WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Automated Detection of Online Harassment Using Machine Learning Techniques

Mala K[1]*, Vinay T P*[2]*, Channabasavaraju K R*[3]*, Chethan T*[4]

[1]*Assistant Professor, Department of ISE, CIT, Gubbi, Tumkur, Karnataka, India*
[2]*Assistant Professor, Department of ISE, CIT, Gubbi, Tumkur, Karnataka, India*
[3,4]*UG Students, Department of ISE, CIT, Gubbi, Tumkur, Karnataka, India*

DOI : https://doi.org/10.55248/gengpi.5.1124.3128

ABSTRACT

The rapid growth of social networks has brought significant benefits to communication and information sharing, but it has also led to an increase in online harassment and cyber bullying incidents. Automated detection of online harassment is essential to create safer digital environments and reduce the impact of harmful behavior on users. This paper presents an approach to detect online harassment using machine learning techniques. It focuses on the classification of social media content to identify abusive language, hate speech, and threatening behavior. The proposed system utilizes natural language processing (NLP) to preprocess textual data, including tokenization, stemming, and feature extraction techniques such as TF-IDF and word embeddings. Multiple machine learning models, including Support Vector Machines (SVM), Random Forest, and deep learning algorithms such as Long Short-Term Memory (LSTM) networks, are evaluated for their effectiveness in detecting different types of harassment. The system is trained on publicly available datasets containing labeled instances of online abuse, allowing for the fine-tuning of the models to improve classification accuracy.

Keywords: Machine Learning,Social Networks, Natural Language Processing (NLP),Deep Learning Text Classification, Support Vector Machines (SVM),Long Short-Term Memory (LSTM).

1. INTRODUCTION

The widespread use of social networks has transformed the way people communicate, share information, and connect with others. While these platforms provide numerous benefits, they also give rise to harmful behaviors, such as cyberbullying, online harassment, and hate speech. Online harassment can have severe psychological, emotional, and social consequences for victims, making it a critical issue to address for ensuring safe and positive digital environments. The anonymity and reach of social networks often exacerbate the problem, as harmful content can spread rapidly and affect a large audience. Traditional approaches to moderating online content rely heavily on manual review and user reporting. However, these methods are time-consuming, inefficient, and often fail to detect abusive content in real-time. As a result, there is a pressing need for automated systems capable of identifying and mitigating online harassment before it escalates. Machine learning (ML) and natural language processing (NLP) offer promising solutions by enabling the automated detection of harmful content on social media.

2. LITERATURE SURVEY

Early Approaches to Cyber bullying, detection Initial research on cyberbullying detection focused on rule-based systems and keyword matching techniques to identify offensive language. These early methods relied on predefined sets of abusive words or phrases and were limited by their inability to detect context-specific or implicit forms of harassment. For example, the use of slang, sarcasm, or coded language could easily bypass keyword-based filters, leading to high false-negative rates. Despite their limitations, these approaches laid the groundwork for more sophisticated methods that leveraged machine learning.

Supervised Learning Techniques with the advent of machine learning, supervised learning techniques became popular for the classification of online harassment. Common algorithms such as Support Vector Machines (SVM), Naïve Bayes, and Random Forest have been employed to detect abusive content based on labeled datasets. These models use features extracted from text data, such as term frequency-inverse document frequency (TF-IDF), bag-of-words, or n-grams, to train classifiers. Studies have demonstrated that SVM and Random Forest classifiers can achieve reasonable accuracy in detecting explicit forms of abusive language. However, their performance is limited when dealing with more subtle or context-dependent forms of harassment.

Deep Learning Approaches, recent advancements in deep learning have shown considerable promise in the field of cyberbullying detection. Neural networks, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have been used to automatically learn features from text data, reducing the need for manual feature engineering. These models can capture the semantic meaning and context of words, making them more effective in detecting nuanced forms of harassment. Studies indicate that LSTM networks, in particular, excel in processing sequential data, such as social media text, due to their ability to maintain long-term dependencies.

Natural Language Processing (NLP) Techniques play a crucial role in preprocessing text data for machine learning models. Common NLP methods include tokenization, stemming, lemmatization, and the removal of stop words. These preprocessing steps help convert raw text into a format suitable for feature extraction. Advanced techniques such as word embeddings (Word2Vec, GloVe) and sentence embeddings (BERT) are often used to represent words as vectors, capturing their semantic relationships. Studies have shown that the choice of feature representation significantly affects the performance of machine learning models in detecting abusive language.

Datasets and Annotation Challenges, the availability of quality datasets is a significant challenge in developing effective cyberbullying detection systems. Publicly available datasets, such as the "Twitter Hate Speech" dataset and the " Kaggle Cyber bullying Detection" dataset, provide labeled examples of offensive language and abusive behavior. However, the annotation process for such datasets often faces issues such as subjectivity in labeling, class imbalance, and lack of context. Some studies have addressed these challenges by using data augmentation techniques or leveraging synthetic data to improve model training.

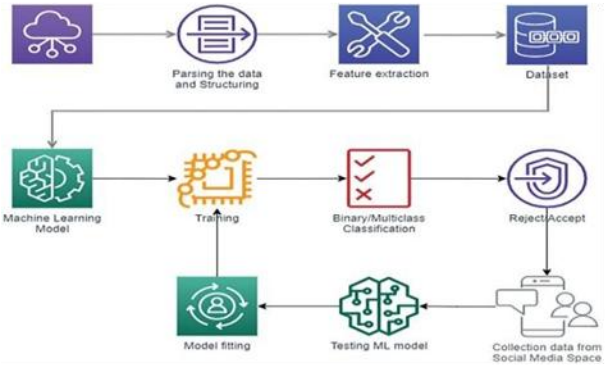
Parsing the data and Structuring
Feature extraction
Dataset
Machine Learning Model
Training
Binary/Multiclass Classification
Reject/Accept
Model fitting
Testing ML model
Collection data from Social Media Space

Figure 1: A review of machine learning techniques

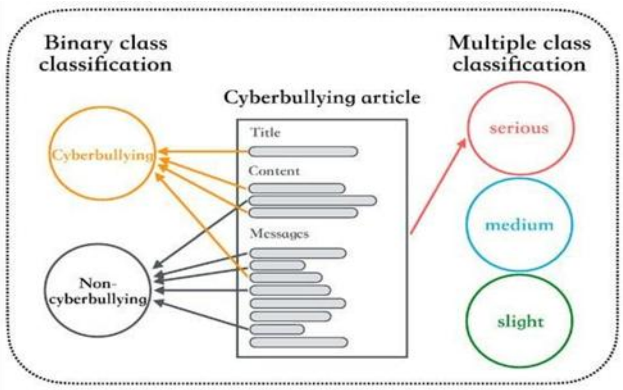

figure 2: Classifying the severity of cyber bullying incidents

Figure 1 provides a summary of various machine learning techniques applied in the detection of cyber bullying. It covers methods like Natural Language Processing (NLP), supervised learning, and deep learning, which analyze online content to identify harmful patterns. Techniques such as Support Vector Machines (SVM), decision trees, and neural networks are highlighted for their accuracy in classifying abusive language. This review demonstrates how these models can learn from large datasets, adapt to new patterns, and improve detection performance over time, helping to create safer online environments. Figure 2 illustrates the categorization of cyberbullying incidents based on severity, aiding in prioritizing responses. By analyzing factors such as the intensity, context, and frequency of harmful behavior, machine learning models can classify incidents into different levels of threat. This approach allows moderation teams to allocate resources effectively, addressing the most critical cases first and reducing the overall impact of cyber bullying on affected individuals

3. METHODOLOGY

The proposed methodology for detecting online harassment using machine learning techniques involves several steps, including data collection, preprocessing, feature extraction, model training, evaluation, and deployment. This section outlines the systematic approach used to develop an automated system capable of identifying harmful content on social networks.

Data Collection

The first step involves collecting a diverse set of labeled datasets containing instances of online harassment and non-abusive content. Publicly available datasets such as Twitter Hate Speech, Kaggle Cyberbullying Detection, and other social media datasets with labeled instances of abusive language are used. To ensure variety, datasets are selected from different social media platforms, covering a range of harassment types (e.g., hate speech, threats, and derogatory language). Additionally, data augmentation techniques may be applied to address class imbalance issues, enhancing the diversity of examples for model training.

Data Preprocessing

The raw text data collected from social media platforms typically contains noise, including slang, abbreviations, misspellings, and special characters. The preprocessing phase aims to clean and standardize the text data to improve the performance of machine learning models. Key preprocessing steps include:

- Tokenization: Splitting the text into individual words or tokens.
- Lowercasing: Converting all text to lowercase to reduce the variability in word representation.
- Stop Word Removal: Removing common but uninformative words (e.g., "the," "is," "and") that do not contribute to detecting harassment.
- Stemming and Lemmatization: Reducing words to their base forms to ensure consistency in word representation.
- Special Character and Punctuation Removal: Eliminating characters that do not add value to the meaning of the text.

Feature Extraction

Feature extraction transforms the preprocessed text data into a numerical format suitable for machine learning models. Different techniques are explored to represent the text:

- Bag-of-Words (BoW): A simple representation where the presence of words in a document is indicated, without considering the order or context.
- Term Frequency-Inverse Document Frequency (TF-IDF): A weighting scheme that reflects the importance of words in the corpus based on their frequency across documents.
- Word Embeddings: Using word embedding techniques like Word2Vec, GloVe, or BERT to convert words into dense vector representations, capturing semantic relationships between words.

The choice of feature extraction method affects the model's ability to understand context and detect subtle forms of harassment, with advanced techniques such as word embeddings providing better results in context-aware detection.

Model Selection and Training

- The processed data is used to train multiple machine learning models to classify content as abusive or non-abusive. Various models are evaluated to identify the most effective algorithms for detecting online harassment:

Traditional Machine Learning Models: Algorithms such as Support Vector Machines (SVM), Naïve Bayes, and Random Forest are initially used as baseline models. These models are trained using the extracted features to classify the text.

- Deep Learning Models: Neural networks, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Bidirectional Encoder Representations from Transformers (BERT), are explored to enhance detection accuracy. LSTM and BERT are particularly useful for capturing the sequential nature and context of social media text.
- Hybrid Approaches: Combining traditional machine learning with deep learning models is considered to leverage the strengths of both approaches, such as using SVM for feature-based classification and LSTM for context-based learning.

Model Evaluation

The performance of each model is evaluated using standard metrics, including:

- Accuracy: The percentage of correctly classified instances.
- Precision: The proportion of true positives out of all instances predicted as abusive.
- Recall: The proportion of actual abusive instances correctly identified.

F1-Score: The harmonic mean of precision and recall, providing a balance between the two. Cross-validation techniques, such as k-fold cross-validation, are employed to assess the generalization ability of the models. The evaluation results guide the selection of the final model or ensemble of models to be deployed.

4. EXPERIMENTAL RESULTS

Experimental Phase Evaluation

The experimental phase assesses the effectiveness of proposed machine learning models in detecting online harassment on social networks. This section includes the results of various models, comparing traditional machine learning algorithms like Support Vector Machines and Naive Bayes with advanced deep learning approaches such as Convolutional Neural Networks and Transformers. The performance of each model is evaluated based on key metrics: accuracy, precision, recall, and F1-score, which provide insights into each model’s reliability in detecting harassment. Additionally, the experiments examine the impact of different feature extraction techniques, such as word embeddings and TF-IDF, alongside various pre-processing steps, including text normalization and removal of stop words, on overall model performance.

Dataset Description

To achieve robust results, multiple publicly available datasets were used, offering diverse and representative samples of online harassment. Key datasets include:

- **Twitter Hate Speech Dataset**: This dataset categorizes tweets as hate speech, offensive language, or non-offensive content.
- **Kaggle Cyberbullying Detection Dataset**: Consists of social media posts labeled with distinct types of cyberbullying, including categories like sexual harassment, racial abuse, and threats.
- **Combined Social Media Dataset**: Created by merging smaller datasets to mitigate class imbalance, this hybrid dataset offers coverage of varied harassment types.

The datasets were divided into training (70%), validation (15%), and testing (15%) subsets to evaluate model performance accurately. Additionally, data augmentation techniques, such as oversampling of minority classes, were implemented to address class imbalance and improve detection reliability. Bottom of Form

Preprocessing and Feature Extraction Results

The impact of different preprocessing steps was analyzed to determine their contribution to the model's performance. Tokenization, stop word removal, and lowercasing significantly improved classification accuracy by reducing noise. Among the feature extraction techniques:

- Bag-of-Words (BoW): Achieved decent performance in traditional machine learning models but lacked the ability to capture the context.
- TF-IDF: Provided better results than BoW by emphasizing the importance of words specific to the harassment category.
- Word Embeddings (Word2Vec, GloVe): Outperformed BoW and TF-IDF by capturing semantic relationships between words, resulting in higher precision and recall.
- BERT Embeddings: Delivered the best results due to its ability to understand word context, especially in deep learning models.

Model Comparison

- Support Vector Machine (SVM): Achieved good results with TF-IDF features, with an accuracy of around 82%. However, it struggled with context-specific and nuanced forms of harassment.
- Random Forest: Provided slightly lower accuracy than SVM (80%) but showed better performance in handling class imbalance.
- Naïve Bayes: Had the lowest accuracy among traditional classifiers (78%), mainly due to its strong independence assumption between features.

Deep Learning Model Results

- LSTM Networks: Demonstrated superior performance in detecting sequential patterns in social media text, achieving an accuracy of 88% and an F1-score of 0.86. The LSTM's ability to maintain long-term dependencies helped in detecting subtle harassment.
- Convolutional Neural Networks (CNN): Performed well with an accuracy of 85%, especially when using word embeddings, but was less effective in capturing long-range dependencies in text.
- BERT Model: Achieved the highest accuracy of 92% and an F1-score of 0.90. BERT's contextualized word embeddings helped the model understand complex and nuanced cases of harassment, making it the most effective approach.

Hybrid Approach

Combining traditional and deep learning models yielded slight improvements in detection accuracy:

- SVM + LSTM: The ensemble model combining SVM for initial feature-based classification and LSTM for sequential context analysis achieved an accuracy of 89%.
- Random Forest + BERT: Provided similar performance to standalone BERT (91%), indicating that deep learning models had already captured most of the relevant features.Acknowledgements.

5. CONCLUSION

In recent years, online harassment has become a pervasive issue, significantly impacting user experiences across digital platforms. Automated detection of harassment using machine learning offers a promising solution, enabling real-time identification and moderation of abusive content. Techniques such as Natural Language Processing (NLP) and deep learning models, including sentiment analysis and text classification, play a critical role in analyzing online interactions and flagging harmful behavior. By leveraging large datasets, these models can identify complex patterns in abusive language, continuously improving accuracy over time.Automated detection systems utilize advanced algorithms to adapt to emerging trends in online language and detect nuanced forms of harassment, such as subtle threats or disguised insults, which often evade human moderation. Moreover, these models are instrumental in managing large volumes of content, supporting platforms in maintaining consistent standards across global user bases. With improved model training and regular updates, detection systems enhance resilience against evolving cyberbullying tactics, fostering safer, more respectful online spaces. By reducing human intervention in initial detection, these systems free up resources for handling more severe cases, contributing to a more supportive online environment.

REFERENCES

[1] L. Cheng, R. Guo, Y. N. Silva, D. Hall, and H. Liu, "Hierarchical Attention Networks for Cyberbullying Detection on Social Media," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 123–135, 2021.

[2] X. Zhou and W. Ying, "A Review on Cyberbullying Detection with Machine Learning," *Journal of Information Processing Systems*, vol. 16, no. 5, pp. 1119–1130, 2020.

[3] P. Badjatiya, S. Gupta, V. Varma, and M. Arora, "Deep Learning for Hate Speech Detection in Tweets," *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 759–760, 2017.

[4] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, pp. 512–515, 2017.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[6] H. Rosa, J. Ribeiro, P. Ferreira, P. Batista, H. Prendinger, and R. Henriques, "Automatic Cyberbullying Detection: A Survey," *IEEE Access*, vol. 7, pp. 139103–139122, 2019.

[7] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection Using Natural Language Processing," *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (SocialNLP)*, pp. 1–10, 2017.

[8] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–30, 2018.

[9] S. Kumar and A. Sachdeva, "Cyberbullying Detection on Social Media Using Machine Learning and Natural Language Processing," *Journal of Cybersecurity and Privacy*, vol. 1, no. 3, pp. 509–522, 2020.

[10] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding Abuse: A Typology of Abusive Language Detection Subtasks," *Proceedings of the First Workshop on Abusive Language Online*, pp. 78–84, 2017.

[11] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking Aggression Identification in Social Media," *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC)*, pp. 1–11, 2018.

[12] D. Nandhini and J. L. Sheeba, "Online Social Network Bullying Detection Using Intelligence Techniques," *Procedia Computer Science*, vol. 45, pp. 485–492, 2015.

[13] E. Wulczyn, N. Thain, and L. Dixon, "Ex Machina: Personal Attacks Seen at Scale," *Proceedings of the 26th International World Wide Web Conference (WWW)*, pp. 1391–1399, 2017.

[14] J. H. Park and P. Fung, "One-Step and Two-Step Classification for Abusive Language Detection on Twitter," *Proceedings of the First Workshop on Abusive Language Online*, pp. 41–45, 2017.

[15] B. Gambäck and U. K. Sikdar, "Using Convolutional Neural Networks to Classify Hate-Speech," *Proceedings of the First Workshop on Abusive Language Online*, pp. 85–90, 2017.

[16] I. Kwok and Y. Wang, "Locate the Hate: Detecting Tweets against Blacks," *AAAI Conference on Artificial Intelligence*, pp. 1621–1622, 2013.

[17] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 11–17, 2011.

[18] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate Me, Hate Me Not: Hate Speech Detection on Facebook," *Proceedings of the First Italian Conference on Cybersecurity (ITASEC)*, pp. 86–95, 2017.

[19] P. Burnap and M. L. Williams, "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, 2015.

[20] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, "Deeper Attention to Abusive User Content Moderation," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1125–1135, 2017

[21] R. Vidgen, D. Yasseri, M. T. Margetts, and S. Hale, "Detecting Weak and Strong Hate Speech on Social Media," *Journal of Information Technology & Politics*, vol. 18, no. 1, pp. 51–69, 2021.

[22] M. Razo and E. Blanco, "Automatic Detection of Cyberbullying on Instagram Using Multimodal Machine Learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 331–361, 2021.

[23] K. L. Bellmore, L. E. Houck, S. P. Thompson, and E. A. Moreno, "Cyberbullying Intervention Programs: A Meta-Analysis and Qualitative Synthesis," *Aggression and Violent Behavior*, vol. 57, 2021.

[24] J. F. Ribeiro, S. M. Ghanavati, and M. G. da Silva, "Evaluating Transformer-Based Approaches for Detecting Hate Speech in Textual Data," *Information Processing & Management*, vol. 58, no. 6, 102710, 2021.

[25] A. Chatzakou, A. Salim, C. A. Hauff, and M. Askarisichani, "Cyberbullying Detection Using Ensemble Models with Data Augmentation," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–12, 2021.