Cyberbullying Detection Approach using Machine Learning Algorithm

Mohammed Shahadat Hossain Talukder

[a]*East Delta University, , Chittagong, 4213, Chittagong, Bangladesh*

1. Introduction

As the usage of social media grows across all age groups, it has become most people's primary means of everyday contact. Social media's common use has created an ideal environment for cyberbullying, which may harm anybody, anywhere, and at any time. As opposed to traditional bullying, the internet's anonymity makes it harder to avoid such personal attacks. The COVID-19 epidemic has worsened this issue, due to increased screen usage, decreased face-toface contact, and the closing of schools creating fears of an increase in cyberbullying incidents. UNICEF issued a warning in April 2020 in response to these risky situations. [1]

Experts describe cyberbullying as the use of digital technologies to constantly damage or harass people. However, because of its digital nature and the comparatively anonymous nature of online experiences, it is difficult to identify the imbalance in power, work, and harmful purpose that is typical of conventional bullying. This is a topic that is still being researched. The statistics on cyberbullying are concerning, with a considerable number of middle and high school children experiencing or witnessing cyberbullying, which has a variety of negative consequences, including worse academic performance and mental health difficulties.
[1]
Combating cyberbullying generally focuses on teaching people about internet safety, spotting warning signals, and giving counseling. Many states in the United States have criminal Consequences for cyberbullying, however, are limited to school-related incidents. Major social media sites have set up tools and passive reporting processes to combat cyberbullying, but None have introduced active anti-cyberbullying features. Given that 90% of cyberbullying situations go unreported, developing active detection technologies is critical. In this kind of situation, we aim to improve machine learning and deep learning models for detecting cyberbullying on social media platforms. This research is part of the larger topic of sentiment evaluation in NLP, which has a significant use in machine learning. Sentiment analysis entails transforming textual input into mathematical representations that future classifiers may use. Detecting cyberbullying, on the other hand, has different challenges, such as the requirement to recognize hidden details and hidden text in language, such as sarcasm, humor, and symbols. Because only vocabulary might not be sufficient for detecting cyberbullying, The effectiveness of NLP or natural language understanding models must be investigated. [2]

Because of the difficulties in getting balanced datasets and publicly available precisely labeled cyberbullying data, we are motivated to explore Dynamic Query Expansion. Class inequalities have also been found in previous cyberbullying datasets. Another concern is the lack of text content, which we want to overcome by capturing links between concepts using a graph structure. The development of a graph structure for social media embedded data, the collection of a balanced multiclass cyberbullying dataset, and the evaluation of multiple posts from embedding methods and classification models are among our contributions to this work. We thoroughly analyze our findings to establish baselines for future study.

2. Literature Review

Cyberbullying is a developing issue in the digital era, posing major hazards to people's mental and emotional health. To address this issue, robust detection systems must be developed. In recent years, researchers have investigated the application of deep and machine-learning approaches to detect cyberbullying situations in various online environments. This overview of the literature looks at much of the research that has contributed to cyberbullying detection, with a focus on the efficiency of deep learning and machine learning systems.

Several studies have investigated the use of deep learning algorithms to identify cyberbullying on various online platforms, notably social media. Iwendi et al.[3] give an empirical examination of deep learning algorithms for identifying insults in social comments, including Bidirectional Long Short-Term Memory, Gated Recurrent Units, Long Short-Term Memory, and Recurrent Neural Network (RNN). The results show that the BLSTM model performs better than other models in terms of accuracy and F1-measure scores, indicating its potential for effective cyberbullying detection.

Al-Ajlan and Ykhlef [4] present the CNN-CB method, which uses convolutional neural networks (CNN) with word embeddings to improve cyberbullying identification. The program achieves an outstanding accuracy of 95% by taking into account the semantics and meanings of terms in cyberbullying content, outperforming previous content-based identification approaches. This study emphasizes the significance of word

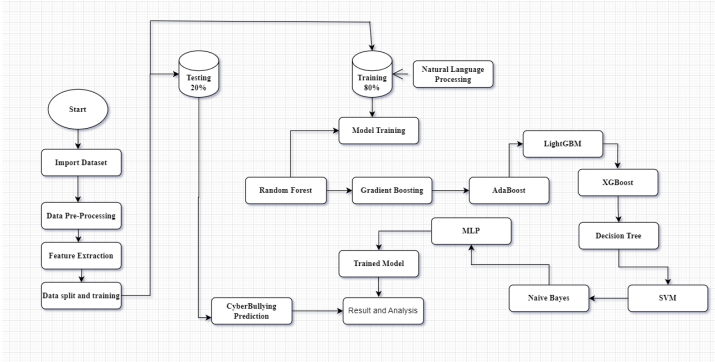

Figure 1: Proposed Methodology.

embeddings in increasing detection accuracy.

Al-Hashedi, Soon, and Goh [5] undertake an experimental study to measure the importance and effectiveness of deep learning methods such as GRU, LSTM, and BLSTM in combination with various word embedding models. Their findings show that when combined with ELMO word embeddings, the BLSTM model outperforms other algorithms in identifying cyberbullying texts. The report also emphasizes GRU's efficiency, which makes it a time-effective choice for detecting cyberbullying.

Banerjee, Telavane, Gaikwad, and Vartak [6]use convolutional neural networks (CNN) to develop a deep neural network-based solution for cyberbullying detection. Their suggested system, built with TensorFlow, achieves 93.97% testing accuracy. This method exhibits the capability of deep learning models in detecting cyberbullying content.

While deep learning models have received a lot of attention, machine learning methodologies have not been forgotten. Islam et al. [1] investigate the use of Bag of Words (BoW) and term frequency-inverse text frequency (TFIDF) characteristics with machine learning techniques to identify cyberbullies on social media. This method highlights the significance of feature engineering and algorithm selection in obtaining high accuracy.

3. Methodology

The research technique utilized in this study is an essential component that defines the systematic approach utilized to separate sentiment analysis in the context of cyberbullying identification in social media discourse. A strong and organized methodology is needed to ensure the study's depth, consistency, and reliability. The process of data collection, prepossessing methods, feature engineering, and the selection and training of machine learning models are all revealed in this section. These processes are all coordinated to understand the complicated sentiment landscape in textual data.

3.1. Data Collection & Preprocessing:

To obtain a large dataset for analysis, the research initiated an organized data collection process. The dataset, obtained from Kaggle, consists of tweets with sentiments related to cyberbullying labeled in them. The primary data source was the 'tweets.csv' file, which included classifications related to both text and cyberbullying. There are about 46017 tweet texts, and five types of cyberbullying are described in the dataset.

The goal of the preparation stage was to guarantee the dataset's consistency and integrity. So, we changed the feature of cyberbullying feature to a number which is numbered 0-4. After that, there were imbalances in the dataset, particularly with regard to "other_cyberbullying." This label's corresponding rows were eliminated in order to rectify class disparities. The column names were updated to improve clarity. 'tweet_text' became 'text' and 'cyberbullying_type' became sentiment. In order to make categorical sentiment labels compatible with machine learning algorithms, they underwent label encoding. To facilitate further modeling, the sentiment classes were numerically encoded.

3.1.1. Text cleaning:

The textual data was prepared for analysis by implementing a thorough text-cleaning procedure. The pipeline's functions included the following methods are used. The 'emoji' library was used to preserve emojis and translate them into text representations. [7] To improve text uniformity, contractions were extended to their whole forms.

	text	sentiment	text_clean
0	In other words #katandandre, your food was cra...	5	word katandandr food crapilici mkr
1	Why is #aussietv so white? #MKR #theblock #ImA...	5	aussietv white mkr theblock today sunris studi...
2	@XochitlSuckkks a classy whore? Or more red ve...	5	classi whore red velvet cupcak
3	@Jason_Gio meh. :P thanks for the heads up, b...	5	meh p thank head concern anoth angri dude twitter
4	@RudhoeEnglish This is an ISIS account pretend...	5	isi account pretend kurdish account like islam...

Figure 2: Cleaned Text.

Meaningful information could be extracted more easily when URLs, mentions, and special characters were removed. To guarantee accurate identification and content extraction, hashtags were processed. Underneath words, special characters like '&' and '$' were found and filtered. To make the text easier to read, several consecutive spaces were combined into one. To standardize word forms and lower dimensionality, the text was subjected to stemming and lemmatization.

3.1.2. Feature Extraction and Data analysis:

An exploratory data analysis step was carried out before the model was trained to obtain an understanding of the distribution of sentiments and text lengths within the dataset. The most common words connected to each sentiment class were shown using visualizations, such as word clouds. [8] [9]

Figure 3: Word Cloud.

The Term Frequency-Inverse Document Frequency (TF-IDF) approach was used to convert textual data into numerical features. The goal of this conversion was to accurately represent a word's significance in relation to the complete dataset. [10]

3.2. Model Selection and Training:

To tackle the sentiment analysis classification task, a variety of machine learning models were selected. These included Decision Tree, Support Vector Machine, Naive Bayes, Random Forest, Gradient Boosting, AdaBoost, LightGBM, XGBoost, and Multilayer Perceptron. All the models you have mentioned are classifiers; they operate by identifying patterns and relationships in the input data, in this case, text data, to predict the target variable, which is probably a binary label, in this case, designating whether or not a text passage is connected to cyberbullying. [11]

3.2.1. Application of Machine learning Models:

During the training of the cyberbullying dataset, RF creates an ensemble of decision trees. A random subset of features and a random subset of data samples are used to build each tree. A tree structure is produced by decision trees, which divide the data according to features. The result of a majority vote from all trees is the final prediction.

Gradient boosting constructs trees in a stepwise manner, fixing mistakes in each tree as it goes. Including weak learners reduces the loss of function in the tweet text dataset for cyberbullying classification. Fitting to the residuals of the previous tree creates new trees. These poor learners are combined to create a powerful predictive model. To produce a powerful classifier, ADA combines many weak ones. It gives each instance a weight, giving misclassified occurrences a higher weight. To increase overall accuracy, weaker models are iteratively blended and reweighted. The gradient-boosting framework LGB employs a learning strategy based on histograms. Building trees leaf-wise instead of level-wise lowers the total complexity. Large datasets are handled with efficiency, and their speed is well known.[11]

An enhanced gradient boosting algorithm is called XGB. To avoid overfitting, it employs a regularized model that effectively manages missing data in the tweet text dataset. To increase prediction performance, weak learners are added repeatedly. To categorize cases, DT makes judgments based on characteristics. It divides the data recursively, depending on the characteristic that best separates the classes. When a given requirement is fulfilled, it stops splitting, resulting in a tree structure.[11]

SVC determines the optimum hyperplane for separating instances of distinct classes. It uses a kernel function to translate input data to a high-dimensional space. It seeks to increase the difference between classes in cyberbullying classification. [12]

NB is based on Bayes' theorem and presupposes feature independence. Given the input features, it computes the probability of each class. It's very useful for text categorization jobs in the Cyberbully dataset.[12]

MLP is a form of neural network that consists of numerous layers of nodes (neurons). During training, it employs backpropagation to update weights and biases. It can capture complicated connections in data, although it may need design and hyperparameter customization.[**?**]

These models seek to learn patterns in tweet text to categorize instances of cyberbullying into one of five categories (religion, age, gender, ethnicity, and general cyberbullying). The effectiveness of each model is determined by factors such as hyperparameter tuning, dataset nature, and cyberbullying characteristics in text data.

3.2.2. Application of SMOTE:

The models were trained on pre-processed and feature-engineered data, with special emphasis on dealing with class imbalances. To correct for imbalances in the training set, the Synthetic Minority Over-sampling Technique (SMOTE) was used. [13]

4. Model Evaluation & Result :

After Preprocessing the data set of cyberbully classification with 46017 unique tweets and with the test set, the accuracy of nine different classifiers is assessed, including Random Forest, Gradient Boosting, AdaBoost, LightGBM, XGBoost, Decision Tree, Support Vector Machine, Naive Bayes, and Multilayer Perceptron. For each classifier, classification reports and confusion matrices are created, offering thorough insights into

precision, recall, and F1-score for each class. The accuracy ratings of the models are shown for easy comparison. Overall, this research seeks to select the most effective model based on performance measures for the given cyberbully categorization job.

Dataset	Training	Testing
Cyberbullying	36813	9204

Table 1: Dataset Distribution

Model	Precision	Recall	F1-Score	Accuracy
RandomForest	0.94	0.94	0.94	0.94
GradientBoosting	0.93	0.93	0.93	0.93
AdaBoost	0.91	0.91	0.91	0.91
LGBM	0.94	0.94	0.94	0.94
XGBoost	0.94	0.94	0.94	0.94
DecisionTree	0.92	0.92	0.92	0.92
SVC	0.93	0.92	0.92	0.93
MultinomialNB	0.86	0.84	0.84	0.85
MLP	0.90	0.90	0.90	0.90

Table 2: Model Comparison

4.1. Result

The RandomForestClassifier, LGBMClassifier, XGBClassifier [14] [15], and DecisionTreeClassifier all produced accuracies of about 94% and showed the best performances overall across a variety of parameters, including precision, recall, and F1-score. These models performed exceptionally well at achieving a balance between accurately classifying cases from various cyberbullying datasets. While displaying somewhat lesser accuracy, the GradientBoostingClassifier still demonstrated strong performance with an accuracy of 93% with precision of 93%.

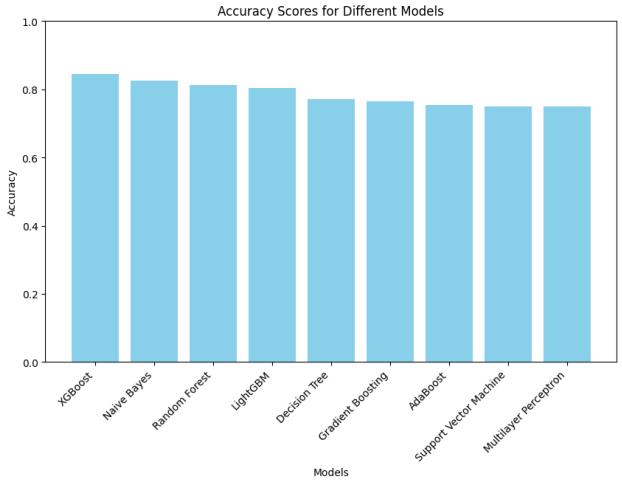

Figure 4: Comparison between models

On the other hand, models like SVC, AdaBoostClassifier, MultinomialNB, and MLPClassifier obtained somewhat lower accuracy, ranging from 85% to 93% accuracy with Precision ranging from 86% to 91%. These models showed mixed strengths and limitations in terms of accuracy, recall, and F1-score, with some finding it difficult to handle particular classes—all while still producing results that were quite acceptable and could be used in real-world scenarios.

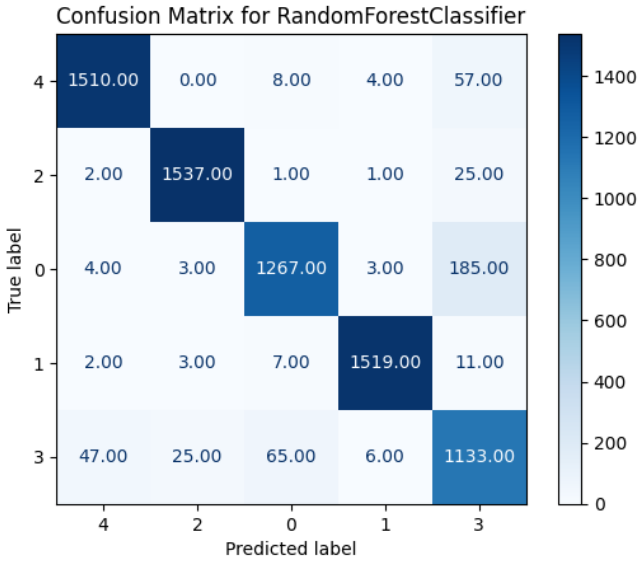

Figure 5: Confusion matrix of random forest classifier

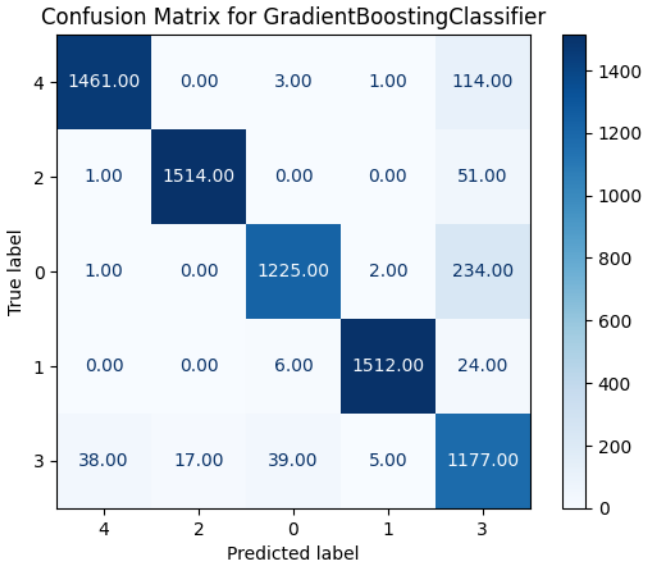

Figure 6: Confusion matrix of Gradient Boost

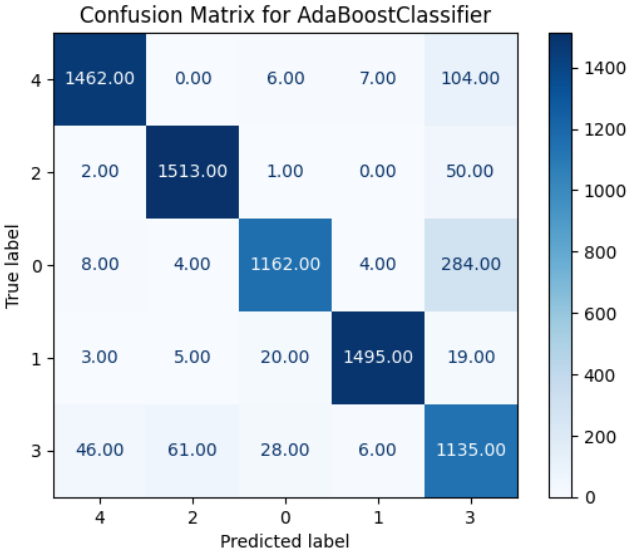

Figure 7: Confusion matrix of Adaboost

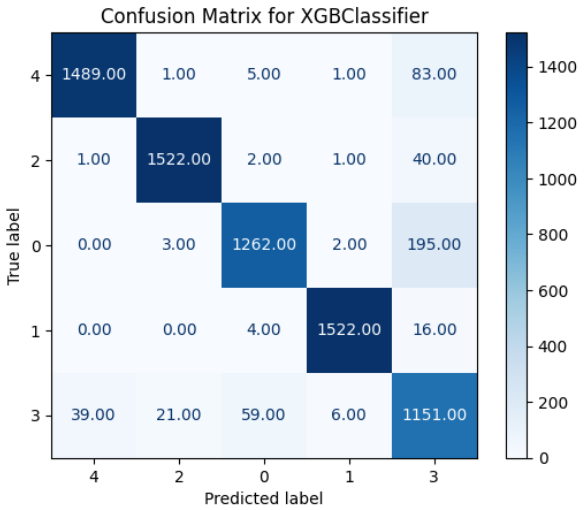

Figure 9: Confusion matrix of XGBoost

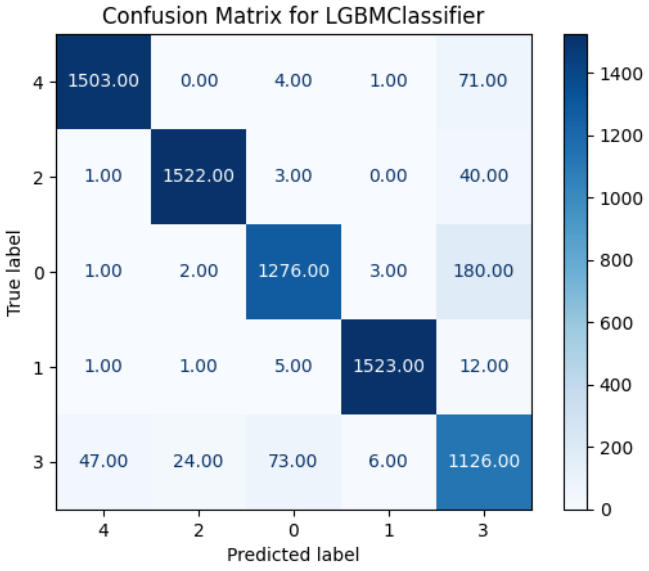

Figure 8: Confusion matrix of LightGBM

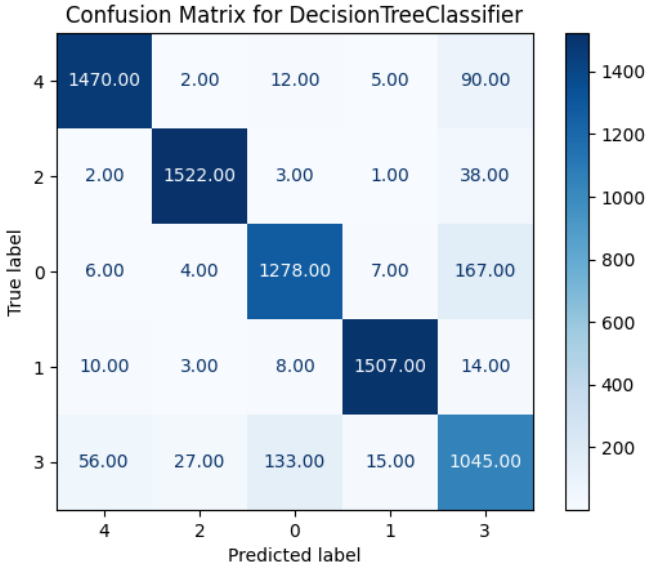

Figure 10: Confusion matrix of Decision Tree

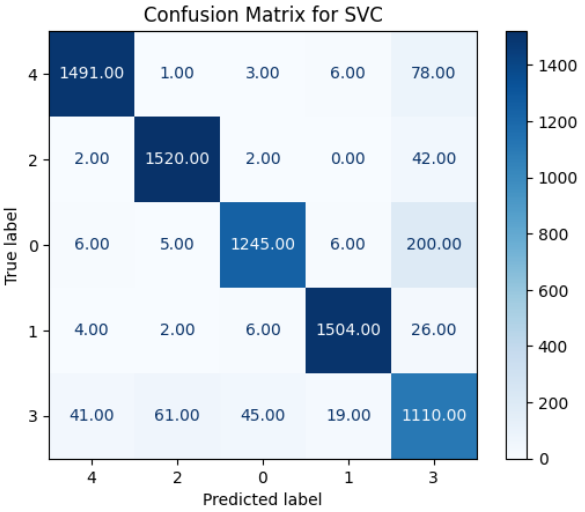

Figure 11: Confusion matrix of Support Vector Machine

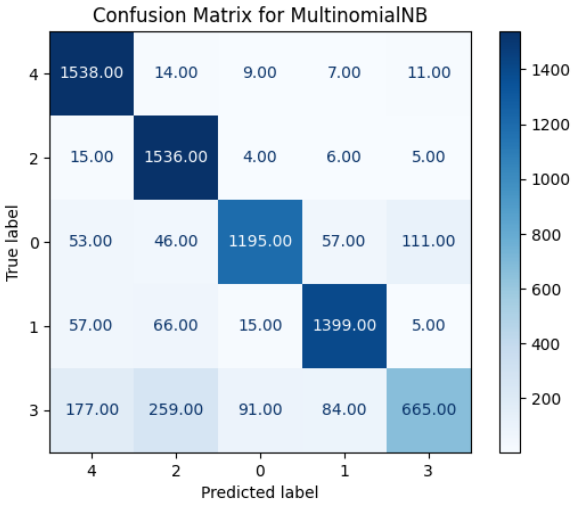

Figure 12: Confusion matrix of Naive Bayes

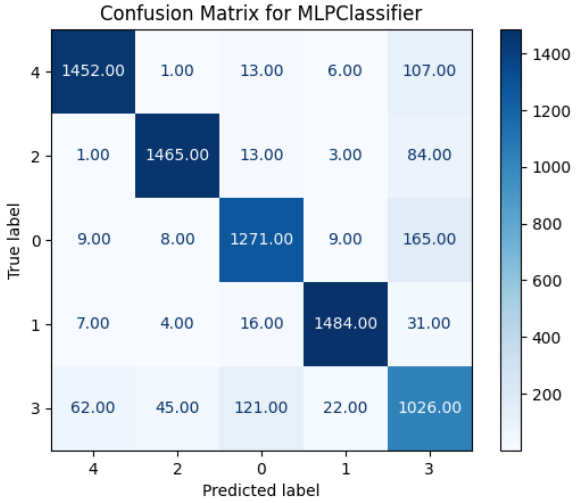

Figure 13: Confusion matrix of Multilayer Perceptron

5. Conclusion & Future Scope

In summary, the analysis of sentiment was used in this work to identify instances of cyberbullying in social media conversations using a systematic method. The study made use of a sizable dataset of tweets analyzed for feelings associated with cyberbullying that were sourced from Kaggle. The preparation and data collection stages corrected imbalances and improved clarity while ensuring the dataset's consistency and integrity. After using text-cleaning techniques, TF-IDF was used to extract features. Many machine learning models were assessed for sentiment analysis categorization, including Random Forest [16], Gradient Boosting, AdaBoost, LightGBM, XGBoost, Decision Tree, Support Vector Machine, Naive Bayes, and Multilayer Perceptron. During the training and testing phases, the Simulated Minority Oversampling Technique (SMOTE) was utilized to target class imbalances.[17] The results showed that the Random Forest, LGBM, XGBoost, and Decision Tree models all outperformed the competition, with an accuracy rate of about 94%. But it's important to recognize the study's shortcomings, which might affect how well the models apply in real-world situations.

This work's future research directions should focus on improving its achievements and resolving its shortcomings. In addition to improving model generalizability through more hyperparameter tweaking and adaptability to changing slang, expanding the dataset with varied linguistic and cultural settings may help reduce bias. Investigating multimodal methods that combine text analysis with visuals and emotions [18] may provide a more profound understanding of cyberbullying identification.[19] In the end, incorporating these developments into ethically sound real-world platforms may result in strong defenses against cyberbullying in online communities.[20]

References

[1] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, U. K. Acharjee, Cyberbullying detection on social networks using machine learning approaches, in: 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1–6. doi:10.1109/CSDE50874.2020.9411601.

[2] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, M. Prasad, Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques, Electronics 10 (22) (2021) 2810.

[3] C. Iwendi, G. Srivastava, S. Khan, P. K. R. Maddikunta, Cyberbullying detection solutions based on deep learning architectures, Multimedia Systems (2020) 1–14.

[4] M. A. Al-Ajlan, M. Ykhlef, Deep learning algorithm for cyberbullying detection, International Journal of Advanced Computer Science and Applications 9 (9) (2018).

[5] M. Al-Hashedi, L.-K. Soon, H.-N. Goh, Cyberbullying detection using deep learning and word embeddings: An empirical study, in: Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems, 2019, pp. 17–21.

[6] V. Banerjee, J. Telavane, P. Gaikwad, P. Vartak, Detection of cyberbullying using deep neural network, in: 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), IEEE, 2019, pp. 604–607.

[7] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, U. Kaymak, Exploiting emoticons in sentiment analysis, Proceedings of the 28th Annual ACM Symposium on Applied Computing (2013) 703–710.

[8] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, arXiv preprint arXiv:1408.3608 (2014) 1–9.

[9] S. Hinduja, J. W. Patchin, Cyberbullying: An exploratory analysis of factors related to offending and victimization, Deviant Behavior 29 (2) (2008) 129–156.

[10] S. Qaiser, R. Ali, Text mining: use of tf-idf to examine the relevance of words to documents, International Journal of Computer Applications 181 (1) (2018) 25–29.

[11] S. Rahman, M. Irfan, M. Raza, K. Moyeezullah Ghori, S. Yaqoob, M. Awais, Performance analysis of boosting classifiers in recognizing activities of daily living, International journal of environmental research and public health 17 (3) (2020) 1082.

[12] A. Talaviya, Cyberbullying detection using topic modeling and sentiment analysis, Medium 22 (12) (2020) 69–78.

[13] B. A. Talpur, D. O'Sullivan, Cyberbullying severity detection: A machine learning approach, PloS one 15 (10) (2020) e0240924.

[14] G. Ke, Q. Meng, Q. Yang, T. Finley, Lightgbm: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, p. 3146.

[15] T. Chen, C. Guestrin, Xgboost: A scalable parallel gradient boosting tree for scigiffi and non-scigiffi applications 785.

[16] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[17] K. B. L. O. H. Chawla, Nitesh V., W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, Applied intelligence 16 (3) (2002) 185–202.

[18] A. Ghosh, D. Roy, D. Sarkar, D. Sarkar, P. Sarkar, Emotion analysis for cyberbullying detection in text and speech, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Vol. 2, Association for Computational Linguistics, 2022, pp. 5094–5106.

[19] Z. Zhang, Y. Wu, M. Zhou, W. Zhao, Y. Guo, J. Xu, L. Xie, S. Ma, Multimodal cyberbullying detection in social media using text, emojis, and images, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 8442–8456.

[20] S. Zuboff, The Social Dilemma: Algorithmic bias, privacy, and the future of online communities, One World/Penguin Random House, 2019.