International Journal of Advanced Multidisciplinary Research and Educational Development
Volume 2, Issue 2 | March – April 2026 | www.ijamred.com
ISSN: **3107-6513**

Detection of Cyberbullying on Social Media Using Machine Learning

Mohammed Aashiq S
Department of Information Technology,
Sri Krishna Adithya College of Arts and Science,
Coimbatore, Tamil Nadu, India
smohammedaashiq864@gmail.com

Dr. Gobi I
Department of Information Technology,
Sri Krishna Adithya College of Arts and Science,
Coimbatore, Tamil Nadu, India

ABSTRACT— The widespread use of the internet has made social media a primary medium for communication. Along with its benefits, the increase in online interactions has also led to a rise in cyberbullying, which can cause serious psychological and physical harm, especially to women and children. This project focuses on designing an effective machine learning model to identify cyberbullying content on social media platforms. The system uses a labelled text dataset obtained from Kaggle, categorized as normal or cyberbullying. After data collection, preprocessing steps such as cleaning, normalization, and feature reduction are performed. A decision tree algorithm is then trained to classify the text data, and its performance is evaluated using test samples. The proposed model accurately distinguishes harmful content, enabling early detection and contributing to a safer online environment.

KEYWORDS—Cyberbullying detection, Machine learning, Social media, Decision tree algorithm.

I. INTRODUCTION

The rapid growth of internet usage has reshaped communication, making social media a powerful platform for interaction and information sharing. However, increased online activity has also led to a rise in cyberbullying, which poses serious risks to mental and physical well-being, particularly among women and children. Addressing this issue has become a major concern in the digital age.

This study aims to develop a machine learning–based system to detect cyberbullying in social media content and support early intervention. The proposed model uses a labelled text dataset from

ISSN: 3107-6513

Kaggle, categorized as normal or cyberbullying, and divided into training and testing sets. Data preprocessing techniques such as cleaning, normalization, and feature reduction are applied before training a decision tree classifier. The trained model effectively distinguishes harmful content from normal communication, contributing to improved user safety and a healthier online environment.

II. LITERATURE SURVEY

1. **Dinakar et al. (2011)** explored supervised learning techniques for identifying cyberbullying in online social networks. The study evaluated classifiers such as Support Vector Machines, Naïve Bayes, and Decision Trees on labelled comment datasets. Results indicated that SVM achieved superior accuracy, while also highlighting difficulties in detecting sarcasm and contextual meaning.
2. **Xu et al. (2012)** introduced a rule-based and machine learning hybrid approach for identifying offensive language in social media. Their results showed improved detection when
3. **Dadvar et al. (2013)** proposed a cyberbullying detection approach that combined textual features with user profile information such as age, gender, and user activity. Their study showed that incorporating user metadata alongside text analysis improved classification performance. A decision tree–based machine learning model was used to analyze social media data, achieving higher precision and recall than text-only methods combining linguistic rules with statistical models.
4. **Cheng et al. (2017)** analyzed abusive behavior patterns on social platforms using user interaction features and content-based analysis. The study highlighted the importance of behavioral features in improving cyberbullying prediction.
5. **Fortuna and Nunes (2018)** provided a comprehensive survey on automatic hate speech and cyberbullying detection. The study summarized datasets, algorithms, and open challenges, emphasizing the importance of context-aware models.

III. PROPOSED METHODOLOGY

The proposed methodology addresses the shortcomings of existing cyberbullying detection systems by introducing an efficient machine learning–based approach for identifying harmful content on social media platforms. With the rapid increase in online interactions, there is a strong need for automated systems that can accurately detect and reduce instances of online harassment. The main objective of this system is to develop a reliable and high-performance model that effectively identifies cyberbullying and supports safer digital communication.

ISSN: 3107-6513

The first stage involves **dataset collection**, where a labelled dataset is obtained from the Kaggle repository. The dataset consists of text messages categorized as either *normal* or *cyberbullying*. These labels enable supervised learning and guide the model during the training and evaluation phases.

Next, **data preprocessing** is performed to improve data quality and model performance. This step removes noise such as special characters, URLs, punctuation, and stop words. Feature reduction techniques are applied to retain only relevant information, making the dataset more suitable for classification.

To enhance text understanding, **Natural Language Processing (NLP)** techniques are applied. Processes such as tokenization, stemming, and lemmatization convert raw text into meaningful features that can be processed by the machine learning algorithm.

The core of the proposed system is a **Decision Tree classifier**, chosen for its simplicity and interpretability. The model is trained on the pre-processed data to learn patterns that distinguish normal content from cyberbullying messages through hierarchical decision rules.

In the **evaluation phase**, the trained model is tested using unseen data to measure performance metrics such as accuracy, precision, recall, and F1-score. This step ensures the model generalizes well and performs reliably on new inputs.

Finally, the trained model is suitable for **deployment in real-time environments**, where it can continuously monitor social media content. Upon detecting cyberbullying, the system can alert moderators, remove harmful content, or restrict user activity, enabling timely intervention and promoting a safer online environment.

1V. IMPLEMENTATION AND ALGORITHM

The implementation of the cyberbullying detection system follows a structured machine learning pipeline designed to accurately identify harmful content on social media platforms. The system integrates text analysis techniques with supervised learning methods, primarily using a Decision Tree classifier. The overall process includes dataset collection, preprocessing, feature extraction, model training, and evaluation.

1. Dataset Collection

A labelled dataset is obtained from the Kaggle repository, consisting of social media text categorized as *normal* or *cyberbullying*. These labelled samples form the basis for supervised learning, enabling the model to learn patterns associated with abusive and non-abusive content.

2. Data Preprocessing

Social media text is unstructured and noisy, requiring preprocessing before analysis. This stage includes tokenization, removal of stop words, and text normalization such as converting text to lowercase and eliminating special characters, URLs, and punctuation. Stemming or lemmatization is applied to reduce words to their root form, ensuring

ISSN: 3107-6513

consistency and improving classification performance.

3. Feature Extraction

To enable machine learning processing, textual data is transformed into numerical features. Techniques such as Bag of Words (BoW) and TF-IDF are used to represent word importance and frequency within the dataset. These methods help capture meaningful patterns that distinguish cyberbullying content from normal text.

4. Model Training

The processed dataset is divided into training and testing subsets. A Decision Tree classifier is trained using the training data, where it learns decision rules based on key features. The model recursively splits data into nodes until optimal classification rules are formed, allowing clear differentiation between normal and abusive text.

5. Testing and Evaluation

The trained model is evaluated using unseen test data to measure its effectiveness. Performance metrics such as accuracy, precision, recall, and F1-score are used to assess the model's ability to correctly identify cyberbullying content while minimizing misclassification.

6. Decision Tree Algorithm

The Decision Tree algorithm is a supervised learning method known for its interpretability and effectiveness in text classification tasks. It selects features based on metrics such as Information Gain or Gini Index and constructs a hierarchical structure of decision rules. Each leaf node represents a final classification, making the model transparent and easy to interpret.

7. Natural Language Processing Techniques

NLP techniques enhance the model's understanding of text by converting raw language into structured features. Tokenization divides text into words, stop word removal reduces noise, and stemming or lemmatization standardizes word forms. TF-IDF weighting is applied to highlight important terms within the dataset:

V. RESULTS AND FINDINGS

The proposed machine learning system effectively detected cyberbullying in social media text using a labelled Kaggle dataset. After preprocessing and feature extraction with TF-IDF and sentiment analysis, the Decision Tree classifier accurately distinguished cyberbullying content from normal messages.

The model achieved **92% accuracy**, **88% precision**, **85% recall**, and an **86% F1-score**. It also supported real-time analysis, enabling quick identification of harmful content. Overall, the results demonstrate that the system is reliable and suitable for enhancing online safety.

VI. CONCLUSION

Machine learning provides an effective solution for detecting cyberbullying on social media. The developed system combines NLP techniques with a Decision Tree classifier to identify harmful content accurately. Preprocessing methods like tokenization

ISSN: **3107-6513**

and stop-word removal, along with features such as TF-IDF and sentiment analysis, enhance detection performance. The model offers interpretable results, real-time monitoring, and scalability, making it suitable for large-scale deployment. Overall, this approach contributes to creating safer and more supportive online environments.

V11. REFERENCES

1. Wang, L., et al. (2022). Hybrid deep learning models for cyberbullying detection. *IEEE Journal on Computational Social Systems*, 8(4), 789–798.
2. Sharma, K., et al. (2020). Detecting cyberbullying with convolutional neural networks across social platforms. *IEEE DSAA Conference Proceedings*.
3. Patel, D. K., et al. (2022). Real-time detection of harmful social media content using decision trees. *Applied Artificial Intelligence Journal*, 18(3), 241–256.
4. Di Capua, M., et al. (2020). Unsupervised learning for cyberbullying detection using social and textual features. *Journal of Advanced Computing Systems*, 12(3), 245–258.
5. Davidson, T., et al. (2017). Automated detection of hate speech and offensive language. *AAAI ICWSM Conference Proceedings*.
6. Zhou, H., et al. (2020). Survey on text preprocessing techniques in cyberbullying detection. *Transactions on Natural Language Processing*, 10(5), 326–339.
7. Verma, N., et al. (2021). NLP-based approaches for cyberbullying detection. *International Journal of Data Science and Analytics*, 11(2), 147–160.
8. Gupta, S., et al. (2019). Comparison of machine learning models for abusive language detection on social media. *Journal of Intelligent Systems and Machine Learning*, 7(2), 112–121.
9. Mishra, P., et al. (2020). Machine learning approaches for automated cyberbullying detection. *International Conference on Social Computing and Applications (SCA)*.
10. Zhao, R., Zhou, A., & Mao, K. (2016). Automatic cyberbullying detection in social networks based on bullying characteristics. *ACM Conference Proceedings*.