Classification and Categorization of Cyberbullying across Social media platforms

Aditi Narendran
Computer Science
Arizona State University
Tempe AZ USA
anarend4@asu.edu

Anjana Ouseph
Computer Science
Arizona State University
Tempe AZ USA
aouseph@asu.edu

Muskan Mehta
Computer Science
Arizona State University
Tempe AZ USA
mmehta33@asu.edu

Vaishnavi Amirapu
Computer Science
Arizona State University
Tempe AZ USA
vamirap1@asu.edu

Vidushi Raturi
Computer Science
Arizona State University
Tempe AZ USA
vraturi1@asu.edu

ABSTRACT

With the increased use of social media platforms in recent years, it has become a platform for bullying and online harassment. CyberBullying, the act of bullying through digital and online communication channels has been a pertinent problem and, thereby needs to be addressed. Machine Learning (ML) can be used to identify patterns, language usage, and behavioral cues indicative of cases of CyberBullying by analyzing vast amounts of textual and visual data. This project is dedicated to utilizing a comprehensive cross-platform CyberBullying dataset to conduct a thorough evaluation of multiple ML classification models, with the objective of distinguishing between abusive and non-abusive content while categorizing the abusive content into specific contexts such as sexism, racism, toxicity, and more. The research uses a comprehensive and organized methodology that includes feature extraction techniques, rigorous preprocessing methods, and data collecting from many social media networks. Then, a variety of ML models are trained and evaluated in order to automate the recognition and accurate classification of CyberBullying content on the internet. The urgent need for potent solutions to stop online harassment and shield users from the damaging effects of CyberBullying is what spurs this research.

KEYWORDS

Machine Learning, Natural Language Processing, Cross-platform dataset, Text Classification, CyberBullying

1 Introduction

The field of Data Science is being driven by Big Data in performing predictive analytics in order to address the problems in the real world. Through improved data collection processes, enhanced computational technology, and innovative data processing techniques, this branch of science is constantly advancing. Widely used Machine Learning models are implemented in a wide range of areas such as Education, Health Care, e-Commerce, e-Learning, and social media, etc.

The principles of Data Mining are used to find correlations, trends, and important information to extract patterns and insights from big datasets. The purpose of this project is to explore the efficacy of various ML classification models to distinguish abusive language from non-abusive content across various social media platforms such as Twitter, YouTube, Instagram, etc. The project also aims at categorizing the abusive content into particular contexts like racism, sexism, toxicity and more.

In this project, we also focus on performing an empirical evaluation of various classification models in order to explore their full potential in classifying abusive content and categorizing them. In this project, dataset from various social-media platforms is taken to identify the patterns of abusive content in CyberBullying. Since the dataset deals vastly with textual data, text-classification techniques using Natural Language Processing (NLP) will be employed in order to process the data before using the ML models. Various ML models such as the VM, Random Forest to classifiers such as Gradient Boost and Transformer models such as BERT and GPT-3 are implemented in classification and evaluation for this project. Essential NLP techniques such as Bag-of-Words (BoW), TF, TF-IDF are used as a part of the Data Processing methods.

2 Related Work

In recent years, CyberBullying Detection has seen a variety of approaches, including the construction of datasets [2], machine learning methods [3, 4], deep learning and language models [5], text mining [9], social network analysis [10], and ensemble learning [12]. Following this, the research papers in this review are organized into groups according to the implementation methods used and the types of datasets utilized for detection.

The papers researched for our literature review collectively explore various methodologies and techniques to tackle the challenging issue of CyberBullying Detection. They extensively evaluate various feature vectors [11], shedding light on their effectiveness in combating CyberBullying on social media platforms. These studies delve into applying advanced deep learning models, such as BERT [5], and present the strengths and limitations of these cutting-edge approaches in addressing CyberBullying concerns. Furthermore, they center their investigations on text mining and NLP [6, 9], leveraging various methods ranging from SVM and Naive Bayes to neural networks for detecting CyberBullying content [3, 4]. These papers comprehensively explore varied text-based techniques for identifying CyberBullying while candidly discussing their merits and drawbacks [7, 8].

One fundamental aspect of CyberBullying detection is the availability of high-quality datasets for training and evaluation. Multiple Social-Media Platform Datasets, which collect data from various online platforms, provide a comprehensive perspective on the issue of CyberBullying [2]. They are a valuable resource for researchers looking to develop versatile models capable of addressing CyberBullying across multiple platforms. These datasets help identify common patterns and differences in CyberBullying behaviors, enabling the creation of generalized detection solutions [8]. In contrast, Single Social-Media Platform Datasets focus exclusively on one platform, offering in-depth insights but limited scope [6]. The strength of Multiple Social-Media Platform Datasets lies in their ability to nurture versatile detection models and facilitate knowledge transfer between platforms, making them a preferred choice for our work [2, 8].

3 Dataset

3.1 Data Collection

The dataset used in this work is a combination of various datasets gathered from various sources [1], all with the unifying topic of automatically detecting CyberBullying. The data comes from a number of social media sites, such as YouTube, Wikipedia Talk pages, Twitter and Kaggle. Each data item is made up of text samples that have been marked as either being devoid of CyberBullying or including instances of it. The dataset covers a wide range of CyberBullying categories, including insults, toxicity, aggression, hate speech, and hate speech.

Altogether 448,880 instances, spread across eight different CSV files, make up the merged dataset, which was carefully selected from a variety of sources. There are 44 features in all, however, some may be removed in preprocessing to improve the effectiveness of the analysis that follows.

3.2 Data Pre-processing

3.2.1 Class Distribution

This refers to the number of instances (data entries or samples) available for each class or category in the dataset. In this context, the classes are related to types of online content:

1. Abusive or Non-Abusive: This classifies content as either being abusive (containing harmful or negative elements) or non-abusive.
2. Subcategories of Abusive Content: This further breaks down the "Abusive" category into specific types, namely:
    a. Sexism: Discrimination or prejudice based on gender.
    b. Racism: Discrimination or prejudice based on race.
    c. Hate Speech: Communication that discriminates against or promotes violence towards a group based on attributes like race, religion, ethnic origin, etc.
    d. Toxicity: Content that is harmful, poisonous, or very unpleasant.

3.2.2 Dataset Splits

This refers to how the data is partitioned for the purpose of training and testing machine learning models. K-Fold Stratification with an 80:20 split is implemented. This means the dataset is split using stratified sampling into K subsets (or "folds"). One-fold is used for validation while the remaining are used for training. This process is repeated K times, using each fold as the validation set once. The 80:20 ratio indicates that 80% of the data is used for training while the remaining 20% is used for validation or testing.

3.2.3 Pre-processing Techniques

The following steps were taken in order to clean and prepare the dataset before implementing the ML models:

1. Outlier Removal: This involves identifying and removing extreme values that deviate significantly from other observations. These can skew or mislead the training process.
2. Handling Missing and Null Values: This ensures that the dataset does not have any missing or undefined data points, which can cause errors during model training.
3. Removing Duplicates: Duplicate entries can bias the model, so they are identified and removed.
4. Performing Normalization on Numerical Values: This process scales numerical features to a standard range (e.g., 0 to 1), ensuring that they have the same scale and hence contributing equally to the model's performance.

4 Methods

4.1 Natural Language Processing

Natural Language Processing (NLP) is essential for handling real-world text data, which often contains irrelevant elements like numbers and punctuation, not pertinent to bullying detection. To effectively apply machine learning algorithms to comments, a preprocessing step is crucial. This phase encompasses various tasks such as eliminating irrelevant characters (stop-words, punctuation, and numbers), tokenization, stemming, and more.

Following the preprocessing, two vital text features are prepared as follows:

1. Bag-of-Words (BoW): Machine learning algorithms cannot work directly with raw text, so the processed data is transformed into a Bag-of-Words representation, where text is converted into numerical vectors for subsequent analysis.
2. TF-IDF (Term Frequency-Inverse Document Frequency): TF-IDF is another feature considered in our model. It's a statistical measure that assesses the relevance of a word to a document within a collection of documents. Unlike BoW, TF-IDF assigns higher importance to words that occur more frequently, as they are more informative for classification.

The Machine Learning module entails the application of various machine learning approaches, including SVM, Random Forest, and Gradient Boosting Classifier, for the detection of bullying messages and text. The classifier with the highest accuracy is identified for a specific public CyberBullying dataset. In the following section, we delve into common machine-learning algorithms used for detecting CyberBullying in social media texts.

4.2 Machine Learning Models

4.2.1 Support Vector Machine

Support Vector Machine (SVM) is a powerful classification algorithm that aims to find an optimal hyperplane in the feature space, maximizing the margin between the bullying and non-bullying classes. SVM can effectively work with high-dimensional data, such as text, by representing it as numerical features like TF-IDF vectors or Bag-of-Words. SVM's goal is to establish the best decision boundary to separate bullying from non-bullying text, making it a strong contender for bullying detection tasks.

4.2.2 Random Forest

Random Forest, on the other hand, is an ensemble learning method that combines multiple decision trees to make predictions. It's particularly suitable for text classification because it can handle high-dimensional data effectively and mitigate overfitting. In the context of bullying detection, Random Forest constructs a forest of decision trees, each trained on a subset of the data and a random subset of features. The final prediction is made through a voting or averaging mechanism, ensuring robustness and the ability to capture complex relationships in text data.

4.2.3 Gradient Boosting Classifier

The Gradient Boosting Classifier is yet another powerful ensemble learning technique for text classification. It builds an ensemble of weak learners, typically decision trees, in an iterative fashion. Each new tree focuses on the mistakes made by the ensemble of trees so far, resulting in a strong learner. This approach is beneficial for bullying detection as it learns from the errors of previous models, continually improving overall performance.

4.2.4 BERT & GPT3

Using BERT and GPT-3 for bullying detection has several advantages. BERT's bidirectional architecture allows it to understand the context in which words and phrases appear, making it highly effective in discerning subtle nuances of bullying language. While GPT-3 is primarily designed for text generation, it can also be used for bullying detection. The approach involves using GPT-3 to generate text that responds to a given input, such as a social media comment and then analyzing the generated response to identify any bullying content. These models can adapt to a wide range of text data and are capable of understanding the context and subtleties of language, which is crucial for identifying bullying content that may not rely solely on explicit keywords. Moreover, they can be used to process and analyze text in real time, making them valuable tools for monitoring and moderating online conversations, forums, and social media platforms.

5 Evaluation and Results

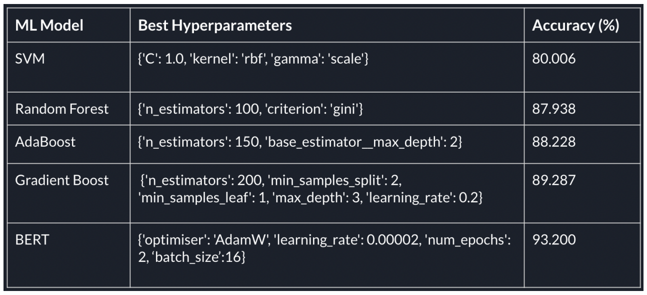
We aim to assess how effectively the aforementioned ML models perform in terms of recognizing and categorizing cyberbullying on various social media platforms. Our evaluation will separate abusive from non-abusive information using basic criteria like accuracy, precision, recall, and F1 score. Furthermore, we seek to quantify the recall and precision of our models for every distinct Cyberbullying “Abusive” category, including racism, sexism, and toxicity etc. We intend to assess the models' adaptability across a variety of datasets using cross-validation approaches such as k-fold validation to provide a robust evaluation. It will be possible to combine the strengths of separate models to produce a final classification system that is more reliable and accurate by combining the findings from several models through ensemble methods, stacking, or meta-classification approaches. The ultimate goal of this course project is to empirically evaluate and analyze the efficacy of various ML classification models in

Cyberbullying detection and categorizing the abusive content into the mentioned classes.

We conducted an empirical evaluation of various ML classification models for the development of a robust system dedicated to CyberBullying Detection across multiple social media platforms. Unlike existing studies that predominantly utilized datasets from a single social media platform, our project addresses this limitation by leveraging a dataset spanning diverse social media platforms such as Twitter, Wikipedia Talk Pages, Youtube and Kaggle. The models under evaluation encompass a spectrum from traditional methods such as Support Vector Machines (SVM), Random Forest, AdaBoost Classifier, Gradient Boost Classifier and BERT. This approach aims to enhance the generalizability and adaptability of the CyberBullying Detection system, considering the varied patterns and behaviors exhibited across different social media environments.

These models were trained with a 5-fold cross-validation splitting technique in order to build robust accuracies for the models on the training dataset (Twitter and Wikipedia Talk Pages) .We also performed Hyperparameter tuning in order to find the best set of hyperparameters for each of the models that would generate higher training and validation accuracies on the training dataset.

The following table depicts the models, along with their best hyperparameters and their best accuracies.



ML Model	Best Hyperparameters	Accuracy (%)
SVM	{'C': 1.0, 'kernel': 'rbf', 'gamma': 'scale'}	80.006
Random Forest	{'n_estimators': 100, 'criterion': 'gini'}	87.938
AdaBoost	{'n_estimators': 150, 'base_estimator__max_depth': 2}	88.228
Gradient Boost	{'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 3, 'learning_rate': 0.2}	89.287
BERT	{'optimiser': 'AdamW', 'learning_rate': 0.00002, 'num_epochs': 2, ‘batch_size’:16}	93.200

Figure 1: Best Hyperparameters & Best Accuracies

The following bar chart visually plots the accuracies of the different classification models.

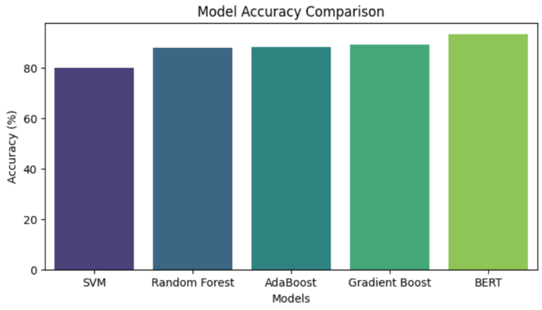
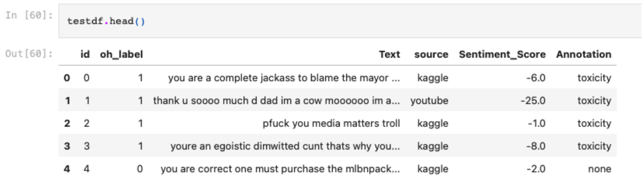


Figure 2: Accuracy v/s Models chart

The BERT model that depicted higher accuracy among the others was used to classify and categorize the test dataset into racism, sexism, toxicity and none (Youtube and Kaggle). Below is the first 5 rows of the categorized data (Youtube and Kaggle data combined).



```
In [60]: testdf.head()

Out[60]:
     id  oh_label                                            Text   source  Sentiment_Score  Annotation
  0   0         1      you are a complete jackass to blame the mayor ...   kaggle             -6.0    toxicity
  1   1         1  thank u soooo much d dad im a cow moooooo im a...  youtube            -25.0    toxicity
  2   2         1                      pfuck you media matters troll   kaggle             -1.0    toxicity
  3   3         1      youre an egoistic dimwitted cunt thats why you...   kaggle             -8.0    toxicity
  4   4         0    you are correct one must purchase the mlbnpack...   kaggle             -2.0        none
```

Figure 3: First 5 rows of the classified test dataset

6 Discussions and Conclusion

In conclusion, our empirical evaluation of the CyberBullying detection models substantiates their effectiveness through a thorough analysis of results on the test dataset. The comparative analysis against baseline models and existing techniques from the literature review across various social media platforms reinforces the novelty and competitive edge of our approach. However, the interpretation of results reveals a crucial insight - negative words alone may not suffice for contextual connotation, emphasizing the significance of considering broader context in differentiating abusive and non-abusive statements. Notwithstanding these achievements, our project encountered challenges, notably with non-English characters in the dataset requiring careful pre-processing. Additionally, limitations in computational resources restricted the handling of large datasets and memory-intensive models. Despite these shortcomings, our work lays a solid foundation for advancing CyberBullying detection methodologies and underscores the importance of nuanced contextual analysis in addressing online abuse.

7 Future Work

In future endeavors, incorporating advanced tools and methodologies will enhance the complexity and effectiveness of systems aimed at identifying and managing aggressive behavior on social media platforms. Utilizing sophisticated deep learning techniques, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), can further elevate the accuracy of aggressive categorization. A key focus should be on increasing granularity in aggression classification, discerning between specific categories such as political, religious, hate speech, and other forms of bullying. This nuanced approach enables more sophisticated and targeted intervention techniques. Moreover, leveraging high-performance computing resources to process larger datasets expeditiously ensures the scalability and robustness of the model to handle substantial amounts of social media data. Additionally, integrating multiple social media sources like Instagram and Pinterest can broaden the model's adaptability across diverse online communication channels, enhancing its overall effectiveness.

REFERENCES

[1] Elsafoury, Fatma. 2020. Cyberbullying datasets, Mendeley Data, V1, doi: 10.17632/jf4pzyvnpj.1

[2] David Van Bruwaene, Qianjia Huang, and Diana Inkpen. 2020. A multi-platform dataset for detecting CyberBullying in social media. *Language Resources and Evaluation* 54, 4 (2020), 851-874. DOI:https://doi.org/10.1007/s10579-020-09488-3

[3] Md Manowarul Islam, Md Ashraf Uddin, Linta Islam, Arnisha Akter, Selina Sharmin, and Uzzal Kumar Acharjee. 2020. CyberBullying Detection on Social Networks Using Machine Learning Approaches. *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (2020). DOI:https://doi.org/10.1109/csde50874.2020.9411601

[4] Amgad Muneer and Suliman Mohamed Fati. 2020. A Comparative Analysis of Machine Learning Techniques for CyberBullying Detection on Twitter. *Future Internet* 12, 11 (2020), 187. DOI:https://doi.org/10.3390/fi12110187

[5] Sayanta Paul and Sriparna Saha. 2020. CyberBERT: BERT for CyberBullying identification. *Multimedia Systems* 28, 6 (2020), 1897-1904. DOI:https://doi.org/10.1007/s00530-020-00710-4

[6] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. *Natural Language Processing and Information Systems, 23rd International Conference on Applications of Natural Language to Information Systems, 23,* (Jun. 2018), 57-64 57-64. DOI:https://doi.org/10.1007/978-3-319-91947-8_6

[7] Dan Ottoson. 2023. CyberBullying Detection on Social Platforms using Large Language Models. Bachelor’s Thesis. Mid Sweden University.

[8] Peiling Yi and Arkaitz Zubiaga. 2022. CyberBullying Detection across Social Media Platforms via Platform-Aware Adversarial Encoding. *Proceedings of the International AAAI Conference on Web and Social Media 16,* (2022), 1430-1434. DOI:https://doi.org/10.1609/icwsm.v16i1.19401

[9] Noviantho, Sani Muhamad Isa, and Livia Ashianti. 2017. CyberBullying classification using text mining. *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)* (2017). DOI:https://doi.org/10.1109/icicos.2017.8276369

[10] Anqi Wang and Katerina Potika. 2021. CyberBullying Classification based on Social Network Analysis. *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)* (2021). DOI:https://doi.org/10.1109/bigdataservice52369.2021.00016

[11] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. 2016. Cybercrime detection in online communications: The experimental case of CyberBullying detection in the Twitter network. *Computers in Human Behavior 63,* (2016), 433-443. DOI:https://doi.org/10.1016/j.chb.2016.05.051.

[12] Kazi Saeed Alam, Shovan Bhowmik, and Priyo Ranjan Kundu Prosun. 2021. CyberBullying Detection: An Ensemble Based Machine Learning Approach. *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (2021), 710-715. DOI:https://doi.org/10.1109/icicv50876.2021.9388499.

APPENDIX

Link to the GitHub source code -
https://github.com/vaamps/cyberbullying-detection