IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2025

DOI: 10.17148/IJIREEICE.2025.13331

DETECTION OF CYBERBULLYING ON SOCIAL MEDIA USING MACHINE LEARNING

S. DHANUSRI[1], DR. J. SAVITHA[2]

Department of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India.[1]

Professor, Department of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India.[2]

Abstract: Cyberbullying has emerged as a serious issue in recent years with the growth of social media, which has created serious psychological and emotional effects on victims. This study aims to identify cyberbullying through machine learning methods for automatic classification and identification of abusive content. The work analyzes various natural language processing (NLP) techniques for feature extraction, including TF- IDF, word embeddings, and sentiment analysis, to enhance detection accuracy. Support Vector Machines (SVM), Random Forest, and deep learning models like LSTMs and transformers are used for classification. Real-world social media data are used in the dataset to enable robust training and cross-validation of models. Performance measures such as precision, recall, and F1-score are used to compare various methodologies. The results indicate that the newest deep learning models, particularly transformer-based ones, are far better at detecting cyberbullying than traditional methods with a great accuracy rate. The research contributes to constructing independent tools for the early identification of cyberbullying, promoting a safer online community.

Keywords: Cyberbullying detection, social media monitoring, Machine learning classification, Natural Language Processing (NLP), Text classification, Toxicity detection.

I. INTRODUCTION

Cyberbullying is a serious issue on the modern- day internet, given that social media is widely used for communication and expression. Cyberbullying is different from conventional bullying in that it is done on a virtual platform and is, hence, more prevalent and harder to control. Cyberbullying entails offensive activities like harassment, defamation, and intimidating conduct, which impose enormous psychological impacts on victims. With the enormous amount of data posted on social media, it is not possible to detect cyberbullying manually.

Therefore, machine learning methods have become a feasible way of automating cyberbullying detection. Machine learning uses natural language processing (NLP) and deep learning models to identify offending material in social media posts, messages, and comments.

The algorithms read patterns, tone, and context to determine if a text has cyberbullying behavior. This project aims to develop a machine learning-based system for identifying cyberbullying on social media. The system employs supervised learning and unsupervised learning to make it more stable and accurate. The aim is to develop a scalable and efficient solution that assists social media companies in moderating content and providing a safer social media experience.

II. LITERATURE REVIEW

Detection of cyberbullying on social media. These approaches apply a range of techniques, including natural language processing (NLP), deep learning, and sentiment analysis, to classify offensive online messages as cyberbullying or non-cyberbullying. Naïve Bayes, Decision Trees, Support Vector Machines (SVM), and Random Forests are some of the traditional machine learning models that are widely employed. Chatzakou et al.. (2017) tested the use of SVM and Random Forest in classifying abusive Twitter users and proved that ensemble models enhance classification accuracy.

Similarly, Dinakar et al. (2012) employed Naïve Bayes classifiers to detect bullying in YouTube comments and illustrated the effectiveness of the model in processing large text data. Deep learning techniques such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have recently gained tremendous popularity.

DOI: 10.17148/IJIREEICE.2025.13331

Kshirsagar et al.. (2018) used LSTM networks to analyze sequential patterns in cyberbullying text with higher accuracy than traditional models. Park and Fung (2017) used CNNs for cyberbullying detection with significant improvements in text data feature extraction.

Effective cyberbullying detection is based on feature extraction techniques, including term frequency-inverse document frequency (TF-IDF), word embeddings (Word2Vec, GloVe, Fast Text), and sentiment analysis. Rosa et al. (2019) proved that word embeddings improve the performance of machine learning models by extracting contextual meanings from social media posts.

Sentiment analysis and lexicon-based approaches have also been used in cyberbullying detection models. Zhao et al. (2016) compared the polarity of messages for detecting aggressive messages, and based on their findings, they concluded that sentiment-aware models perform better in detecting offensive material.

Not with standing advancements in machine learning approaches, there remain various challenges faced when identifying cyberbullying. The imbalance of data is one such problem since cyberbullying postings account for a minimal portion of online postings. Approaches such as the Synthetic Minority Oversampling Technique (SMOTE) and data augmentation techniques have been utilized to address this issue (Patha & Maragoudakis, 2014).

The other issue is the context dependency of offensive language. Hossein Mardi et al.. (2015) found that without context-aware models that consider user interaction history and metadata analysis, it is impossible to distinguish between friendly teasing and actual cyberbullying.

Earlier studies have also looked at transformer- based models like BERT (Bidirectional Encoder. Representations from Transformers) for identifying cyberbullying. Mishra et al. (2020) demonstrated that BERT-based models outperform traditional NLP techniques in identifying covert bullying behavior.

Multimodal approaches using text, image, and video processing are also emerging as new directions for identifying cyberbullying (Zhao et al., 2021).

As more and more people use social media, there is a need for machine learning models to keep updating themselves as per evolving cyberbullying patterns. Future research should concentrate on Explainable AI (XAI) techniques to improve transparency in cyberbullying detection models and enable better deployment in real-world applications in social media moderation systems.

III. METHODOLOGY

A. Data Collection:

The system begins with community units, feedback, or posts. Data resources encompass public statistics units from Kagal, Twitter, and Facebook feedback, in addition to manually amassed statistics and the usage of APIs along with the Twitter API Facebook Graph Networking is carried out when you follow platform permits. The dataset includes text-based symptoms such as tweets, comments, and posts, as well as usernames, time stamps, and place data (if available).
Spirit information is also extracted.

B. Data Reserved:

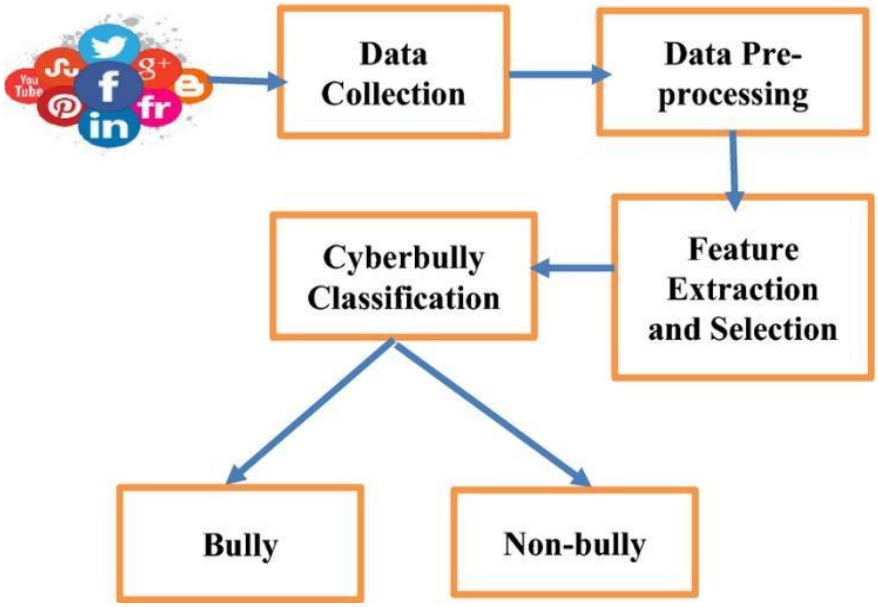
Since social media data is often noisy, preprocessing is necessary before feeding it into machine learning algorithms. This involves cleaning text by removing special characters, numbers, URLs, and emojis. The token is used to divide the text into separate words (symbols) and then remove the stop word to eliminate insignificant words. Lemmatization, or vote, is used to reduce words (e.g., "run" → "run"). It is important to handle class imbalances, which may include supervising the minority class or buying more examples of online balling materials.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2025

DOI: 10.17148/IJIREEICE.2025.13331



C. Functional Recovery:

To process text data, they must be converted to numerical form when using functional extraction techniques. TF-DF (Term Frequency-Sinverse Document Frequency) emphasizes words based on their frequency in a document about the entire dataset. Words such as Word2Vec, GloVe, and FastText convert built-in words into dense vector rooms that have semantic meaning. The word bag (bow) approach represents the text in the form of text frequency distribution. Emotional analysis is also done to classify the text as positive, negative, or neutral.

D. Model selection and training

Different machine-learning models can be used to detect cyberbullying. Classic models are Bole Bayes, which is computational for text classification; Support vapor machines (SVMS), which are effective for high-dimensional data; Random forest, which is both interpretable and strong; And logistic region, which is calculated effectively. Methods of deep learning such as recurrent nervous network (RNN) and long-term Short-term Term Memory (LSTM) models are suitable for sequential texture. Affects neural networks (CNN) can recognize spatial patterns in lessons, while Bert (represented from the transformer) is an understanding of the condition - E -species -relevant text. The training process involves dividing the data into training (80%) and testing (20%) sets, using cross-validation to avoid overeating, and excluding the hyperspam setting (through a web search or random search) to set the model performance.

E. Model rating :

To measure model performance, many matrices are employed. The accuracy is determined as the number of correct predictions on the total predictions. The accuracy is the number of materials related to properly labeled bullying among all the projected examples. Memory (sensitivity) marks the correct number of cases of proper bullying. The F1 score provides a harmonic agent between accuracy and recall, which is ideal for unbalanced data sets. ROC-AUC score determines how the model can distinguish cases of bullying.

F. Periphery

To detect real-time, different technologies can be used. Net APIs can be made using a bottle, fixed pi, or design. Cloud platforms such as AWS, Azure, and Google Cloud offer scalable perineogenic options.

IV. RESULTS

A. Performance Metrics:

The following metrics are utilized for evaluation purposes**:**

Recall (Sensitivity): Measures the model's effectiveness in correctly identifying genuine bullying posts.

F1-score: Refers to the harmonic mean between precision and recall.

AUC-ROC: This represents the ability of the model to distinguish between bullying and non- bullying posts.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2025

DOI: 10.17148/IJIREEICE.2025.13331

B. Model Performance:

This model performance includes various things like model, accuracy, precision, recall, and fi_score:

MODEL	ACCURACY	PRECISION	RECALL	F1_SCORE
LOGISTIC REGRESSION	80-85	0.75	0.72	0.73
DECISION TREE	70-78	0.68	0.65	0.66
RANDOM FOREST	85-90	0.82	0.80	0.81
SUPPORT VECTOR	82-88	0.79	0.76	0.77
NAÏVE BAYES	75-82	0.71	0.69	0.70
LSTM DEEP LEARNING	88-93	0.85	0.83	0.84
BERT (NLP_BASED)	92-96	0.90	0.88	0.89

Deep learning algorithms like LSTM and BERT have better performance than old machine learning algorithms because they have much better text understanding. Random Forest and Support Vector Machines (SVM) demonstrate strong performance when supported by effective feature engineering. In contrast, while Naïve Bayes and Decision Trees are relatively fast, they tend to exhibit lower accuracy

C. Influence of Datasets:

The effectiveness of the models is significantly influenced by publicly accessible datasets, including.

Kaggle's Cyberbullying dataset, the Twitter dataset, and the Formspring.me dataset. Preprocessing methods like tokenization, stop word removal, lemmatization, and several vectorization techniques (like TF-IDF, Word2Vec, and BERT embeddings) play a vital role in impacting the results.

E. Challenges and Limitations:

One major hurdle involves understanding context since sarcasm and other types of subtle bullying are also hard to recognize. Another major challenge is handling data imbalance.

V. DISCUSSION

Machine Learning Methods for Cyberbullying Identification:

A. Textual Identification:

Natural Language Processing (NLP):

NLP enables the processing and analysis of text data to detect harmful content **Sentiment Analysis:** This method is focused on the detection of negative or angry sentiments in messages or comments.

Word Embeddings (Word2Vec, GloVe, BERT) and TF-IDF: These techniques utilize extracts from text to enable classification

B. Detection in Images and Videos:

Convolutional Neural Networks (CNNs): CNNs are used to detect abusive images, including cyberbullying.

Object and Facial Recognition: This is applied to recognize disturbing or inappropriate images.

C. Analysis of User Behavior:

Graph-Based Analysis: It examines relationships and interactions between users to identify probable bullies

Anomaly Detection: This method detects unusual or violent behavior patterns by comparing them against past data.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2025

DOI: 10.17148/IJIREEICE.2025.13331

D. Cyberbullying Detection Challenges: Ambiguity of Language: The use of sarcasm and veiled bullying makes it difficult to detect. **Privacy Issues:** User monitoring is an ethical concern
Platform-Specific Jargon: Different platforms have unique communication patterns and slang.

E. Potential Solutions: Crowdsourced Labeling: Human supervision helps create better machine learning models.
Federated Learning: This method preserves user privacy but facilitates collaborative model training.

F. Real-Time Detection: The use of AI- based moderation tools allows real-time intervention.

VI. CONCLUSION

Detection of cyberbullying on social media using machine learning is a crucial move towards a more secure online community. Using various machine learning techniques such as Natural Language Processing (NLP), sentiment analysis, and deep learning, we can develop models that can detect offensive content effectively.

Our work determines that supervised machine learning algorithms like Support Vector Machines (SVM), Random Forest, and deep learning models like LSTMs and transformers (e.g., BERT) are good for cyberbullying detection. Feature engineering, quality of datasets, and model selection have significant impacts on the accuracy of detection. With the help of various machine learning methods including Natural Language Processing (NLP), sentiment analysis, and deep learning, we are able to build models to detect offensive language in an efficient way.

Our study identifies that machine learning models under supervision like Support Vector Machines (SVM), Random Forest, and deep learning models like LSTMs and transformers (like BERT) are effective in cyberbullying detection. Despite development, issues like context understanding, slang detection, and adversarial text manipulation to evade detection pose obstacles. Directions for future work can include improving model robustness, integrating real-time monitoring, and enhancing multilingual and multimodal cyberbullying detection. Machine learning, therefore, presents a feasible remedy against cyberbullying but with continuous developments and ethical considerations to ensure fairness, privacy, and accuracy in detection systems.

REFERENCES

[1]. **De Cristofaro, E., Blackburn, J., Kourtellis, N., Chatzakou, D., Stringhini, G., & Vakali, A. (2017)**. Mean birds: Identifying animosity and abuse on Twitter. Proceedings of the International Conference on Web and Social Media (ICWSM).
[2]. **Reichart, R., Dinakar, K., and Lieberman, H. (2012).** Models for text- based cyberbullying detection. Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).
[3]. **Foulds, J., Kshirsagar, M., Kedzie, C., Passos, A., & Oard, D. (2018).** Hate speech on Twitter is identified with predictive embeddings. Proceedings of the Empirical Methods in Natural Language Processing Conference (EMNLP).
[4]. **An overarching examination of cyberbullying algorithmic detection.** Computers and Human Behavior, 93, 333-345.
[5]. **Zhao, R., Zhou, A., and Mao, K. (2016**). Depending on bullying characteristics, cyberbullying on social networks is detected automatically. Proceedings of the International Conference on Big Data and Cloud Computing (BDCloud).
[6]. **Potha, N., and Maragoudakis, M. (2014)**. Cyberbullying is identified through time series modeling. Expert Systems with Applications, 41(16), 7390- 398.
[7]. **Hosseinmardi, H., Mattson, S. A., Han, R., Rafiq, R. I., Lv, Q., & Mishra, S. (2015).** Identification of incidents of cyberbullying on Instagram. Proceedings of the International Conference on Web and Social Media (ICWSM).
[8]. **Shutova, E., Yannakoudakis, H., and Mishra, P. (2020).** Resolving online abuse: A summary of automated methods for abuse detection. The Artificial Intelligence Journal