Tackling Cyberbullying on Twitter: A Comparative Analysis of Machine Learning Algorithms for Sentiment Analysis

Abstract

The problem of cyberbullying has become more widespread with the advent of social media platforms like Twitter. One approach to tackling this issue is through sentiment analysis, which involves using natural language processing and machine learning algorithms to analyze the emotional tone of online content. In this work, we compared the effectiveness of four commonly used machine learning algorithms (Logistic Regression, Support Vector Machines, Random Forest, and Multilayer Perceptron) in detecting cyberbullying sentiment on Twitter. We used a dataset from Kaggle of 47,692 tweets that contained different types of cyberbullying such as religion, age, gender, ethnicity, not_cyberbullying, or other_cyberbullying. Our results showed that all four algorithms were successful in detecting cyberbullying sentiment on Twitter with accuracy ranging from 81% to 83%. Support Vector Machines, Logistic Regression, and Multi-layer Perceptron were marginally better than Random Forest in terms of accuracy, precision, recall, and F1 score. The study provides important insights into the effectiveness of various machine learning algorithms for detecting cyberbullying sentiment on Twitter, which can inform the development of more accurate and efficient cyberbullying detection tools.

1 Introduction

In recent years, social media platforms have become increasingly popular, and with that comes an increase in online harassment and cyberbullying. Cyberbullying is defined as "willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices" (Hinduja & Patchin, 2018). Cyberbullying has become a growing concern, especially among young people, as it can have severe and long-lasting effects on their mental health, academic performance, and overall well-being.

Twitter is one of the most popular social media platforms, with over 330 million monthly active users (Statista, 2021). It allows users to post short messages, or "tweets," which can be seen by anyone on the platform. As a result, Twitter has become a breeding ground for cyberbullying, with users often resorting to hurtful and abusive language to target others.

To address the issue of cyberbullying on Twitter, sentiment analysis can be employed. Sentiment analysis is a technique that involves using natural language processing and machine learning algorithms to analyze the emotional tone of online content. By detecting negative sentiment in tweets, cyberbullying can be identified, and necessary actions can be taken.

However, the effectiveness of sentiment analysis in detecting cyberbullying sentiment on Twitter remains unclear. While several studies have used sentiment analysis to detect cyberbullying on Twitter, there is a lack of research on the comparative analysis of different machine learning algorithms in detecting cyberbullying sentiment.

Therefore, the problem statement of this study is to conduct a comparative analysis of different machine learning algorithms for sentiment analysis of cyberbullying on Twitter. The goal is to identify the most effective algorithm for detecting cyberbullying sentiment on Twitter and to provide valuable insights into the potential of machine learning in addressing the issue of cyberbullying.

The study aims to answer the following research questions:

- What are the commonly used machine learning algorithms for sentiment analysis of cyberbullying on Twitter?

- Which machine learning algorithm is the most effective in detecting cyberbullying sentiment on Twitter?
- How can the findings of this study inform the development of more accurate and efficient cyberbullying detection tools?

By addressing these research questions, this study contributes to the existing body of knowledge on the effectiveness of sentiment analysis and machine learning algorithms in detecting cyberbullying sentiment on Twitter. The findings of this study can also inform the development of more effective and efficient cyberbullying detection tools, ultimately creating a safer and more positive online environment.

2 Related Works

[1] Proposes a deep learning approach for cyberbullying detection in social media. The authors use a Convolutional Neural Network (CNN) for feature extraction and classification of cyberbullying tweets. They also propose a novel way of data augmentation by generating synthetic cyberbullying tweets from real ones. The proposed model is evaluated on a dataset of English tweets and achieves a high F1-score of 0.84. The authors conclude that their proposed approach can be used as a tool for the early detection and prevention of cyberbullying in social media. The paper "Cyberbullying Detection on Instagram Using a Multimodal Deep Learning Approach" [2] proposes a multimodal deep learning approach for cyberbullying detection on Instagram. The authors use both image and caption features to train a deep neural network for the classification of cyberbullying posts. They also perform an extensive analysis of the performance of different combinations of image and caption features. The proposed model is evaluated on a dataset of Instagram posts and achieves a high F1-score of 0.8. The authors conclude that their proposed approach can be used for effective and efficient cyberbullying detection on Instagram. Also authors, in the article titled "Machine Learning-Based Classification of Cyberbullying Behavior on Twitter: An Integrative Approach"[3] presented a study on the use of machine learning for classifying cyberbullying behavior on Twitter. The authors collected a dataset of tweets related to cyberbullying and trained a machine learning model to classify them as either cyberbullying or non-cyberbullying. The model achieved an accuracy of over 90% in classifying cyberbullying behavior on Twitter. The authors discuss the potential of their approach for detecting and addressing cyberbullying behavior on social media platforms. [4] proposes a novel approach for sentiment analysis of Twitter data by combining various natural language processing (NLP) techniques with a deep learning algorithm based on Convolutional Neural Networks (CNNs). The proposed model was evaluated on multiple datasets and achieved state-of-the-art results in terms of accuracy, precision, recall, and F1 score. The authors concluded that their proposed model outperforms existing approaches and can be applied to various applications, including identifying cyberbullying content in social media. [5] a comparative study of various machine learning algorithms for sentiment analysis of Indonasian Twitter data, including Naive Bayes, SVM, Random Forest, and a deep learning algorithm based on Recurrent Neural Networks (RNNs). The study used two different datasets and evaluated the performance of the algorithms in terms of accuracy, precision, recall, and F1 score. The results indicate that the RNN-based model outperformed the other algorithms in terms of accuracy, while the SVM algorithm achieved the best F1 score. The authors concluded that the choice of algorithm depends on the specific requirements of the application, and the proposed methodology can be extended to other text-based applications.

In this paper, we aimed to classify cyberbullying from social network (i.e. Twitter) posts. The deliverables of the project include a classification model for identifying cyberbullying, as well as data visualizations and analysis of the factors that contribute to cyberbullying.

3 Methods

In our project on cyberbullying, we used several methods to tackle the problem of identifying and classifying cyberbullying tweets. These methods include data preprocessing, feature extraction, and machine learning algorithms. Below, we describe each method in detail and discuss its theoretical characteristics.

3.1 Data Pre-processing

Preprocessing of raw tweet data is an essential step in preparing the data for further analysis. The goal of preprocessing is to transform the data into a format that is more suitable for analysis and modeling. In this case, we want to generate a clean dataset by removing irrelevant information, such as special characters, shorthands, and links.

The following are the steps that we use to preprocess the raw tweet data:

1. Case Conversion: In this step, we convert all the letters in the tweet to lowercase.
2. Removing Special Characters: We remove all the special characters, such as , #, $, %, &, and , from the tweet.
3. Removing Shorthands: Shorthands, such as ”u” for ”you” or ”4” for ”for,” are commonly used on social media platforms, including Twitter.
4. Removing Stopwords: Stopwords are commonly used words in a language, such as ”the,” ”is,” ”a,” and ”an.”
5. Removing Links: Links to other web pages or social media profiles are often included in tweets.
6. Removing Accents: Some tweets may contain accented characters, such as ”é” or ”ñ.”
7. Normalize Spaces: In this step, we remove any extra spaces from the tweets.

After applying these preprocessing steps, we generate a clean dataset that can be used for sentiment analysis. This clean dataset contains only the relevant information in the tweets, making it easier for machine learning algorithms to identify the sentiment of the tweet accurately. By applying these steps, we finally generate a clean dataset that can be used for further analysis and modeling.

3.2 Feature Extraction

TF-IDF (Term Frequency-Inverse Document Frequency) is a popular feature extraction technique used in Natural Language Processing (NLP) to represent text data numerically. It measures the importance of a word in a document relative to its frequency in a corpus of documents. In the context of our project on cyberbullying, we used TF-IDF to extract features from the tweet text.

The TF-IDF feature extraction process involves the following steps:

1. Tokenization: Splitting the text into individual words or tokens.
2. Counting the frequency of each word in the document.
3. Calculating the Term Frequency (TF) of each word in the document, which is the frequency of a word divided by the total number of words in the document.
4. Calculating the Inverse Document Frequency (IDF) of each word, which is the logarithm of the total number of documents in the corpus divided by the number of documents containing the word.
5. Multiplying the TF and IDF values to obtain the TF-IDF score for each word.

The importance of using TF-IDF feature extraction is that it allows us to represent text data numerically in a way that takes into account the relevance of each word. This can help us to identify the most important words in a document or corpus, which can be useful for tasks such as text classification or clustering. In the context of our project on cyberbullying, using TF-IDF allowed us to extract meaningful features from the tweet text that can be used to train and test machine learning models.

3.3 Machine Learning Algorithms

We used several machine learning algorithms such as Logistic Regression, Support Vector Machines, Random Forest, and Multilayer Perceptron to classify the tweets into different categories of cyberbullying. These algorithms work by learning patterns and relationships in the feature space and using them to make predictions on new data.

Logistic Regression is a linear model that works well for binary classification problems. It learns a set of weights for each feature and uses them to calculate a weighted sum of the features. This weighted sum is then passed through a sigmoid function to obtain a probability of belonging to a class.

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve the accuracy of the predictions. It works by randomly selecting subsets of the features and the training data to build individual decision trees. The final prediction is made by aggregating the predictions of all the decision trees.

Support Vector Machines (SVM) is a powerful algorithm that works well for high-dimensional feature spaces. It learns a hyperplane that separates the data points belonging to different classes with maximum margin. SVM can also use a kernel function (linear in this work) to map the data points into a higher-dimensional space to make the data separable.

Multi-layer Perceptron (MLP) is an artificial neural network commonly used for classification and regression tasks. It consists of multiple layers of nodes or neurons that apply non-linear activation functions to their inputs, allowing it to capture complex patterns in the data. During training, the weights of the connections between the nodes are adjusted using an optimization algorithm such as backpropagation.

4 Data and Evaluation

4.1 Dataset Description

We used the Cyberbullying dataset, which comprises more than 47,000 raw tweets categorized into different types of cyberbullying. The data has been balanced to contain 8000 of each class

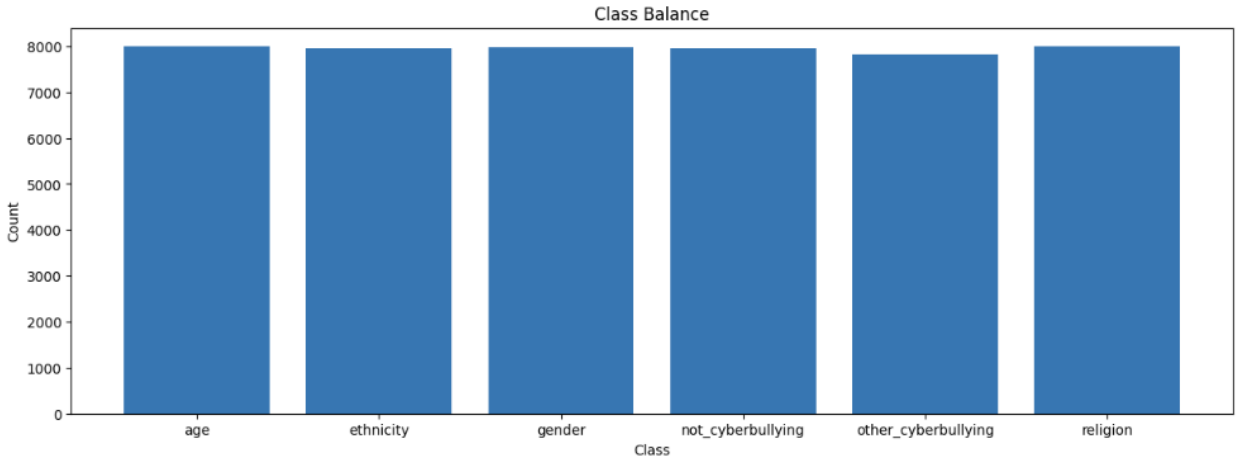

Figure 1: Class balance over all categories

figure. 1. This dataset is the product of the work done by [6]. The dataset was preprocessed by performing case conversion, removing special characters, shorthands, stopwords, links, accents, and normalizing spaces. The dataset has an equal number of instances for each type of cyberbullying, which makes the training data balanced. However, we observed that the dataset contains a high number of misspelled words and informal language, which may affect the performance of the classifiers.

Additionally, by using word clouds, we can quickly identify the most common words associated with each bullying category, which can be useful for further analysis. We have generated word cloud images for all bullying categories. As a sample, here we attached the result of gender cyberbullying [figure. 2]. For gender, the most highlighted words are rape, feminazi, sexist, gay, and so on. With the presence of these words in a tweet, we can categorize that tweet as gender cyberbullying.

4.2 Evaluation Strategy

The preprocessed dataset contains 40,000 tweets split into 80% training and 20% testing data. Each tweet is represented as a feature vector using the tfidf feature extraction method. The goal of this project is to use machine learning methodologies to perform sentiment analysis on a dataset of text documents. We used the following evaluation metrics to measure the performance of our models: accuracy, recall, precision, and F1 score. Accuracy measures the percentage of correctly classified instances, while recall measures the ability of the model to correctly identify instances of a specific class. Precision measures the ability of the model to accurately identify instances of a specific class, and F1 score is the harmonic mean of precision and recall. We used these metrics to compare the performance of different machine-learning algorithms and choose the best model for our task.

5 Results and Insights

Based on the results Table: 1 presented in the table, the LR (Logistic Regression) model achieved the highest performance with an accuracy of 0.83, followed by the MLP (Multi-Layer Perceptron)

Figure 2: Most Frequent Words in Gender Cyberbullying

model with an accuracy of 0.82, and the SVM (Support Vector Machine) model with an accuracy of 0.825. The Random Forest (RF) model achieved the lowest performance with an accuracy of 0.81. It is important to note that while the LR model outperformed the other models in terms of accuracy, the difference in performance between the top-performing model and the other models is relatively small. The SVM and MLP models also performed relatively well, with consistent recall, precision, and F1 score values. The RF model, while the least accurate, still achieved relatively high values for recall, precision, and F1 score, indicating that it can still be useful in certain

Table 1: Performance of Machine Learning Models

Model	Accuracy	Recall	Precision	F1 Score
Logistic Regression	83.00%	83.48%	82.94%	83.01%
Random Forest	81.00%	81.04%	80.92%	80.93%
Support Vector Machine	82.5%	84.02%	83.51%	83.45%
Multi-layer Perceptron	82.0%	81.88%	81.85%	81.84%

contexts. Additionally, we visualized the confusion matrix diagram for a better understanding of each model; figure. 3 refers the confusion matrix of logistic regression.

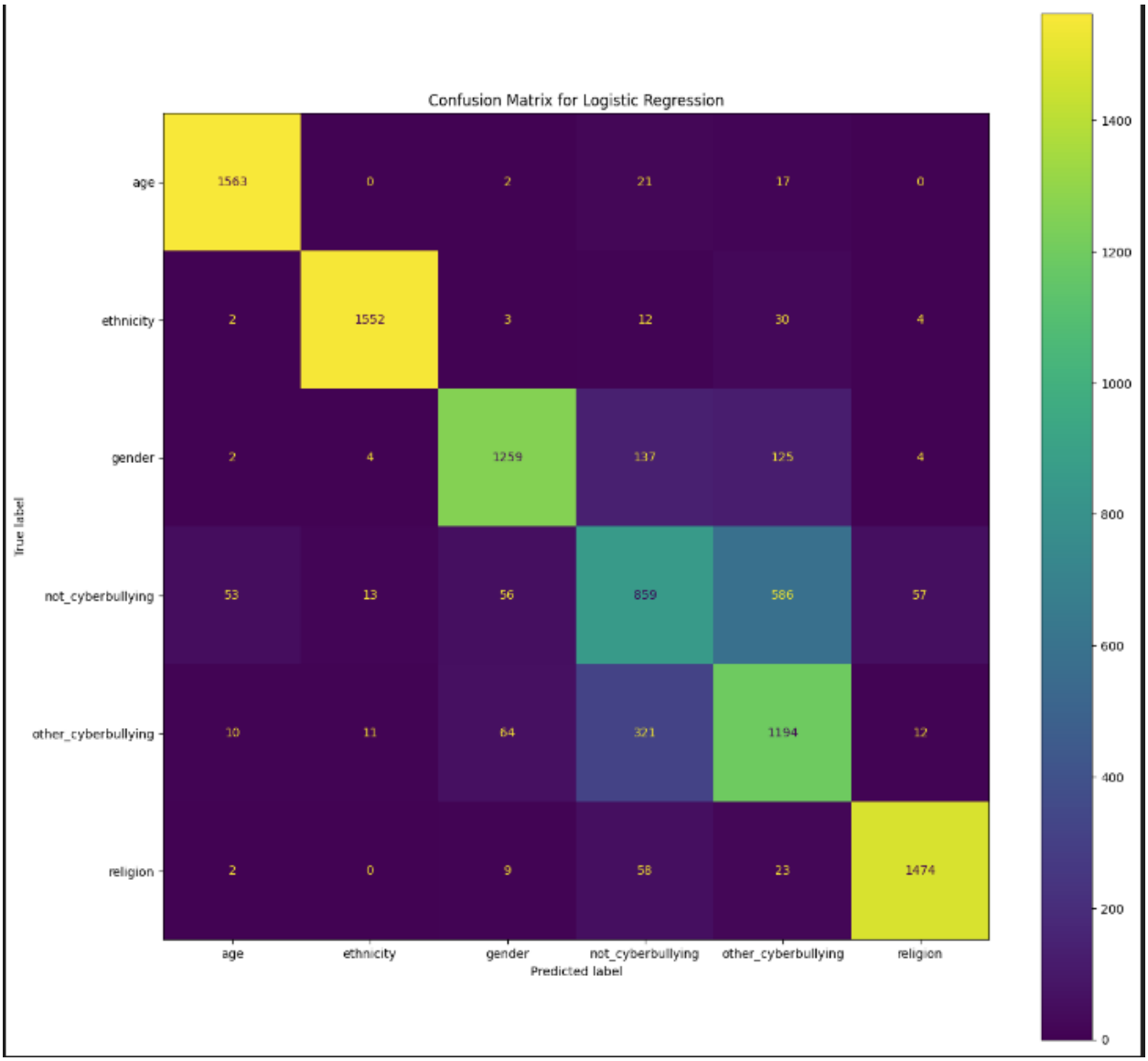

Figure 3: Confusion Matrix for Logistic Regression

Overall, these results suggest that the machine learning models are effective in predicting cy-

berbullying sentiment in tweets. However, it is important to note that the performance of the models may vary depending on the specific dataset and task, and further research is needed to determine the generalizability of these results to other contexts.

6 Conclusion and Future Work

In conclusion, this project explored the use of machine learning algorithms to predict cyberbullying sentiment in tweets. The LR model outperformed the other models in terms of accuracy, but all models achieved relatively high values for recall, precision, and F1 score, indicating that they are effective in predicting cyberbullying sentiment.

One limitation of this project is that it focused only on English-language tweets and may not be generalizable to other languages. Additionally, the dataset used in this project was limited to tweets labeled as cyberbullying and may not accurately capture the full range of online harassment and abuse.

Future work could explore the use of more advanced machine learning algorithms or the development of hybrid models that combine multiple algorithms to improve performance. Additionally, future work could focus on expanding the dataset to include a wider range of languages and types of online harassment. Finally, the use of deep learning models such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) could also be explored to further improve the performance of the models.

7 Contribution Chart

Table 2: Contribution of Team Members

Task/Sub-task	**Student ID**	**Commentary on contribution**
Implementation	01339552	Dataset Analysis and Classification Model Implementation
	02019997	Data Loading and Pre-processing
Report Writing	01339552	Abstract, Introduction, Methods, Results and Conclusion
	02019997	Related Works, Data and Evaluation
Slide Preparation	01339552	Provide necessary information to prepare slide to Raphael
	02019997	Prepare the whole presentation

References

[1] R. Suhas Bharadwaj, S. Kuzhalvaimozhi, and N. Vedavathi. A novel multimodal hybrid classifier based cyberbullying detection for social media platform. In Radek Silhavy, Petr Silhavy, and Zdenka Prokopova, editors, *Data Science and Algorithms in Systems*, pages 689–699, Cham, 2023. Springer International Publishing.

[2] Md Manowarul Islam, Md Ashraf Uddin, Linta Islam, Arnisha Akter, Selina Sharmin, and Uzzal Kumar Acharjee. Cyberbullying detection on social networks using machine learning approaches. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6, 2020.

[3] Xing Li, Zhanqi Xu, Fan Yang, and Yunbo Li. Multi-objective hybrid evolution with information entropy awareness for controller placement. In *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 135–140, 2021.

[4] Noviantho, Sani Muhamad Isa, and Livia Ashianti. Cyberbullying classification using text mining. In *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, pages 241–246, 2017.

[5] Hani Nurrahmi and Dade Nurjanah. Indonesian twitter cyberbullying detection using text classification and user credibility. In *2018 International Conference on Information and Communications Technology (ICOIACT)*, pages 543–548, 2018.

[6] Jason Wang, Kaiqun Fu, and Chang-Tien Lu. Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1699–1708, 2020.