ORIGINAL RESEARCH ARTICLE

SAMRIDDHI Volume 17, Issue 1, 2025 Print ISSN: 2229-7111 Online ISSN: 2454-5767

Filtering Online Harassment: ML based Cyberbullying Detection

Ayushman Singh[1], Hari Om Shankar Mishra[2], Sumit Kumar Jha[3]

[1,2]Electronics and Communication Engineering Department, MNNIT Allahabad, Prayagaj, India.

[3]Electronics and Communication Engineering Department, MNNIT Allahabad, India.

Abstract

Cyber bullying has emerged as a great threat to the people on the internet.The platform, which was made for good use, is being used by some to harass people. This is actually a misuse of a great invention. Also, the nature of social media is such that these things spread very quickly due to the online communications. Many times, this spreads anonymously. Manual way of detection of cyber bullying will be very inefficient and a lot of time consuming.Thus, an automated cyber bullying detection using machine learning will come to help. This paper explores the effectiveness of various machine learning algorithms in classifying the tweets into different types of cyber bullying, including age based, ethnicity based, gender based, religious based and non cyber bullying. This automated detection using machine learning will offer a great approach to mitigate these bullying and its after effects by identifying and isolating the harmful texts and messages in real time.

Keywords: Cyber bullying, Machine learning, Multinomial Naïve Based, Logistic Regression, Support Vector Machine.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2025); DOI: 10.18090/samriddhi.v17i01.01

Introduction

The rapid use of social media applications and websites and digital platforms have made a huge revolution of how the users share the information and interact with each other. However, as every coin has two sides, this digital revolution has also given rise to a very serious issue of cyber bullying. Defined as the digital use of communication to bully, harass, or disturb individuals. It can be of various means like derogatory comments, threats and the spread of fake information. The difference between traditional bullying and cyber bullying is that cyber bullying can occur at any time and place, which can eventually lead to emotional and psychological distress for victims.

Solving the problem of detecting cyber bullying is difficult, it is because the nature of the crime and the very high volume of user generated content. Manual detection and monitoring are very much impractical given the vast amount of data generated on social media platforms like Twitter, FB, Instagram etc. This there is a need for automatic detection, therefore study needs to be done in this area to develop the automated systems capable of detecting and categorizing cyber bullying in real time allowing in time detection and mitigation.

ML will offer a very better solution to the challenge; it will enable automatic detection of cyber bullying by investigation of text and classifying it into predefined categories. By the help of ML algorithms, pattern can be identified and the bullying can be predicted which will definitely provide a scalable and effective approach to monitor online interactions.We shall focus on applying the ML techniques to classify tweets into various types of cyber bullying, including age based, ethnicity based, gender based, religion based and other forms of cyber bullying and non cyber bullying.

Traditional approaches to detect this have relied upon keyword and lexicon-based approaches. These methods involved matching the text against a list of offensive words, if there exists any content that is offensive, it will show a red flag. However, these can miss the case of indirect bullying. For instance, a comment using sophisticated language or even a sarcasm might be missed despite its harmful intention. Recent advancements in Natural Language Processing and ML have created a way for more better detection methods. Techniques such as sentiment analysis, feature extraction and deep learning have given improved accuracy because they consider the context, sentiment and linguistic patterns also in the text. However, these methods also introduce challenges

Corresponding Author: Hari Om Shankar Mishra, Electronics and Communication Engineering Department, MNNIT Allahabad, Prayagaj, India., e-mail: hari.2021rel05@mnnit.ac.in

How to cite this article: Singh, A., Mishra, H.O.S., Jha, S.K. (2025). Filtering Online Harassment: ML based Cyberbullying Detection. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* 17(1), 1-6.

Source of support: Nil

Conflict of interest: None

related to data requirements and other complexities.

Authors in[1] investigate the role of empathy in encouraging the young and adolescent to support the victims of cyber bullying. Study shows the importance of decision support systems.[2] explores various ML algorithms like Support Vector Machine and NLP to detect cyber bullying. The study has focused on challenges of processing of data which are multi lingual like Hinglish, i.e. the mixture of English and Hindi which is a popular way people in India communicate. The different ML approaches were compared in[3] and the study evaluates the effectiveness of Bag of Words methods and tokenization. [4] studies a risk mitigation approach for mobile cyber bullying through digital forensic readiness. Their study addresses the challenge of cyber bullying on mobile platforms.

[5]studies data analysis techniques to identify and predict cyberbullying incidents. They have analysed user behavior on social media sites to understand patterns of cyber bullying. [6] has studied the rapid detection method for cyber bullying using BERT model. The study focuses on effectiveness of transfer learningand focal loss in improvingdetection accuracy. [7] discusses the application of ML algorithms for detecting the cyberbullying on social media digital platforms. They have specifically highlighted the role of exploratory data analysis (EDA) in enhancing the prediction accuracy. The authors in[8] propose an AI method for detecting cyberbullying. The study used LIME(Local Interpretable Model-agnostic Explanations) to provide transparency in ML decisions. Authors in[9] uses NLP and text analytics to detect cyberbullying. The paper works on the use of Python and Twitter API for real time analysis of data.[10] presents a ML approach for detecting cyberbullying using various classifiers, including Naïve Bayes. Their study has focused mainly on supervised learning techniques for effective detection. Authors in[11] introduce a hybrid classier combining deep learning and machine learning methods for cyber bullying detection on social media. The paper presents experimental results demonstrating the classifier's effectiveness and accuracy.

[12]focuses on content-based cyber bullying detection using machine learning and deep learning techniques. The study evaluates various methods for effective detection on social media networks.[13] has studied the unsupervised learning model approach for detecting cyberbullying in social networks. The study uses techniques like self-organizing maps and sentiment analysis for detection. The paper shows good results.[14] has worked on creating a system called Cybersafe for detecting the cyber bullying, the study employs convolutional neural networks and SVM machine for text and image processing. Authors in[15] have explored the use of ML methods including deep learning for detecting cyber bullying in social networks, the study has highlighted the significance of big data and multimedia analysis.[16] has presented the ML based approach for detection of cyberbullying and toxicity. The study focuses the importance of real time detection and presents various classifier performances.

Authors in[17] have used labelled data with ML techniques using Weka tool kit, to train and recognize bullying content. They were able to identify the true positives with 78.5% accuracy.[18] have presented two new hypotheses for feature extraction to detect offensive things using supervised learning techniques. In[19], authors have studied the method of NLP with ML and they have tried to design a model which is inspired by Growing Hierarchical SOMs.

Research Gap

While many studies have explored ML based techniques for cyber bullying detection, there is a gap in research in real time detection system which can be integrated in systems and can be used effectively. We have tried to implement considering various categories of bullying which includesthe tweets into different types of cyber bullying, including age based, ethnicity based, gender based, religious based and non cyber bullying.

Contribution

This study contributes to the existing body of research by offering a comparison of traditional ML models in the cyberbullying detection. Unlike Deep Learning approach which requires very high computation resources and very large amount of data, our model aims to balance the performance with practical implementation considerations making them usable for real world applications.

Problem Statement and formulation

Cyber bullying these days have emerged as a very significant issue on digital platforms which involves digital or electronic means of communication to bully, harass or threaten individuals which can lead to severe psychological and emotional stress for the victims. Traditional methods for detection are not very good due to their inability to handle the large and dynamic data. The challenge is to develop an effective, real-time, and multilingual system for detecting cyber bullying.

Support Vector Machine, Logistic Regression and Naïve Bayes

Support Vector Machine (SVM) is a frequently used supervised learning algorithm which is widely used in text classification. SVM works by finding the most optimal hyperplane which is the main boundary which separated the different classes. SVM works by making the hyperplane which will maximize the margin between the given classes. While generally SVM works with linear hyperplane, but the kernel trick allows it to handle the nonlinear also. This is very crucial for text where the complexity can he high.

Logistic Regression method is used for the binary classification problems, with some modifications it can also be used to handle multi-class classification. It predicts the category of a given text based on the features given. It works in three steps, the text is first represented numerically which is known as text vectorization, after that feature extraction

happens and the final step involves calculation of weighted sum of the input features and applied the logistic function to obtain the class of the text.

Naïve bayes classifier is a classification algorithm which is based on Bayes' Theorem of probability. Here every feature is independent of each other. As it is a classification algorithm, it is highly used in text classification. It can be used extensively in spam filtering, sentiment detection and rating classification. The main advantage of this algorithm is that it is fast in making prediction with high dimension of data. Naïve bayes predicts the probability whether the following object belong to the particular class with a given set of value.

METHODOLOGY

This section will Aim at outlining the methodology which is being used for the cyber bullying detection. The methodology includes data collection, preprocessing of the data, feature extraction, model training, evaluation and comparison with previous studies.

Data Collection

We need to have some data set which will be used to train and test our model. The dataset used in our study consists of the tweets which are labelled with different categories of cyberbullying. The dataset here used we have loaded it in the form of a CSV (comma separated variable) file. Our dataset provides with very diverse range of examples thus ensuring that the model which will be trained can be generalized well to use in the real-world scenarios. Data set used in this study consists of 47692 tweets labelled into six-categories viz,

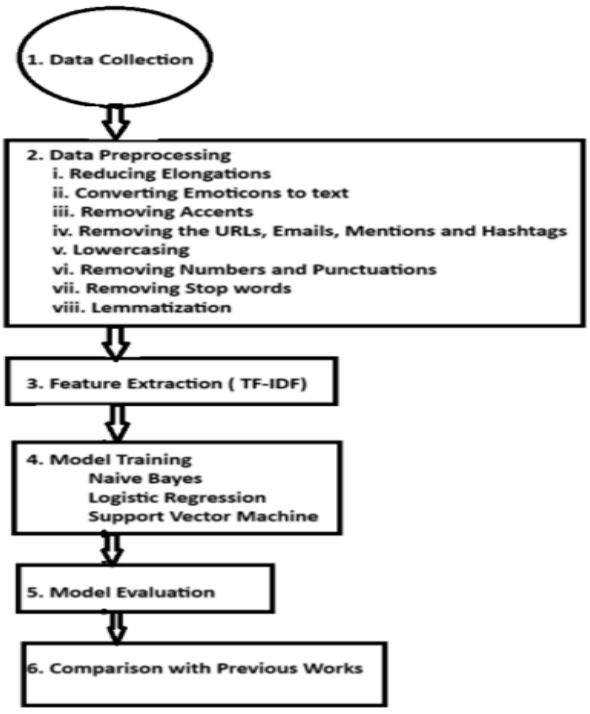

Figure 1: Methodology of the Project

age-based, ethnicity based, gender based, religions based, and not cyber bullying. The data set used was obtained for various twitter accounts ensuring the diversity in content and the language to make our study better.

Data Preprocessing

This data preprocessing is a very important step that will transform our raw data/text into a clean and structured format which will be suitable for the machine learning models. The steps will include.

Reducing Elongations

This step will reduce the size, in other words will normalizes the elongated words. For example: The word "looooooove" will be reduced to "love".

Converting Emoticons to Text

Replaces the emoticons, which are commonly used in text these days, to descriptive text. For example, ":)"will get replaced to "smile."

Removing Accents

This action will replace the accented characters present in the text which their non accented forms. Which means texts containing characters like é, ï, ç, é, ö, ñ etc. The output text will be non-accented. These characters will be replaced with normal characters like e, i, c, e, o, n.

Removing the URLs, Emails, Mentions and Hashtags

These are not important in our study, this it is important to remove these items. We need to focus on the tweet content and not on these things

Lowercasing

Converting our whole text to lowercase, it makes operations and other analysis easier, otherwise there would be complexity in our analysis involving both lowercase and uppercases.

Removing Numbers and Punctuations

This is done to make the text standard and which helps to make consistent text analysis. Numbers and punctuations generally do not contribute to the meaning of the text and act as a noise which leads to less accurate models. Removing irrelevant characters can improve the performance of ML models by focusing on the most meaningful parts.

Removing Stop Words

This step removes the common stop words because stop words like (is, the, and) etc do not contribute significantly to the meaning of the text. Removing these again helps ML models to focus on important part of text thus better accuracy.

Lemmatization

Reduces the words to their base or root form. This ensures that the different forms of the same word are treated as

```
def preprocess_text(text):
    text = re.sub(r'http\S+|www\S+', '', text)  # Remove URLs
    text = re.sub(r'\S+@\S+', '', text)  # Remove emails
    text = re.sub(r'@\w+', '', text)  # Remove user mentions
    text = re.sub(r'#(\w+)', r'\1', text)  # Remove hashtags
    text = reduce_elongations(text)  # Reduce elongated characters
    text = convert_emoticon_to_text(text)  # Convert emoticons to text
    text = remove_accents(text)  # Replace accented characters
    text = text.lower()  # Convert to lowercase
    text = re.sub(r'\d+', '', text)  # Remove numeric characters
    text = re.sub(r'[^\w\s]', '', text)  # Remove punctuation
    text = ' '.join([word for word in text.split() if word not in stop_words])
    text = ' '.join([token.lemma_ for token in nlp(text)])  # Lemmatize text
    return text
```

Figure 2: Preprocessing Steps followed in coding

same and single item. For example, “running” and “ran” are converted to “run”. Thus, creating more accurate and meaningful feature

Feature Extraction

After the necessary preprocessing, the cleaned text obtained is then transformed into numerical features using the commonly used TF-IDF (Term Frequency- Inverse Document Frequency) vectorizer. TF-IDF is a statistical measure which is used to evaluate the importance of a word in a document in relative to the collection of documents. Maximum Features set to 5000 to balance the computational efficiency and the feature representation.

Model Training

We train on the three ML models which were discussed before on the pre-processed dataset. They are Naïve Bayes, Logistic Regression and Support Vector Machine. The data set is first split into training and testing sets to evaluate the model’s performance.

Model Evaluation

The performance of each model is then evaluated considering in context various metrices like accuracy, precision, recall score and F-1 Score. These metrices provides a comprehensive understandingof how well the models perform on the test data.

Hyperparameters used

For Naïve Bayes we go with the default parameters, in Logistic Regression we use max_iter = 200, which specifies the maximum number of iterations the algorithm will take to evaluate the best solution. Sometimes the algorithm needs more time to find the optimal solution, and thus increasing this ensures that the model reaches convergence.

In SVM, we used ‘kernel’=’linear’. The kernel determines how the input data is transformed, the linear kernel means the algorithm looks for a straight line to separate the classes. Thus, using the linear kernel reduces the computational complexity compared to more complex polynomial or radial basis function.

```
# Initialize TF-IDF Vectorizer
vectorizer = TfidfVectorizer(max_features=5000)

# Fit and transform the data
X = vectorizer.fit_transform(df['cleaned_text']).toarray()
y = df['cyberbullying_type']  # Assuming 'cyberbullying_type' is the target column
```

Figure 3: Feature Extraction using TF-IDF

Comparion of our model with other studies

To check our results, we compare the findings of our current study with the other previous studies. This comparison gives us an idea about improvements in our results and performance.

By following this methodology, our study aims at developing an effective cyberbullying detection system using ML techniques. The use of multiple models and effective preprocessing ensures the accuracy and reliability of our results. The comparison with previous researches helps to demonstrate the advancements made in this research.

Simulation and Results

The model was run and the results obtained then studied. Each model was evaluated using precision, accuracy, recall, and f1 score.

The Naïve Bayes performed with 75.91%, it showed high precision for ‘ethnicity’ and lower recall for ‘not cyberbullying.’The Logistic Regression was performing with accuracy of 81.81%, it showed high precision and recall for ‘ethnicity’ and good results in ‘religion’ due to the ability of Logistic Regression to handle linear separability effectively. The SVM showed accuracy of 82.56%, it was best overall performance with high precision and recall across most of the classes.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train Naive Bayes
print("Training Naive Bayes...")
nb_model = MultinomialNB()
nb_model.fit(X_train, y_train)
nb_preds = nb_model.predict(X_test)
print("Naive Bayes Accuracy:", accuracy_score(y_test, nb_preds))
print(classification_report(y_test, nb_preds))

# Train Logistic Regression
print("Training Logistic Regression...")
lr_model = LogisticRegression(max_iter=200)
lr_model.fit(X_train, y_train)
lr_preds = lr_model.predict(X_test)
print("Logistic Regression Accuracy:", accuracy_score(y_test, lr_preds))
print(classification_report(y_test, lr_preds))

# Train Support Vector Classifier
print("Training Support Vector Classifier...")
svc_model = SVC(kernel='linear')
svc_model.fit(X_train, y_train)
svc_preds = svc_model.predict(X_test)
print("SVC Accuracy:", accuracy_score(y_test, svc_preds))
print(classification_report(y_test, svc_preds))
```

Figure 4: Training of Models

Table 1: Accuracy, Precision and Recall Score

Class	*Metric*	*Naïve Bayes*	*Logistic Regression*	*SVC*
Overall Accuracy		75.91%	81.81%	82.56%
Age	Precision	0.79	0.96	0.95
	Recall	0.95	0.96	0.97
Ethnicity	Precision	0.86	0.98	0.98
	Recall	0.90	0.96	0.97
Gender	Precision	0.76	0.90	0.90
	Recall	0.80	0.81	0.83
Not Cyberbullying	Precision	0.65	0.58	0.62
	Recall	0.41	0.56	0.51
Other Cyberbullying	Precision	0.60	0.59	0.59
	Recall	0.54	0.68	0.75
Religion	Precision	0.82	0.95	0.96
	Recall	0.96	0.94	0.94

Comparison with Previous Works

Previous study of[20] reported their accuracy of around 75% in Naïve Bayes while our current research achieved an accuracy of 75.91%. Thus, our research showed slight improvement which is mostly due to the preprocessing techniques used. The studies of[21] reported the accuracy of Logistic Regression of less than 80% while our methodology showed significant improvement of 81.81% which is due to effective feature extraction and preprocessing. In SVC, the previous study of[22] reported accuracy of around 80% while our current study achieved an accuracy of 82.56% which is due to better performance of SVC with linear kernels in handling the data with high dimensional text.

Although we did not use Deep learning models in our findings, papers suggest that integrating these deep learning models can enhance detection accuracy. Also, studies have showed the benefit of combining of textual, visual, and social features, by integrating these multi-modal data sources, the performance can be improved further.

Conclusion

Our findings show that Logistic Regression and SVC provide a good performance close to deep learning models of other researchers. They are computationally more efficient. Naïve Bayes can be better used in places where simplicity is prioritized over the need for the better accuracy. The improvements in our study can be because of the fact that we are doing very good preprocessing and very effective feature extraction using TF-IDF and our selection of optimised hyperparameters. Naïve Bayes while having the lowest overall accuracy, but it provided good precision for 'ethnicity', while it performed not so good for 'not cyberbullying' likely due to its pre assumption that features are independent, which however may not have held true in complex data.

References

[1] Owusu, S., & Zhou, L. (2015, May). Positive bystanding behavior in cyberbullying: The impact of empathy on adolescents' cyber bullied support behavior. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 163-165). IEEE.

[2] Singla, S., Lal, R., Sharma, K., Solanki, A., & Kumar, J. (2023, August). Machine learning techniques to detect cyber-bullying. In *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 639-643). IEEE.

[3] Rohini, D. S., & Ramchander, M. (2023, November). A Comparative Study of Machine Learning Approaches for Cyber bullying Detection in Digital Forums. In *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)* (pp. 332-338). IEEE.

[4] Serra, S. M., & Venter, H. S. (2011, August). Mobile cyber-bullying: A proposal for a pre-emptive approach to risk mitigation by employing digital forensic readiness. In *2011 Information Security for South Africa* (pp. 1-5). IEEE.

[5] Nakano, T., Suda, T., Okaie, Y., & Moore, M. J. (2016, February). Analysis of cyber aggression and cyber-bullying in social networking. In *2016 IEEE tenth international conference on semantic computing (ICSC)* (pp. 337-341). IEEE.

[6] Behzadi, M., Harris, I. G., & Derakhshan, A. (2021, January). Rapid Cyber-bullying detection method using Compact BERT Models. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)* (pp. 199-202). IEEE.

[7] Jain, V., Saxena, A. K., Senthil, A., Jain, A., & Jain, A. (2021, December). Cyber-bullying detection in social media platform using machine learning. In *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 401-405). IEEE.

[8] Pawar, V., Jose, D. V., & Patil, A. (2022, December). Explainable AI method for cyber bullying detection. In *2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC)* (pp. 1-4). IEEE.

[9] Hsien, Y. K., Salam, Z. A. A., & Kasinathan, V. (2022, April). Cyber bullying detection using natural language processing (NLP) and text analytics. In *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)* (pp. 1-4). IEEE.

[10] Siddhartha, K., Kumar, K. R., Varma, K. J., Amogh, M., & Samson, M. (2022, August). Cyber Bullying Detection Using Machine Learning. In *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)* (pp. 1-4). IEEE.

[11] Aqeel, H., & Kamble, A. (2022, December). A Hybrid Classifier of Cyber Bullying Detection in Social Media Platforms. In *2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)* (pp. 1-5). IEEE.

[12] Rajeevan, A., & Krishnaraj, N. (2023, January). Detection Of Cyberbullying based On Online Social Networks. In *2023 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-4). IEEE.

[13] Di Capua, M., Di Nardo, E., & Petrosino, A. (2016, December). Unsupervised cyber bullying detection in social networks. In *2016 23rd International conference on pattern recognition (ICPR)* (pp. 432-437). IEEE.

[14] Jia, H. L., Hameed, V. A., & Rana, M. E. (2022, January).

CyberSaver–A Machine Learning Approach to Detection of Cyber Bullying. In *2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (pp. 1-5). IEEE.

[15] Altay, E. V., & Alatas, B. (2018, December). Detection of cyberbullying in social networks using machine learning methods. In *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)* (pp. 87-91). IEEE.

[16] Jadhav, R., Agarwal, N., Shevate, S., Sawakare, C., Parakh, P., & Khandare, S. (2023, June). Cyber bullying and toxicity detection using machine learning. In *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)* (pp. 66-73). IEEE.

[17] Di Capua, M., Di Nardo, E., & Petrosino, A. (2016, December). Unsupervised cyber bullying detection in social networks. In *2016 23rd International conference on pattern recognition (ICPR)* (pp. 432-437). IEEE.

[18] Nandhini, B. S., & Sheeba, J. I. (2015). Online social network bullying detection using intelligence techniques. *Procedia Computer Science*, *45*, 485-492.

[19] Njuguna, L. W. (2024). AI-Assisted Digital Forensics for National Security Investigations. International Journal of Technology, Management and Humanities, 10(01), 125-146.

[20] Chavan, V. S., & Shylaja, S. S. (2015, August). Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 2354-2358). IEEE.

[21] Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 5, No. 3, pp. 11-17).