
[5]Sistem Informasi, Fakultas Teknologi Informasi dan Komunikasi, Universitas Teknologi Mataram, Indonesia
Email: [1]zlpnhadi@gmail.com, [2]emisuryadi@gmail.com, [3]ardiyallah_akbar@ymail.com, [4]zen3d.itb@gmail.com, [5]rudimuslim93@gmail.com

SEJARAH ARTIKEL
Diterima: 22.07.2024
Direvisi: 26.07.2024
Diterbitkan: 30.07.2024

ABSTRAK

Penelitian ini mengevaluasi berbagai model pembelajaran mesin dalam mengklasifikasikan sentimen dalam data perundungan siber yang terdiri dari enam kategori: *not_cyberbullying, gender, religion, other_cyberbullying, age,* dan *ethnicity*. Menggunakan pendekatan *Bag of Words* yang dikombinasikan dengan pemilihan fitur *Chi-Square* (1000 fitur), model yang diuji termasuk *SVM*, *Logistic Regression*, *Naïve Bayes*, *KNN*, dan *Random Forest*. Hasil menunjukkan bahwa *SVM* dan *Logistic Regression* mencapai akurasi tertinggi sebesar 83%, menunjukkan efektivitasnya dalam prediksi. *Naïve Bayes* menunjukkan kinerja terendah dengan akurasi hanya 62%, mengindikasikan ketidakcocokan dengan data atau perlu penyesuaian lebih lanjut. *KNN* dan *Random Forest* menunjukkan kinerja yang baik dengan akurasi masing-masing 75% dan 81%, meskipun tidak setinggi *SVM* dan *Logistic Regression*. Pendekatan multi-algoritma ini memberikan wawasan tentang efektivitas dan perilaku setiap model pada karakteristik data yang beragam, yang penting untuk memahami nuansa unik dari setiap kategori perundungan siber. Pemilihan model harus mempertimbangkan akurasi, interpretabilitas, biaya komputasi, dan kecocokan dengan karakteristik masalah tertentu. Penelitian ini bertujuan untuk memperdalam pemahaman tentang perundungan siber guna mendukung strategi mitigasi yang lebih efektif

Kata Kunci: analisis sentimen, *cyber bullying, chi-square, bag of words,* klasifikasi.

ABSTRACT

This research evaluates various machine learning models in classifying sentiment in cyberbullying data across six categories: not_cyberbullying, gender, religion, other_cyberbullying, age, and ethnicity. Using a Bag of Words approach combined with Chi-Square feature selection (1000 features), models tested include SVM, Logistic Regression, Naïve Bayes, KNN, and Random Forest. Results show SVM and Logistic Regression achieving the highest accuracy at 83%, indicating their effectiveness in prediction. Naïve Bayes performed the poorest with 62% accuracy, suggesting a mismatch with the data or need for further tuning. KNN and Random Forest showed good performance with 75% and 81% accuracy respectively, though not as high as SVM and Logistic Regression. This multi-algorithm approach provides insights into each model's effectiveness and behavior on diverse data characteristics, essential for understanding the unique nuances of each cyberbullying category. Model selection should consider accuracy, interpretability, computational cost, and suitability to specific problem characteristics. This research aims to deepen understanding of cyberbullying to support more effective mitigation strategies.

Keywords: *sentiment analysis, cyber bullying, chi-square, bag of words, classification.*

1. INTRODUCTION

Social media is an online platform that allows users to easily participate, share and create content on blogs, social networks, wikis, forums and virtual worlds [1]. The use of social media has changed the way we communicate, share information, and build social relationships in everyday life. According to the latest report from We Are Social and Hootsuite, social media users worldwide have reached 4.76 billion, equivalent to 59.4% of the world's current population. In Indonesia, the number of users reached 60.4% of the country's total population in the same month [2]. Social media is a forum for positive and negative interactions. One negative example of the use of social media is cyberbullying, namely the deliberate sending of electronic text messages to hurt, harass, threatening, or disturbing other users [3]. This phenomenon often has negative impacts on victims, such as loss of self-confidence, stress, and depression. In tragic cases, cyberbullying can even lead to death [4]. Given the serious impact of cyberbullying, it is important to research to understand this phenomenon in depth and find effective solutions to overcome it. Therefore, it is important to investigate the issue of cyberbullying comprehensively.

Research conducted by [4] discusses cyberbullying sentiment analysis on the Instagram platform. The method used involves combining TF-IDF and n-grams as well as feature selection, which is then classified using the Naive Bayes algorithm. The best results show an accuracy of 91.25%.

Research by [5] discusses cyberbullying using the chi-square method and the Multinomial Naive Bayes algorithm and Support Vector Machine. The two algorithms were compared, and the best results were obtained by Naive Bayes with an accuracy of 90.77%.

Research by [6], also discusses cyberbullying using the Support Vector Machine algorithm, and the approach is carried out using the Support Vector Machine kernel. The best results were obtained with the sigmoid kernel, achieving an accuracy of 83.85%.

Research by [7] discusses fake reviews, where the method used to increase accuracy is chi-square. The best result obtained using the chi-square method was an accuracy of 92.19%

Therefore, this study aims to explore the use of the chi-square test in sentiment analysis to detect cyberbullying. In this way, it is hoped that it can make a significant contribution to efforts to prevent and overcome cyberbullying through a more sophisticated and evidence-based data analysis approach.

2. RESEARCH METHODS

Generally, in research, there is a methodology that outlines the steps or flow of the research. In this research, the author presents the method or model used in Figure 1

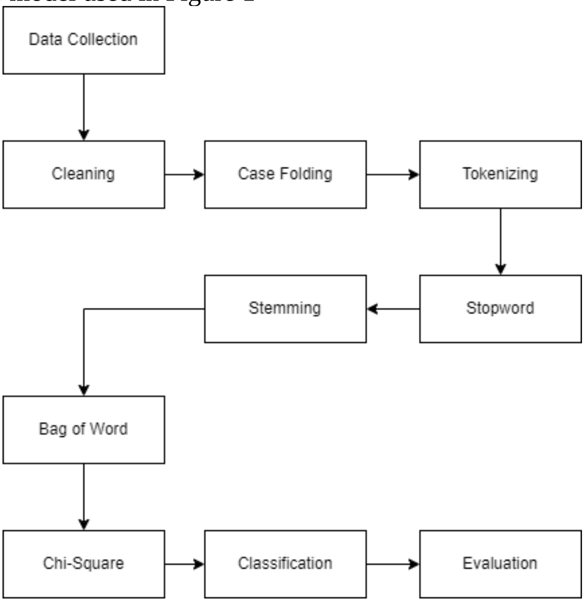


Figure 1. Research Flow

2.1. Data Collection

The data used in this study are secondary data where the data is taken from the kaggle.com website page obtained from previous research conducted by [8]. These data have been collected, processed, and analyzed in the context of initial research, thus providing a strong and relevant basis for further analysis in this study. This dataset contains a classification of cyberbullying based on various factors such as age, ethnicity, gender, and religion.

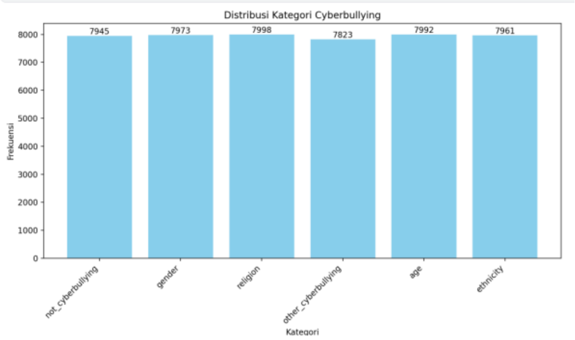


Figure 2. Distribution of Cyberbullying Categories

This bar diagram shows that the frequency of cyberbullying incidents is relatively balanced in various categories, with the following details: not_cyberbullying (7945 incidents), gender (7973 incidents), religion (7998 incidents), other_cyberbullying (7823 incidents), age (7992 incidents), and ethnicity (7961 occurrences). The "religion" category has the highest frequency, while "other_cyberbullying" has the lowest frequency. Overall, the frequency distribution of cyberbullying incidents shows that these incidents are spread almost evenly across various aspects such as religion, age, gender, and ethnicity. This indicates that cyberbullying is a complex problem that touches various aspects of an individual's life and identity. Thus, efforts to address cyberbullying must consider these factors to be effective in reducing and preventing future incidents.

2.2. Preprocessing Data

The data preprocessing process in this research involves several important stages to ensure data quality and consistency before further analysis is carried out.

1. Cleaning Data
2. At this stage, the data that has been collected is cleaned to remove noise such as punctuation, numbers, symbols, and other irrelevant elements. This cleaning process aims to ensure that the data is ready for further processing.
3. Case Folding
   This process changes all text to lowercase. This is done to ensure that uppercase and lowercase words are considered the same. For example, "Data" and "data" will be considered the same after case folding.
4. Tokenizing
   Tokenization is the process of breaking down text into its smallest units called tokens (for example, words or sentences). This token will later be used in further analysis and processing
5. Stopword Removal
   Stopwords are common words that appear frequently in the text but do not have significant meaning for analysis (for example, "and", "or", "but"). This process removes these words to improve the quality of text analysis
6. Stemming
   Stemming is the process of changing words to their basic form. For example, the words "playing", "played", and "plays" will be changed to "play". The goal is to reduce variations in words that are different but have the same meaning.

2.3. Bag of Word

The Bag of Words (BoW) model, abbreviated as BoW, is a model that represents text based on the appearance of various words in a document. BoW consists of two parts: a dictionary containing known words and a measurement of the occurrence of known words. According to [9], this model is called "bag" because it does not pay attention to the order or structure of words in the document. The BoW approach was initially based only on a binary vector representation that only contained the numbers 0 and 1, which is called a one-hot representation. The formula used can be seen in (1) and (2).

$$V = \{w_1, w_2, \dots . w_{lvl}\} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots . (1)$$

$$w = [0, \dots . ,0,1,0, \dots ,0] \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots . . (2)$$

where V is the dictionary (vocabulary) of the words used and w is a one-hot representation where code 1 indicates if the document in question contains words contained in the dictionary and code 0 if the document does not contain words contained in the dictionary. The development of this one-hot representation is by calculating the frequency of occurrence of words in sentences and not just based on codes 1 and 0 as can be seen in equation (3).

$$s = \sum_{k=1}^{l} w_1 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots . (3)$$

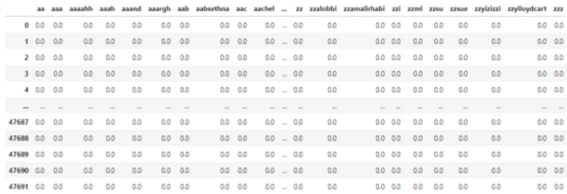
where l denotes the length of sentence s. The sentence representation s is the sum of the one-hot representations of n words in the sentence, that is, each element in s represents the Term Frequency (TF) of the corresponding word. Then Inverse Document Frequency (IDF) is used to measure the importance of words in V in equation (4):

$$idf_{wi} = \log \frac{|D|}{df_{wi}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots . (4)$$

where |D| is the number of all documents in the corpus D and $idfwi$ represents the Document Frequency of wi . At the end, the tf-idf value is calculated using the formula in Equation 5 [10]

$$\hat{S} = s\ x\ idf_{wi} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots .. (5)$$

Research conducted by [11] states that the Bag of Words model is better at classifying than the tf-idf method.



	aa	aaa	aaaahh	aaah	aaand	aaargh	aab	aabsethna	aac	aachel	...	zz	zzalobbi	zzamalirhabi	zzi	zzml	zzsu	zzsue	zzyizizzi	zzylloydcart	zzz
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
47687	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
47688	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
47689	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
47690	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
47691	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 3. Example of Result BOW feature extraction values

2.4. Chi-Square

Chi-square is a statistical technique used to select the most significant features in text. It helps in reducing the dimensionality of the data by selecting only the most relevant features for analysis. Research conducted by [12], used Multinomial Naïve Bayes to classify question sentences, and the Chi-Square algorithm was used for feature selection. The dataset used in this research is a collection of question sentences in Indonesian, consisting of 519 sentences labeled factoids, 491 sentences labeled non-factoids, and 185 sentences labeled other. The test results show that using feature selection with Chi-Square increases classification accuracy by 0.1.

2.5. Classification

At this stage, the processed data is used to train a classification model. This model can be a machine learning algorithm such as Naive Bayes, SVM, Random Forest, or others, which is used to classify text into predetermined categories. Research shows that using machine learning algorithms for text classification can increase accuracy in identifying the correct category.

2.6. Support Vector Machine

Vladimir Vapnik created the Support Vector Machine classification algorithm which can predict classes based on patterns based on the results of machine learning (supervised learning) [13]. The concept of SVM can be simply interpreted as an effort to find the best hyperplane to separate two classes in the input space. By measuring the margin value and finding its maximum point, we can find the best hyperplane [14]

In the SVM algorithm, there is a kernel function that functions to solve non-linear problems to become linearly separable. Several kernel functions used in classification are Polynomial, Linear, Sigmoid, and Radial Basis Function (RBF) kernels. Each kernel trick has its parameter values. The Grid Search method is used to find the best parameter values for the kernel function used.

2.7. Random Forest

The Random Forest algorithm was designed by J. Ross Quinlan, called Random Forest because it is a descendant of the ID3 approach to building decision trees. Random Forest is an algorithm that is suitable for classification problems in machine learning and data mining.

Random Forest maps the attributes of classes so that it can be used to find predictions for data that has not yet appeared. The decision tree itself is a "divide and conquer" approach in studying problems from a set of independent data depicted in a tree chart [15]. A decision tree is also a collection of questions that are arranged systematically, where each question determines a branch based on the attribute value and stops at the leaf of the tree which is a prediction of the variable class [16]

The following are the algorithm stages in making a decision tree using the Random Forest Algorithm:

1. The following are the algorithm stages in making a decision tree using the Random Forest Algorithm.
2. Calculate the value of the information using all existing data with the formula:

$$Info(D) = -\sum_{i=1}^{m} pi \log 2(pi) \dots\dots\dots\dots\dots\dots\dots\dots (6)$$

Where is the probability of a tuple in D becoming a class with assumptions or also called the entropy of D is the average information needed to identify a tuple in D [17].

If value A is a discrete value then data D will be separated by several data values A so that the value of each branch will be pure and similar. After the first branch, the number of possible branches is measured by the equation:

$$InfoA(D) \sum_{j}^{v} \frac{|Dj|}{|D|} xInfoA(Dj) \dots\dots\dots\dots\dots\dots\dots (7)$$

3. Calculating the value of information with formulas.
4. For each attribute, pay attention to the data content of the attribute. where $|Dj|$ $|D|$ is the weight of partition j. InfoA(D) is the information needed to classify tuples from D in partition A. The smaller the result of this equation, the better the resulting partition. The value of an attribute determines whether or not the attribute is important in preparing a decision tree. If the attribute has a continuous value, then the split_point will be searched by sorting all the data according to the attribute from small to large, then averaging one data with the data after it. The information value will be calculated according to the split_point candidates one by one and the smallest split_point value will be selected. (4) The gain value for each attribute will be calculated using formula (7.8), the value with the highest gain will be used as a branch in the decision tree.

$$Gain(A) = info(D) - InfoA(D) \dots\dots\dots\dots\dots\dots (8)$$

5. After the decision tree branch is formed, the calculation is carried out again as in stages 1 to 4. However, if the branch has reached the maximum allowed branches, leaves will be formed with the majority value of the data value.

2.8. KNN

In the book Data Mining Algorithms, Kusrini explains that the K-Nearest Neighbors Algorithm is an approach to finding cases by calculating the proximity between new cases and old cases, which is based on matching the weights of several existing features. [17].

Another view states that KNearest Neighbors is an algorithm for classifying objects based on data that is closest to the object. Data is described in a multidimensional space, where each dimension reflects a feature of the data. The best k value for this algorithm depends on the data. In general, a high k value will reduce the effect of noise on classification, but make the boundaries between each classification more blurred [18]. The main goal of this algorithm is to classify an object based on attributes and training samples. The K-Nearest Neighbor (K-NN) algorithm uses point proximity classification as an estimated value for a new query instance. The following is a way to determine the closeness distance of data in the K-NN method [19]:

$$d = \sqrt{\sum_{i=1}^{n} (a_i - b_1)^2 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (10)}$$

Description:
d= distance
a= test/testing data
b= sample data
i = data variables
n = data dimension

Naïve Bayes

The Naive Bayes classifier is one of the simplest and most commonly used classifiers in the field of machine learning. The Naive Bayes classification model calculates the posterior probability of a class based on the distribution of words in a document. This approach relies on a very simple representation of the document, namely as a Bag of Words. Bag of Words is a text representation method that ignores the order of words in a document and only considers the frequency of occurrence of each word.

The Naive Bayes model works by extracting features from a Bag of Words and using Bayes' Theorem to predict the probability that a given set of features falls under a particular label. Bayes' theorem allows this model to combine information from prior probability and conditional probability to make classification decisions. In the context of Twitter sentiment analysis, bigrams from Twitter data are used as features in Naive Bayes. Bigrams are pairs of two consecutive words in text, which provide richer context information than unigrams. This process classifies tweets into positive, negative, and neutral labels [20].

Naïve Bayes is a machine learning algorithm that uses probability calculations with a Bayes approach. The use of

Bayes' theorem in the Naïve Bayes algorithm involves combining prior probability and conditional probability into a formula that can be used to calculate the probability of each possible classification. The Bayes Theorem formula used is as follows:

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(9)$$

Description:

- $P(H|X)$ is the posterior probability of the hypothesis H provided evidence X.
- $P(H)$ is the prior probability of the hypothesis H, namely the initial probability of the hypothesis before there is evidence.
- $P(X|H)$ is the conditional probability of evidence X given a hypothesis H.
- $P(X)$ is the marginal probability of evidence X, which acts as a scaling factor to ensure that the posterior probabilities add up to 1.

2.9. Logistic Regression

Logistic Regression technique is a statistical method that works based on the relationship between one or more variables. This method is beneficial in various applications, including in classifying sentences as positive or negative, as stated in several studies [21]. Logistic Regression has a dependent variable, denoted as $P(Y)$, with a measurement scale that is nominal or categorical. In this context, the value of $P(Y)$, can only be 0 or 1, representing two possible categories, positive or negative.

The formula for Logistic Regression is:

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k)}} \ldots\ldots\ldots\ldots\ldots..(10)$$

The explanation of equation (10) is that in Logistic Regression calculations, the sigmoid function is used. This sigmoid function has the characteristic that it converts linear input into probabilistic output which is non-linear, producing an output value of $P(Y)$ which can only be binary numbers, namely 0 and 1.

The sigmoid function is used because of its ability to map each input value to a range of values between 0 and 1, making it very suitable for binary classification tasks. The parameters $b_0, b_1, b_2, \ldots., b_k$ in this equation are the regression coefficients estimated from the training data, where $X_1, X_2, \ldots, X_k$ in this equation are the regression coefficients estimated from the training data, where Y.

The Logistic Regression calculation process begins with parameter estimation using the Maximum Likelihood Estimation (MLE) method, which aims to find the parameter values that are most likely to produce the observed data. Once the parameters are estimated, the model can be used to predict the probability of an event, and based on this probability value $P(Y)$ we can carry out classification. If the probability is greater than 0.5, then the output is categorized as 1, while if it is less than 0.5, the output is categorized as 0.

By using Logistic Regression, we can build a model that is not only able to classify data well but also provides an understanding of the relative influence of each predictor variable on the probability of an event. This makes Logistic Regression a very powerful and widely used technique in data analysis, especially in fields that require binary classification such as sentiment analysis, fraud detection, medical diagnosis, and many more.

2.10. Evaluation

To measure the performance of the classification model, we compare the actual values with the predicted values using the Confusion Matrix. Confusion Matrix is an important tool used in classification problems to evaluate model performance, whether on two or more classes. This matrix presents four combinations of predicted and actual values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP represents the number of positive cases correctly predicted, TN is the number of negative cases correctly predicted, FP is the number of negative cases incorrectly predicted as positive, and FN is the number of positive cases incorrectly predicted as negative. By using the Confusion Matrix, we can deeply understand model performance and identify areas for improvement. This matrix also allows the calculation of other performance metrics such as accuracy, precision, recall, and F1-Score, which provides a more comprehensive picture of the model's effectiveness. This evaluation is critical to ensure that the model not only provides accurate predictions but also maintains balance in dealing with different types of prediction errors.

Table 1. Confusion Matrix

	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

For example, in the pregnancy test analogy:

- True Positive (TP): Positive test, indeed pregnant.
- False Positive (FP): Positive test, not pregnant.
- False Negative (FN): Negative test, actually pregnant.
- True Negative (TN): Negative test, indeed not pregnant.

The model will be evaluated using a confusion matrix to determine how well it performs in data classification.

1. Accuracy
   Accuracy is the level of closeness between the predicted value and the actual value. If the accuracy value is high then a system will be better at making predictions, accuracy can be formulated as the following
   $$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \dots\dots\dots\dots.(11)$$
2. Recall
   Recall is a calculation of prediction accuracy that is used as a measure of the system's level of success in retrieving information. Recall can be calculated using a formula as follows:
   $$Recall = \frac{TP}{TP + FN} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots..(12)$$
3. Precision
   Precision is the level of accuracy between the information requested by the user and the precision system answer which can be calculated using the following formula:
   $$Precision = \frac{TP}{TP + FP} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.(13)$$
4. F1-Score
   F1-Score is the harmonic average of Precision and Recall, providing a measure of model performance that takes both metrics into account. F1-Score is very useful in situations where we want to balance between Precision and Recall, especially when there is a class imbalance in the data.
   $$F1 - Score = 2\ X\ \frac{Precision\ X\ Recall}{Precision + Recall} \dots\dots..(14)$$

3. RESULTS AND DISCUSSION

In the introduction to this research, we focus on the complex classification of cyberbullying, where we have identified and divided it into six different categories. We believe a deep understanding of the full range of cyberbullying behavior is key to providing appropriate and effective responses. Therefore, we apply a multi-algorithm approach using five different classification methods. This step allows us to comprehensively explore and compare the strengths and weaknesses of each algorithm, helping us choose the method that best suits our classification goals.

Table 2. Performance Evaluation

Model	accuracy	Recall	Precision	F1-Score
SVM	0.83	0.83	0.84	0.82
Naïve Bayes	0.62	0.62	0.72	0.61
KNN	0.75	0.75	0.79	0.76
Random Forest	0.81	0.81	0.82	0.82
Logistic Regression	0.83	0.83	0.83	0.83

Based on the accuracy results that have been provided for various machine learning models, it can be seen that SVM (Support Vector Machine) and Logistic Regression show excellent performance with the highest accuracy reaching 83%. This shows that these two models can provide accurate predictions in the test scenarios used. On the other hand, the Naïve Bayes model showed the lowest accuracy among all models, reaching only 62%, indicating that this model may need adjustments or may not fit the data used. Meanwhile, KNN (K-Nearest Neighbors) and Random Forest show quite good performance with an accuracy of 75% and 81% respectively. These results indicate that both models are quite effective in overcoming prediction challenges in the given dataset, although not as good as SVM and Logistic Regression in terms of pure accuracy. Selection of an appropriate model does not only depend on accuracy alone, but also considers other factors such as interpretability, computational cost, and suitability to the specific characteristics of the problem at hand.

4. CONCLUSION

Based on the performance evaluation of various machine learning models, it can be concluded that the SVM (Support Vector Machine) and Logistic Regression models show the best performance with the highest accuracy, each reaching 83%. This shows that these two models are very effective in providing accurate predictions in the test scenarios used. On the other hand, the Naïve Bayes model had the lowest performance with an accuracy of only 62%, indicating that it may be a poor fit for the data used or requires further adjustments. The KNN (K-Nearest Neighbors) and Random Forest models show quite good performance with accuracy of 75% and 81% respectively. These results show that these two models are also quite effective in making predictions, although not as good as SVM and Logistic Regression in terms of pure accuracy. In conclusion, the selection of an appropriate model should consider not only accuracy but also other factors such as interpretability, computational cost, and suitability to the specific characteristics of the problem at hand..

REFERENCES

[1] H. Guntoro, D. Rikardo, Amirullah, A. Fahrisani, and I. P. Suarsana, “Analisa Hubungan Kebersihan Cargo Bilges dengan Cargo Hold dalam Mendukung Kelancaran Proses Bongkar Muat,” *E-Journal Mar. Insid.*, vol. 1, no. 2, pp. 1–32, 2022, doi: 10.56943/ejmi.v1i2.9.

[2] Cindy Mutia Annur, “Pertumbuhan Melambat, Jumlah Pengguna Media Sosial Global Capai 4,76 Miliar hingga Awal 2023,” databooks. Accessed: Jun. 10, 2024. [Online]. Available: https://databoks.katadata.co.id/datapublish/2023/02/07/pertumbuhan-melambat-jumlah-pengguna-media-sosial-global-capai-476-miliar-hingga-awal-2023

[3] M. R. Kurniawanda and F. A. T. Tobing, “Analysis Sentiment Cyberbullying In Instagram Comments with XGBoost Method,” *IJNMT (International J. New Media Technol.*, vol. 9, no. 1, pp. 28–34, 2022, doi: 10.31937/ijnmt.v9i1.2670.

[4] Fauzan Baehaqi and N. Cahyono, “Analisis Sentimen Terhadap Cyberbullying Pada Komentar Di Instagram Menggunakan Algoritma Naïve Bayes,” *Indones. J. Comput. Sci.*, vol. 13, no. 1, pp. 1051–1063, 2024, doi: 10.33022/ijcs.v13i1.3301.

[5] S. Riadi, E. Utami, and A. Yaqin, “Comparison of NB and SVM in Sentiment Analysis of Cyberbullying using Feature Selection,” *Sinkron*, vol. 8, no. 4, pp. 2414–2424, 2023, doi: 10.33395/sinkron.v8i4.12629.

[6] S. S. Wijayanti, E. Utami, and A. Yaqin, “Comparison of Kernels on Support Vector Machine (SVM) Methods for Analysis of Cyberbullying,” *2022 6th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng.*, 2023, doi: 10.1109/ICITISEE57756.2022.10057761.

[7] Z. Hadi and A. Sunyoto, “Detecting Fake Reviews Using N-gram Model and Chi-Square,” *2023 6th Int. Conf. Inf. Commun. Technol. ICOIACT 2023*, pp. 454–458, 2023, doi: 10.1109/ICOIACT59844.2023.10455895.

[8] J. Wang, K. Fu, and C. T. Lu, “SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection,” *Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020*, pp. 1699–1708, 2020, doi: 10.1109/BigData50022.2020.9378065.

[9] J. Brownlee, “Deep Learning for Natural Language Processing : Develop Deep Learning Models for Natural

Language in Python,” *Mach. Learn. Mastery*, p. 414, 2017, [Online]. Available: http://web.stanford.edu/class/cs224n/readings/cs224n-2019-notes06-NMT_seq2seq_attention.pdf

[10] O. Mogren, *Representation Learning for Natural Language*. 2018.

[11] Dedy Sugiarto, Ema Utami, and Ainul Yaqin, “Perbandingan Kinerja Model TF-IDF dan BOW untuk Klasifikasi Opini Publik Tentang Kebijakan BLT Minyak Goreng,” *J. Tek. Ind.*, vol. 12, no. 3, pp. 272–277, 2022, doi: 10.25105/jti.v12i3.15669.

[12] N. Yusliani, S. A. Q. Aruda, M. D. Marieska, D. M. Saputra, and A. Abdiansah, “The effect of Chi-Square Feature Selection on Question Classification using Multinomial Naïve Bayes,” *Sinkron*, vol. 7, no. 4, pp. 2430–2436, 2022, doi: 10.33395/sinkron.v7i4.11788.

[13] G. R. Ditami, E. F. Ripanti, and H. Sujaini, “Implementasi Support Vector Machine untuk Analisis Sentimen Terhadap Pengaruh Program Promosi Event Belanja pada Marketplace,” *J. Edukasi dan Penelit. Inform.*, vol. 8, no. 3, p. 508, 2022, doi: 10.26418/jp.v8i3.56478.

[14] Y. X. Chu, X. G. Liu, and C. H. Gao, “Multiscale models on time series of silicon content in blast furnace hot metal based on Hilbert-Huang transform,” *Proc. 2011 Chinese Control Decis. Conf. CCDC 2011*, pp. 842–847, 2011, doi: 10.1109/CCDC.2011.5968300.

[15] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. 2011. doi: https://doi.org/10.1016/C2009-0-19715-5.

[16] K. Dinas *et al.*, “Prediksi Jumlah Penggunaan BBM Perbulan Menggunakan Algoritma Decition Tree (C4.5) Pada,” *J. Inform. dan Teknol.*, vol. 1, no. 1, pp. 56–63, 2018.

[17] L. T. E. . Kusrini, *Algoritma Data Mining. Buku Algoritma Data Mining*, I. Yogyakarta: C.V ANDI, 2009. [Online]. Available: https://books.google.co.id/books?id=-Ojclag73O8C&printsec=frontcover&hl=id#v=onepage&q&f=false

[18] N. T. Romadloni, I. Santoso, and S. Budilaksono, “Perbandingan Metode Naive Bayes, Knn Dan Decision Tree Terhadap Analisis Sentimen Transportasi Krl Commuter Line,” *J. IKRA-ITH Inform. J. Komput. dan Inform.*, vol. 3, no. 2, pp. 1–9, 2019.

[19] A. Tanggu Mara, E. Sediyono, and H. Purnomo, “Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Metode Pembelajaran Dalam Jaringan (DARING) Di Universitas Kristen Wira Wacana Sumba,” *Jointer - J. Informatics Eng.*, vol. 2, no. 01, pp. 24–31, 2021, doi: 10.53682/jointer.v2i01.30.

[20] R. Jose and V. S. Chooralil, “Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach,” *IEEE*, 2016, doi: https://doi.org/10.1109/SAPIENCE.2016.7684133.

[21] P. K. Sari and R. R. Suryono, “Komparasi Algoritma Support Vector Machine Dan Random Forest Untuk Analisis Sentimen Metaverse,” *J. Mnemon.*, vol. 7, no. 1, pp. 31–39, 2024, doi: 10.36040/mnemonic.v7i1.8977.