Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.4, August 2025, pp. 1681~1686
e-ISSN: 2548-6861

Prediction of Cyberbullying in Social Media on Twitter Using Logistic Regression

Santi Prayudani[1*], Lilis Tiara Adha[2*], Tika Ariyani[3*], Arif Ridho Lubis[4*]
*Teknik Komputer dan Informatika, Politeknik Negeri Medan, Medan
santiprayudani@polmed.ac.id[1], lilistiaraadha@students.polmed.ac.id[2], tikaariyani@students.polmed.ac.id[3], arifridho@polmed.ac.id[4]

Article Info

Article history:

Received 2025-06-16
Revised 2025-07-07
Accepted 2025-07-19

Keyword:

Cyberbullying,
Logistic Regression,
SMOTE Method,
Twitter

ABSTRACT

As cases of cyberbullying on social media increase, there is a need for efficient measures to detect the vice. This research aims to establish the application of machine learning algorithms in analyzing text on social media to determine potentially harmful comments using logistic regression. The first and most important research question of this study is to assess the extent to which the model is capable of correctly identifying the comments that contain features of cyberbullying and those that do not. The data set included comments from different social media sites and was preprocessed before further analysis was conducted on it. Exploratory Data Analysis was applied in the study to establish relationships and textual features with bullying behavior. As with any other model, after training and testing the model, the results were analyzed using parameters like precision, precision, gain, and F1 statistics. The outcomes of this study revealed that the use of logistic regression models can give a fairly satisfactory level of accuracy in identifying cyberbullying. In light of this, this study underscores the need to use machine learning algorithms to minimize negative actions in cyberspace.

I. INTRODUCTION

Over the last few years, technological development and the growing popularity of the internet have led to the development of various severe social issues, such as cyberbullying, privacy violations, and fake news dissemination [1]. One of them is social media which is often used since it offers many benefits to its users and allows them to interact with others using technological items in their everyday lives. [2]. The main problem that is currently emerging is cyberbullying [3]. New developments in social media have altered the way people use Social Media to post unethical information including the use of Twitter, Instagram, and Facebook to spread hatred and cyberbullying [4]. Twitter is one of the platforms that we can use to share ideas with others and express ourselves, but it is also an environment where trolls operate [5].

Cyberbullying can be in the form of sending nasty or threatening messages, posting embarrassing pictures or videos, gossiping/posting rumors, blocking someone from any form of communication, and impersonation by posting something using another person's identity online [6]. Cyberbullying is the process of using information technology to threaten, tease, or terrorize another person with the effect of causing untold harm to the victim [7]. Certain forms of cyberbullying include unlawful actions or can be deemed as a crime [8]. In addition, interactions on social media platforms are done through screens hence making the process of cyberbullying different from bullying that is face to face [9]. Therefore, this study considers cyberbullying as a classification problem, where it is necessary to determine whether an event is classified as cyberbullying or not. To this end, we use machine learning algorithms, specifically the logistic regression method, which has been proven effective in solving binary classification problems, to improve the accuracy and effectiveness of cyberbullying prediction on social media, specifically on Twitter [10].

This study aims to identify cyberbullying on different social media sites through the classification of comments as either bullying or not. We apply a machine learning technique known as the logistic regression method, which is suitable for solving binary classification. The model is trained to detect

potentially toxic comments by processing the text gathered from social media sites where each comment is labeled. In this study, the model is tested on a diverse data set to establish how well it can predict cyberbullying comments, which is useful for platform developers in creating safer online environments.

The purpose of this study is to assess the accuracy of the model in identifying the comments that are likely to contain cyberbullying content in social media using logistic regression. Therefore, by applying the mentioned model, this study aims to enhance the community's awareness of how to recognize potentially toxic comments such as sarcasm and insults in normal conversation. This is to ensure that the users are in a position to understand the impacts of cyberbullying, refrain from cyberbullying, as well as improve their online image. Therefore, it is postulated that this study will assist in the development of better detection algorithms and protective measures against cyberbullying behavior.

Research conducted by Aaminah et al. [11] has studied the detection of cyberbullying using different machine learning techniques, and the differences in accuracy. In some studies, Support Vector Machine (SVM) and Ensemble models outperformed other models with overall accuracy of 79%. However, the performance of Logistic Regression was also good with the accuracy rate of about 78 % while that of the Random Forest was 76.7 %. However, Naïve Bayes gave a relatively good accuracy of 76% which was almost as good as the more complicated models. This research also affirms that several machine learning algorithms can be used to predict cyberbullying with variations in the precision of the used techniques. In our study the only algorithm used was logistic regression but for accuracy our results revealed that the model was an average of 80.83% which we consider slightly better than other similar studies that have reported Logistic Regression accuracy of approximately 78%.

The second research conducted by Nureni et al. [12] has highlighted the effectiveness of different machine learning techniques with a special focus on Random Forest Classifier in identifying cyberbullying on social media platforms including Twitter. Random Forest was seen to perform well in the classification of bullying-related text with median values of 0.77, 0.73 and 0.94 on the several datasets which clearly shows the versatility of Random Forest in handling text classification. However, our research using the Logistic Regression model had Training accuracy of 89.82% and Testing accuracy of 71.85%. While Random Forest in prior studies employed the ensemble method to enhance the outcomes, the present study demonstrates that Logistic Regression can also offer satisfactory performance, particularly when class balance methods are employed.

Research conducted by Pranathi et al. [13] combats cyberbullying on social media using machine learning and both controlled and uncontrolled surveillance techniques. One of the emerging factors for enhancing the sentiment analysis was the proper identification of the right keywords. According to the test, the model that was based on the Logistic Regression had the best performance: F1 Score was equal to 0.93 after including the custom user data. However, in the present work, Logistic Regression was used with SMOTE for handling class imbalance issues, without incorporating any additional user data and achieved Testing Accuracy of 71.85%. Our work concentrates on class balancing and cross-validation techniques to improve model stability rather than on data augmentation.

Research conducted by Khalid et al. [14] applied a method to identify cyberbullying from textual data available on Kaggle. The study identified three main topics in the data set: 30.4% for religious tweets, 39.3% for sexist tweets, and 30.3% for general tweets. The accuracies of the applied machine learning algorithms that include SVM, Naïve Bayes, Random Forest, and Logistic Regression were 94.9%, 93.1%, 94.66%, and 95% respectively. More work is suggested to enhance the precision by using LDA-based technique and computing cosine similarity between clusters. Compared to the above study, our study only applied Logistic Regression with SMOTE to cope with the class imbalance issue and achieved a Testing Accuracy of 71.85%. Our concern is more towards the class balance and model validation by cross validation while the previous researchers have concern towards exploring the different algorithms and text topics and proposing LDA methods for better accuracy.

II. METHODOLOGY

Figure 1 represents a framework of procedures and steps for studying cyberbullying prediction on social media platforms with a focus on the Twitter application. To maintain reliability and validity, this research adopts a systematic approach. The following research workflows are proposed to enlighten the process so that it can be accomplished properly. Therefore, all the research activities and procedures are performed systematically in order to obtain the best outcome.

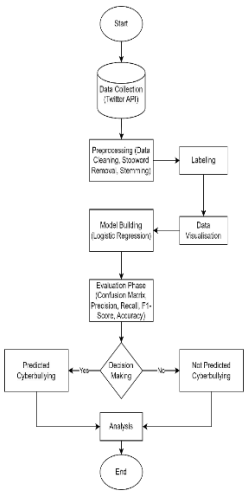


Figure 1. Work Flow Diagram

A. Dataset

The dataset that we employ is a dataset with tweets of cyberbullying; it has two columns: Text and CB_Label. The Text column refers to the main column containing the tweet text. The CB_Label is a column that contains a binary label that defines whether the tweet has indications of cyberbullying (1) or not (0). There are a total of 11101 tweets in this dataset. With this dataset, we shall be able to predict cyberbullying in this Twitter application.

TABEL I. Top Words

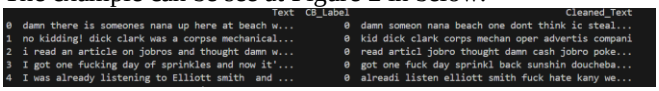
No.	Text	CB_Label
1.	damn there is someones	0
2.	no kidding! dick clark	0
3.	i read an article on jobros	0
4.	I got one fucking day of	0
5.	…	0
6.	I told Derek to go fuck	0
7.	I'm watching the new	0
8.	My mom didn't like	0
9.	…	0
10.	thanks! this is the only	0
11.	…	0
12.	Looks like that little Cut-	0
13.	damn there is someones	0
14.	no kidding! dick clark	0
15.	i read an article on jobros	0
16.	I got one fucking day of	0
11101.	…	1

B. Preprocessing Data

After data cleaning and exploratory data analysis, this research moves to the data preprocessing stage where preprocessing is done using the Python programming language in Visual Studio Code to ensure that the dataset to be used for modeling has been properly cleaned. This process is very significant in the case of detection of cyberbullying because the quality of data determines the quality of the model to be developed. Several things were involved in this research, including :

1. First, clean the data by removing all non-letters from the text and making all letters lowercase and other features that are in the text, such as URLs, mentions, and hashtags.
2. After cleaning the text an additional step of tokenization is performed on it to split the text into individual words. This process divides the text into tokens and eliminates other words known as stop words that do not hold much significance to the context of the analysis.
3. Applying stemming to the dataset in order to have all the different forms of a word in the same group in order to improve the analysis.
4. Once all the cleaning steps are done then merge the cleaned words back into a single text string, making the final result easy to read and using the Clean_Text function which will return the text that has undergone the cleaning process.
5. To prepare the cleaned text data [15] for use in the classification model training process, the Bag of Words model is developed.

The example can be see at Figure 2 in below:



	Text	CB_Label	Cleaned_Text
0	damn there is someones nana up here at beach w...	0	damn someon nana beach one dont think ic steal...
1	no kidding! dick clark was a corpse mechanical...	0	kid dick clark corps mechan oper advertis compani
2	i read an article on jobros and thought damn w...	0	read articl jobro thought damn cash jobro poke...
3	I got one fucking day of sprinkles and now it'...	0	got one fuck day sprinkl back sunshin douchebа...
4	I was already listening to Elliott smith and ...	0	alreadi listen elliott smith fuck hate kany we...

Figure 2. Preprocessing Data

C. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA), which is a process of data preparation, came to be because of the first and most important step to be followed in processing data, including the data related to cyberbullying, before one proceeds to process the data in other complicated ways. EDA enables the discovery of the characteristics of data that are not immediately visible, as well as detection of outliers, and distribution of all the variables and their mutual dependencies. First, the structure of the dataset is preprocessed; its dimensionality, data types, and other peculiarities, as well as missing data treatment, are discussed. Following that, bar charts are applied to demonstrate the words that appeared most frequently in the analyzed comments or tweets.

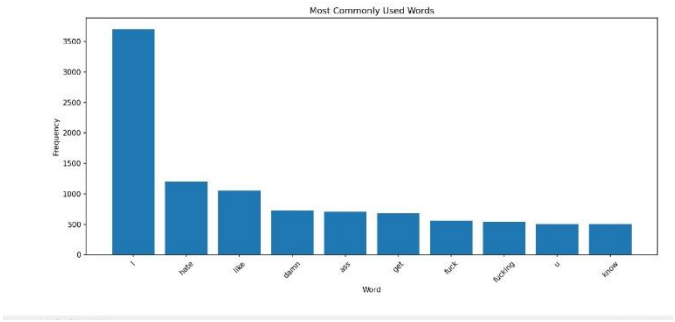


Figure 3. Most Commonly Used Words

The image (Figure 3) above shows a count plot for the most used word frequency in the comments or tweets dataset analyzed above. On the x-axis (word) are the words that are most frequently used in the text. Some of the words that were written are “I”, “hate”, “like”, “damn”, “ass”, “get”, “fuck”, “fucking”, “u”, “know”. On the x-axis there is the number of appearances of each word in the given data set. The number that is observed on the graph shows the frequency of each word in the text.

D. Model Building

Logistic regression algorithm used in model building of this case study is quite appropriate for binary classification, which is whether a particular tweet has cyberbullying or not. The data used in this study is 11100 with the training data being 80% and the testing data 20%. The Logistic Regression algorithm used for classification of cyberbullying yields an

accuracy of 80.83% which means that the model predicts the class correctly for total data used in 80% of the cases. This result depicts that the model is quite effective in classifying the existence of cyberbullying in a tweet or comment or not depending on the patterns from the training data.The label we use initializes the label "0" for not cyberbullying and "1" for cyberbullying, where the algorithm we use is logistic regression:

The equation (1) above represents the Logistic Regression is based on a logistic function, mapping any real number to a number between 0 and 1. The formula for the logistic function is:

$$P(Y = 1|X) = \frac{1}{1+e-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n} \quad (1)$$

Where:

- P(y=1|X) represents the probability that the outcome y equals 1 (positive class/cyberbullying) given the input features X.
- X1,X2,...,Xn are the independent features or input variables.
- β0 is the bias or intercept.
- β1,β2,...,βn are the regression coefficients that determine the influence of each feature on the outcome probability.
- e is the base of the natural logarithm

What is, in fact, the basis of Logistic Regression, is the sigmoid function that maps the output of a linear model into a probability value ranging between 0 and 1.
The sigmoid function is expressed as:

$$\sigma(z) = \frac{1}{1+ e^{-z}} \quad (2)$$

The equation (2) above represents, the sigmoid function restricts the output between 0 and 1 making the algorithm ideal for use in binary classification problems.

E. Evaluation Phase

Once the logistic regression model was developed and optimized the final phase of the study consisted of an evaluation of the performance of the established machine learning model in detecting cyberbullying on the twitter site. These include confusion matrix, F1 score, Precision and Recall. This enables us to know how well our model is in classifying comments as containing cyberbullying or not containing it. F1 score balances precision and recall, this logistic regression model was evaluated using the accuracy measure, below:

1. Precision

Precision measures how many positive predictions are correct compared to the total positive predictions made by the model. The formula for precision is given in Equation (3):

$$Precision = \frac{\text{True Positives TP}}{\text{True Positives TP+False Positives FP}} \quad (3)$$

Where:

- TP = True Positives (number of positives correctly predicted)
- FP = False Positives (number of negatives predicted as positive).

2. Cross Validation

Performed after every model adjustment to better assess how well a model works and if it predicts correctly.

3. Recall

Recall measures how many positive predictions are correct compared to the total actual positives. The for recall is given in Equation (4):

$$Recall = \frac{\text{True Positives TP}}{\text{True Positives TP+False Negatives FN}} \quad (4)$$

Where:

- FN = False Negatives (number of positives predicted as negative)

4. F1-Score

F1-score is the harmonic mean of precision and recall, providing a single number that reflects the balance between the two. The F1 Score, as expressed in Equation (5), is calculated using the formula:

$$\text{F1 Score} = 2 \text{ X } \frac{\text{Precision x Recall}}{\text{Precision+Recall}} \quad (5)$$

5. Support

Support is the actual number of a particular class in the dataset. the support for class 0 is 1128 and for class 1 is 1092.

6. Accuracy

Accuracy measures how many predictions are correct compared to the total predictions made. The formula for accuracy is provided in Equation (6):

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{Total Samples}} \quad (6)$$

III. RESULTS AND DISCUSSION

By applying the logistic regression model to training and testing, the results are obtained as in the Table 2 below:

TABLE 2.
TRAINING AND TESTING SCORES

Model Name	Training Score	Testing Score
Logistic Regression	89,8%	71,8%

For data visualization using a confusion matrix, TN cells (881) are dark blue, FP cells (247) are gray, FN cells (372) are sky blue, and TP cells are bright blue.

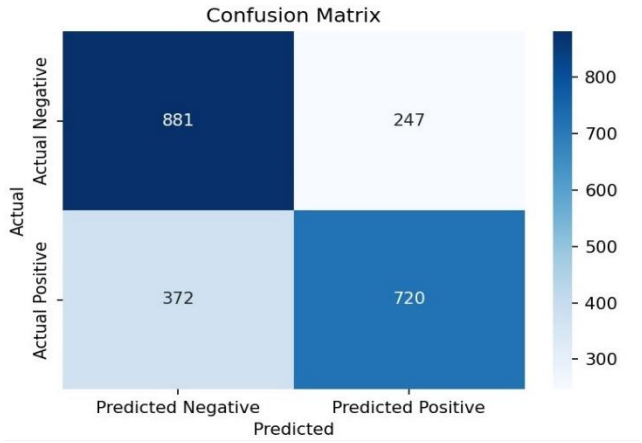


Figure 4. Confusion Matrix

Figure 4 shows the confused matrix obtained from the classification model, we can calculate the evaluation matrix shown below. In this confused matrix, 881 actual negative cases, and 881 negative cases predicted by the model are considered. These are known as True Negatives (TN). There were 247 actual negative cases that were classified as positive by the model, which are called False Positives (FP) or type 1 errors. Furthermore, there were 372 cases that were actually positive but were classified as negative by the model and therefore False Negatives (FN) or type 2 errors. In contrast, there were 720 cases that were actually positive, and the model predicted them to be positive; these are called True Positive (TP).

Below are the results of the classification of evaluation methods.

TABLE 3.
EVALUATION RESULT

Class	Precision	Recall	F1-Score
0	0.70	0.78	0.74
1	0.75	0.65	0.69
Accuracy			0.72
Macro Avg	0.72	0.72	0.72
Weighted Avg	0.72	0.72	0.72

As show in Table 3 above, the precision for the negative class is 0.70 and for the positive class it is 0.75. The recall for the negative class is 0.78 and for the positive class is 0.65. They also had almost similar values of the F1-score for both classes, which is quite good for the model.

In SMOTE implementation, the Training Logistic Regression was 89.82% and the Testing Logistic Regression was 71.85%, thus showing good effectiveness. These accuracy figures suggest that the model is better suited for training data rather than for testing data. The use of SMOTE is intended to address the issue of the existence of many samples within the minority class in the dataset.

In the case of the testing dataset, the positive and negative labels are distributed with 49.19% positive labels and 50.81% negative labels. This indicates that the ratio of the label distribution is reasonably balanced after applying the SMOTE method.

Cross validation is done in order to check the robustness of the model generated. The cross-validation score also provides some indication of how the model behaves with different portions of the data set. In the present analysis, the cross-validation score is calculated at 80.83% on average, meaning that the model is stable and performs well when assessed with different data.

and the final result states

```
Dataset ini mengandung cyberbullying berisi 5550 data.
```

with 80.83 percent accuracy according to the model and function stated above.

IV. CONCLUSION

This research aims to collect and categorise access permissions from APK files distributed via WhatsApp, grouping them into malware and non-malware categories. The study identified patterns and correlations that could indicate whether a distributed file is malware or not. The findings highlighted differences in accuracy among various classification algorithms used, including random forests, logistic regression, and gradient boosting machines. Notably, the GBM algorithm demonstrated superior performance, achieving a 100% accuracy rate in identifying harmful applications, although the difference to other algorithms was not significant, it proved to be the most reliable method among those tested.

This study emphasises the importance of implementing robust security measures when using WhatsApp, in addition to educating users about the potential risks associated with granting app permissions. The primary objective is to empower the general public with the necessary tools and knowledge to identify potentially malicious files shared by compromised contacts, such as seemingly innocuous files like wedding invitations or misleading content, which may actually pose a malware threat. Ultimately, the research aims to mitigate hacking incidents caused by the installation of malware and contributes to the field of cybersecurity by demonstrating the effectiveness of machine learning techniques in combating the spread of malware on social media platforms.

A potential enhancement for the future is to develop a Graphical User Interface (GUI) to improve user accessibility, ensuring that the general public can interact with the system more clearly. Continuously refining these machine learning models holds the promise of creating stronger defences capable of adapting to the changing landscape of cyber threats.

REFERENCES

[1] A. S. G. Tabares, J. E. Restrepo, and G. Zapata-Lesmes, "The effect of bullying and cyberbullying on predicting suicide risk in adolescent females: The mediating role of depression," *Psychiatry Res*, vol. 337, Jul. 2024, doi: 10.1016/j.psychres.2024.115968.

[2] N. Chamidah and R. Sahawaly, "Comparison Support Vector Machine and Naive Bayes Methods for Classifying Cyberbullying in Twitter," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 7, no. 2, p. 338, Sep. 2021, doi: 10.26555/jiteki.v7i2.21175.

[3] A. Almomani, K. Nahar, M. Alauthman, M. A. Al-Betar, Q. Yaseen, and B. B. Gupta, "Image cyberbullying detection and recognition using transfer deep machine learning," *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 14–26, Jan. 2024, doi: 10.1016/j.ijcce.2023.11.002.

[4] C. Poh Theng, N. Fadzilah Othman, R. Syahirah Abdullah, S. Anawar, Z. Ayop, and S. Najwa Ramli, "Cyberbullying Detection in Twitter Using Sentiment Analysis," *IJCSNS International Journal of Computer Science and Network Security*, vol. 21, no. 11, 2021, doi: 10.22937/IJCSNS.2021.21.11.1.

[5] D. Musleh *et al.*, "AMachine Learning Approach to Cyberbullying Detection in Arabic Tweets," *Computers, Materials and Continua*, vol. 80, no. 1, pp. 1033–1054, 2024, doi: 10.32604/cmc.2024.048003.

[6] W. Xiao and M. Cheng, "The Relationship between Internet Addiction and Cyberbullying Perpetration: A Moderated Mediation Model of Moral Disengagement and Internet Literacy," *International Journal of Mental Health Promotion*, vol. 25, no. 12, pp. 1303–1311, 2023, doi: 10.32604/ijmhp.2023.042976.

[7] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, "Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT," *Information (Switzerland)*, vol. 14, no. 8, Aug. 2023, doi: 10.3390/info14080467.

[8] P. Yi and A. Zubiaga, "Session-based cyberbullying detection in social media: A survey," *Online Soc Netw Media*, vol. 36, Jul. 2023, doi: 10.1016/j.osnem.2023.100250.

[9] C. Marinoni, M. Rizzo, and M. A. Zanetti, "Social Media, Online Gaming, and Cyberbullying during the COVID-19 Pandemic: The Mediation Effect of Time Spent Online," *Adolescents*, vol. 4, no. 2, pp. 297–310, Jun. 2024, doi: 10.3390/adolescents4020021.

[10] A. Raza, M. Bilal, and M. Fahad Rauf, "Comparative Analysis Of Machine Learning Algorithms For Fake Review Detection," *International Journal of Computational Intelligence in Control Copyrights @Muk Publications*, vol. 13, no. 1, 2021.

[11] A. Ali and A. M. Syed, "Cyberbullying Detection Using Machine Learning."

[12] N. A. Azeez, S. O. Idiakose, C. J. Onyema, and C. Van Der Vyver, "Cyberbullying Detection in Social Networks: Artificial Intelligence Approach," *Journal of Cyber Security and Mobility*, vol. 10, no. 4, pp. 745–774, 2021, doi: 10.13052/jcsm2245-1439.1046.

[13] P. Pranathi, V. Revathi, P. Varshitha, S. Shaik, and S. Bhutada, "Logistic Regression Based Cyber Harassment Identification," *Journal of Advances in Mathematics and Computer Science*, vol. 38, no. 8, pp. 76–85, Jun. 2023, doi: 10.9734/jamcs/2023/v38i81792.

[14] K. M. O. Nahar, M. Alauthman, S. Yonbawi, and A. Almomani, "Cyberbullying Detection and Recognition with Type Determination Based on Machine Learning," *Computers, Materials and Continua*, vol. 75, no. 3, pp. 5307–5319, 2023, doi: 10.32604/cmc.2023.031848.

[15] D. E. Kurniawan, A. Dzikri, E. Br. Sembiring, N. Ardi, H. Mochtoha, J. Friadi, and P. Prasetyawan, Kecerdasan Bisnis: Literasi Data untuk Pengambilan Keputusan. Media Sains Indonesia, 2024.