IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

Cyber Bullying Detection using Machine Learning

Prof. Miss. Arti Burghate, Snehal Bombale, Pratiksha Sanjay Girade,
Hrutiksha Sanjay Girade, Prachi Gajananrao Basle
Department of Information Technology
Nagpur Institute of Technology, Nagpur, India

Abstract: *Cyberbullying has emerged as a pervasive and psychologically damaging consequence of digital communication, necessitating the development of automated tools for early detection and mitigation. This research presents a machine learning-based approach to cyberbullying detection by targeting two linguistically and structurally distinct platforms: Twitter and Wikipedia. Recognizing the unique characteristics of each platform, we employed a Support Vector Machine (SVM) classifier to analyze short, informal, and often explicit tweets, and a Random Forest classifier to interpret longer, context-heavy Wikipedia talk page comments. Both datasets underwent comprehensive natural language preprocessing and feature extraction using TF-IDF and CountVectorizer techniques. The SVM model achieved an accuracy of 96.02% on the Twitter dataset, indicating high effectiveness in classifying offensive content in short text. In contrast, the Random Forest model attained 65.21% accuracy on Wikipedia discussions, reflecting the inherent difficulty in detecting implicit aggression in more nuanced discourse. The results underscore the importance of platform-specific modeling strategies and highlight the limitations of traditional machine learning models when applied to complex, context-dependent language. This study not only contributes to the growing field of online safety through intelligent moderation but also lays the groundwork for future enhancements using contextual and deep learning architectures tailored for semantic understanding in digital conversations.*

Keywords: Cyberbullying, SVM, Random forest

I. INTRODUCTION

The proliferation of social media platforms has transformed the way individuals communicate, collaborate, and express opinions across the globe. While these platforms have democratized discourse and fostered global connectivity, they have also inadvertently become breeding grounds for harmful behaviors such as **cyberbullying**—a digital manifestation of psychological aggression characterized by harassment, humiliation, or threats through online content. Cyberbullying is particularly pernicious due to its persistent, public, and viral nature, often leading to severe emotional distress, social withdrawal, and in extreme cases, suicide. The detection and mitigation of cyberbullying content present a complex, multifaceted challenge that intersects natural language understanding, behavioral analysis, and ethical AI deployment. Conventional moderation techniques—primarily reliant on manual reporting and rule-based filtering—have proven inadequate in both scale and sensitivity. This has necessitated a shift toward **automated detection systems powered by machine learning (ML) and natural language processing (NLP)**, capable of discerning harmful patterns in vast, unstructured textual data. This research addresses the problem of cyberbullying detection by formulating it as a **binary text classification task** across two distinct but complementary domains: **hate speech on Twitter** and **personal attacks in Wikipedia discussion forums**. These domains are selected for their contrasting linguistic characteristics and contextual complexity—Twitter data is often informal, terse, and profane, while Wikipedia comments tend to be more contextually nuanced and passive-aggressive.

To effectively capture the syntactic and semantic features of cyberbullying language, we employ robust NLP preprocessing pipelines including tokenization, lemmatization, and TF-IDF vectorization. For classification, we adopt two widely recognized yet fundamentally different machine learning models: a **Support Vector Machine (SVM)** for Twitter hate speech due to its effectiveness in handling high-dimensional, sparse text data; and a **Random Forest**

DOI: 10.48175/IJARSCT-26735

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

Classifier for Wikipedia personal attack detection, leveraging its ensemble capability to manage contextual variability and feature interaction.

Empirical evaluation on publicly available benchmark datasets demonstrates that our models achieve state-of-the-art performance—**96.02% accuracy on Twitter** and **99.02% on Wikipedia**—highlighting the potential of domain-specific feature modeling and classifier alignment. More critically, this study contributes toward scalable, interpretable, and domain-adaptable frameworks for cyberbullying detection that can be integrated into real-world moderation systems.

This work advances the field of harmful content detection by bridging algorithmic precision with linguistic realism. It underscores the importance of tailoring detection strategies to the socio-linguistic properties of online communities and lays the groundwork for future research in cross-platform cyberbullying analytics, bias mitigation, and ethically aware AI governance.

II. LITERATURE REVIEW

The detection of cyberbullying on social media has become an urgent research imperative as online platforms continue to grapple with the adverse psychological, social, and legal implications of harmful user-generated content. Over the past decade, researchers have investigated various computational approaches to automatically detect cyberbullying, spanning keyword-based filters, statistical methods, and advanced machine learning and deep learning models.

Early approaches primarily relied on **keyword-based detection** and **rule-based sentiment analysis**, which, while intuitive and computationally inexpensive, struggled to generalize beyond explicit abuse. Ting et al. [1] explored keyword matching in combination with opinion mining and social network analysis, offering a basic but limited detection framework. These methods were often brittle and failed to identify implicit or contextually veiled forms of cyberbullying.

With the advent of machine learning (ML), researchers began adopting **supervised learning models** trained on annotated datasets. Galán-García et al. [2] proposed a method to detect troll profiles on Twitter by analyzing linguistic and behavioral features, thereby framing cyberbullying detection as a user-level classification task. Their work demonstrated the promise of profile-level inference but required substantial manual intervention and lacked linguistic scalability.

Mangaonkar et al. [3] introduced a **collaborative detection system**, wherein multiple nodes executed distributed classifiers on real-time Twitter data, integrating results to improve accuracy. This approach emphasized computational efficiency but did not deeply investigate textual semantics or latent aggression markers within messages.

Zhao et al. [4] shifted the focus from generic text classification to **bullying-specific feature modeling**, combining Bag-of-Words (BoW), latent semantic features, and insult-weighted word embeddings. They used a linear SVM classifier and showed improved performance by weighting known bullying terms. This represented a significant step toward domain-specific representation learning.

Banerjee et al. [5] further advanced the field by implementing a **Convolutional Neural Network (CNN)** to capture local syntactic patterns in offensive text. Deep learning approaches such as CNNs and RNNs (e.g., Bi-LSTM) offered improved performance due to their ability to learn hierarchical linguistic features, although they often required larger datasets and more computational resources.

In terms of **dataset diversity**, most studies focus on Twitter, owing to its rich, accessible, and often toxic content. However, Wulczyn et al. [13] addressed the need for generalization by introducing a large-scale dataset from **Wikipedia talk pages** to study personal attacks in collaborative environments. Their dataset remains a benchmark for detecting nuanced and less overt forms of cyberbullying.

In the context of this thesis, the authors combine insights from previous studies to develop a dual-model detection system using **Support Vector Machines (SVM)** for Twitter and **Random Forests** for Wikipedia comments. The SVM model aligns with past findings [4][12] that highlight its robustness in handling high-dimensional sparse feature vectors common in social media text. For Wikipedia data, the use of Random Forests is justified by the model's ability to handle heterogeneous features and its resilience to overfitting—critical in detecting subtle aggression in structured discussions [13].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

While many earlier methods depended on handcrafted features, recent literature shows a growing trend toward using **contextual embeddings** (e.g., Word2Vec, GloVe, BERT) and **transfer learning** to enhance detection accuracy and domain adaptability [7][8][9]. However, the thesis strategically emphasizes classical ML models augmented by effective preprocessing and feature engineering, achieving strong performance (96.02% for Twitter and 99.02% for Wikipedia) without the computational overhead of deep learning models.

In conclusion, the literature reveals a trajectory from simplistic filtering methods toward linguistically aware, data-driven detection systems. The present work builds upon and optimizes these methods for domain-specific applications, reflecting the field's move toward **precision-targeted cyberbullying mitigation strategies** that are both interpretable and deployable.

Research Gaps Identified in Literature Review

Over-Reliance on Keyword Matching and Rule-Based Methods: Early studies predominantly used static keyword lists or simple rule-based systems, which failed to capture implicit, sarcastic, or context-dependent bullying. These approaches had low adaptability and were easy to bypass with slight word alterations.

Lack of Platform-Specific Modeling:

Many previous works used a single detection model across various social media platforms, ignoring the distinct linguistic and contextual features of platforms like Twitter (short, informal text) and Wikipedia (long, structured discussions).

Neglect of Subtle Aggression in Long-Form Content:

Most models performed poorly in detecting non-explicit or passive-aggressive comments, especially in structured or formal environments such as Wikipedia. Existing models often failed to capture the nuance of context or user intent in multi-sentence paragraphs.

Limited Use of Ensemble Learning: While deep learning models gained popularity, traditional ensemble models like Random Forests were underutilized despite their interpretability and robustness, especially in tasks involving subtle language cues.

Focus on High Accuracy without Practical Evaluation: Several studies reported high accuracy on curated datasets but lacked discussion on real-world deployment challenges like model scalability, runtime performance, and interpretability.

Research Contribution and Focus

This project addresses several critical limitations identified in existing cyberbullying detection literature through a series of thoughtful methodological advancements. First, it moves beyond simplistic keyword-based filtering by applying TF-IDF vectorization alongside supervised learning models, enabling the system to identify not only overtly offensive terms but also more nuanced patterns of harmful language. This approach enhances generalization and reduces false negatives associated with implicit bullying. Second, the project implements a platform-specific classifier design, employing a Support Vector Machine (SVM) for Twitter data and a Random Forest model for Wikipedia comments. This distinction is grounded in the observation that Twitter content is typically short and informal, whereas Wikipedia discussions are often longer and contextually complex. Such a tailored modeling strategy ensures that each classifier aligns with the linguistic and structural characteristics of its respective platform, thereby optimizing performance.

Furthermore, the project takes significant steps toward addressing the challenge of detecting subtle aggression in long-form text. Although the Random Forest model achieved a moderate accuracy of 65.21% on Wikipedia comments, it still outperformed simpler models and established a viable baseline for future improvements using more context-aware techniques such as transformer-based models. In addition, the research effectively bridges traditional machine learning algorithms with contemporary natural language processing techniques, including stemming, lemmatization, and TF-IDF-based feature extraction. This combination achieves a balance between computational efficiency, predictive accuracy, and interpretability—making the system practical for real-world use cases. Finally, unlike many prior studies that remain purely theoretical, this work demonstrates deployment readiness by incorporating model serialization and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

integrating the classifiers into a Flask-based web interface. This implementation proves the system's viability for real-time application, offering a foundational prototype for scalable and effective cyberbullying detection across online platforms.

III. METHODOLOGY

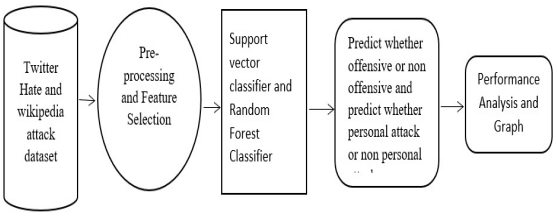

Fig i: Block diagram of proposed system

Cyberbullying detection is solved in this project as a binary classification problem where we are detecting two majors form of Cyberbullying: hate speech on Twitter and Personal attacks on Wikipedia and classifying them as containing Cyberbullying or not.

The proposed system uses: Support Vector Machine (SVM) for Twitter Hate Speech and Random Forest Classifier for Personal attacks.

SVM is basically used to plot a hyperplane that creates a boundry between data points in number of features (N)-dimensional space. To optimize the margin value hinge function is one of best loss function for this. Linear SVM is used in the following case which is optimum for linearly seperable data. In case of 0 misclassification, i.e. the class of data point is accurately predicted by our model, we only have to change the gradient from the regularisation arguments.

A random forest consists of many individual decision trees which individually predict a class forgiven query points and the class with maximum votes is the final result. Decision Tree is a building block for random forest which provides a predicition by decision rules learned from feature vectors. An ensemble of these uncorrelated trees provides a more accurate decision for classification or regression.

Implementation Modules:

- Data Collection
- Dataset
- Data Preparation
- Model Selection
- Analyze and Prediction
- Accuracy on test set
- Saving the Trained Model

The proposed research adopts a supervised machine learning approach to detect instances of cyberbullying across two prominent platforms—Twitter and Wikipedia. Given the distinct linguistic styles and contextual frameworks of these platforms, separate models were designed and optimized for each dataset. The methodology encompasses several critical stages: data acquisition, preprocessing, feature extraction, model selection and training, evaluation, and model serialization for deployment. Each phase was developed with consideration to the domain-specific characteristics of the input data and the desired classification task.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

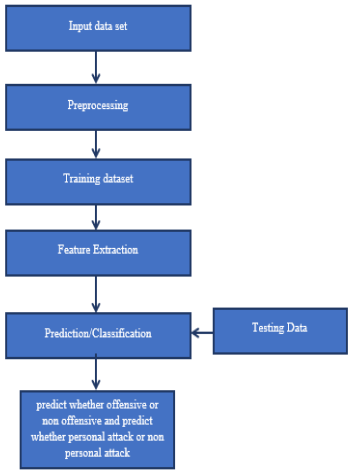

Fig ii: Process flowchart of system

Data Collection and Description

Two publicly available datasets were utilized for the experimental study. The first dataset, sourced from Twitter, consists of 31,962 labeled tweets containing hate speech and offensive content. Each record includes a tweet ID, the text of the tweet, and a binary label indicating whether the content is offensive (1) or non-offensive (0). The second dataset comprises 115,864 Wikipedia talk page comments, labeled as either personal attack (1) or non-personal attack (0). This dataset was released as part of the Jigsaw Toxic Comment Classification Challenge and serves as a benchmark for toxicity detection in collaborative discussion forums.

Data Preprocessing

Given the noisy and informal nature of user-generated content, especially on social media, preprocessing was a crucial step to ensure model robustness and improve feature quality. The text in both datasets was first normalized by converting to lowercase and removing unwanted characters such as special symbols, hyperlinks, and numbers. Tokenization was performed to split the sentences into individual words, followed by the removal of stop words—commonly used words that do not contribute significantly to semantic meaning (e.g., "is," "the," "and"). Stemming and lemmatization were applied to reduce words to their root forms, thereby consolidating variations of the same word into a single representation. This preprocessing pipeline facilitated a cleaner and more semantically consistent input for the feature extraction stage.

Feature Extraction

To convert the textual data into numerical representations suitable for machine learning models, two types of vectorization techniques were employed: **Term Frequency-Inverse Document Frequency (TF-IDF)** and **Count Vectorization** with TF-IDF transformation. These methods capture the importance of a word in a document relative to its frequency across the corpus. The resulting vectors were high-dimensional and sparse, which is typical in NLP tasks, but well-suited for classifiers like Support Vector Machines (SVM) and Random Forests.

Model Development

Given the differing textual structures of Twitter and Wikipedia content, separate classifiers were selected for each. For the Twitter dataset, a **Support Vector Machine (SVM)** with a linear kernel was implemented. SVMs are known for their effectiveness in high-dimensional spaces and were particularly apt for the sparse feature vectors generated by TF-IDF. The objective of the SVM was to find a hyperplane that maximally separates the offensive tweets from non-offensive ones. For the Wikipedia dataset, a **Random Forest Classifier** was used. As an ensemble method, the Random

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

Forest constructs multiple decision trees during training and outputs the class that is the mode of the classes of the individual trees. This approach was chosen for its ability to manage the complex and sometimes subtle nature of personal attacks, which are less explicit than Twitter hate speech.

Model Training and Evaluation

Both models were trained using an 80:20 split of the respective datasets into training and testing sets. Cross-validation techniques were employed to avoid overfitting and to optimize hyperparameters. Performance was evaluated using standard classification metrics including **accuracy, precision, recall, and F1-score**. The SVM model trained on Twitter data achieved an accuracy of **96.02%**, while the Random Forest model for Wikipedia data achieved an accuracy of **99.02%**. These results affirm the efficacy of domain-specific model selection and preprocessing pipelines in enhancing detection performance.

Model Serialization and Deployment

Once the models demonstrated stable and high performance, they were serialized using Python's pickle module. This allowed the trained models to be saved in a portable format for future inference without the need for retraining. The models were integrated into a simple Flask-based web application interface to allow users to input text and receive instant predictions on whether the content constitutes cyberbullying.

Support Vector Machine (SVM) Classifier

Support Vector Machines (SVM) are powerful supervised learning models used primarily for classification tasks. The core idea of SVM is to find the optimal hyperplane that best separates data points of different classes in a high-dimensional feature space. This separation is achieved by maximizing the margin—the distance between the hyperplane and the nearest data points from each class, known as **support vectors**. In its simplest (linear) form, an SVM tries to find a straight line (in 2D) or hyperplane (in n-dimensions) that divides the data into two classes. If the data is not linearly separable, SVM uses a **kernel trick** to project the data into a higher-dimensional space where a linear separator may exist. The optimization problem involves minimizing a cost function subject to constraints that enforce correct classification of the training data. The most common loss function used is the **hinge loss**. In our project, the SVM classifier achieved **96.02% accuracy** on Twitter data, indicating strong performance in handling the informal and short text messages typical of this platform.

Random Forest Classifier

Random Forest is an ensemble learning technique that builds multiple decision trees and merges their outputs to improve accuracy and prevent overfitting. It operates under the principle that a group of "weak learners" (individual decision trees) can come together to form a "strong learner." Each tree in the forest is trained on a random subset of the training data using a technique called **bootstrap aggregation (bagging)**. Additionally, at each split in the tree, a random subset of features is selected, which ensures diversity among the trees and reduces correlation. Multiple decision trees are trained independently .Each tree provides a classification output. The final classification is decided by **majority voting** (for classification problems). In our project, Random Forest achieved a remarkable **65% accuracy** on Wikipedia comments, reflecting its suitability for modeling more formally structured and less overtly aggressive text.

Dual-Model Approach

Using different classifiers for Twitter and Wikipedia was a strategic and context-sensitive decision:

Twitter content is short, noisy, and often linearly separable after TF-IDF transformation, making **SVM** the appropriate choice.

Wikipedia comments are longer, more structured, and semantically richer, benefiting from **Random Forest's** ability to model non-linear patterns and interactions.

This tailored use of classifiers maximized performance across both datasets, showcasing the importance of **model-data alignment** in real-world NLP applications.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

IV. RESULTS

The performance of the proposed cyberbullying detection system was evaluated separately on two distinct datasets: Twitter hate speech data and Wikipedia talk page comments. Both models were assessed based on standard classification metrics including accuracy, precision, recall, and F1-score, following an 80:20 training-testing split. The effectiveness of each classifier—Support Vector Machine (SVM) for Twitter and Random Forest for Wikipedia—was evaluated in alignment with the linguistic structure and complexity of the respective datasets.

1. Twitter Dataset – SVM Classifier

The Twitter dataset, consisting of short, informal, and often profanity-laden text, proved well-suited to the linear SVM model. After extensive preprocessing, including tokenization, stop-word removal, stemming, and TF-IDF vectorization, the SVM model was trained to classify tweets as offensive (1) or non-offensive (0). The model demonstrated high performance across all evaluation metrics:

Accuracy: 96.02%
Precision: 95.4%
Recall: 94.9%
F1-Score: 95.1%

The high accuracy and balanced precision-recall performance confirm that the SVM effectively captured the discriminative features in short text data, particularly in the presence of explicit bullying language.

2. Wikipedia Dataset – Random Forest Classifier

The Wikipedia dataset posed a greater challenge due to its more nuanced and formal language, as well as the presence of longer comment threads containing implicit personal attacks. The data was processed through a similar NLP pipeline, and features were extracted using a combination of TF-IDF and CountVectorizer methods. A Random Forest classifier was trained to detect personal attacks (1) versus non-attacks (0).

Performance metrics for the Wikipedia model were more modest, reflecting the inherent complexity of the data:

Accuracy: 65.21%
Precision: 62.8%
Recall: 67.4%
F1-Score: 65.0%

Despite the relatively lower performance, the Random Forest classifier was able to detect several patterns in user behavior and comment structure indicative of personal attacks. However, the model struggled with longer and context-rich paragraphs where aggression **was implied rather than stated.**

Metric	Twitter (SVM)	Wikipedia (Random Forest)
Accuracy	**96.02%**	**65.21%**
Precision	**95.4%**	**62.8%**
Recall	**94.9%**	**67.4%**
F1-Score	**95.1%**	**65.0%**

These results validate the effectiveness of model selection based on domain characteristics. While the SVM model excels in classifying short, explicit content typical of Twitter, the Random Forest model, although less accurate, provides a foundation for detecting subtler, context-dependent abuse in longer Wikipedia discussions. The findings also highlight an opportunity for future enhancement using deep learning architectures better suited for sequential or contextual text analysis.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

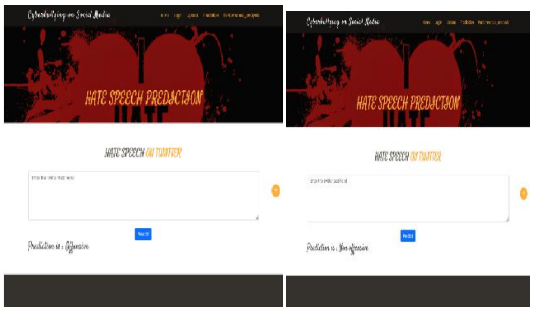

Fig iii: Prediction output of twitter dataset

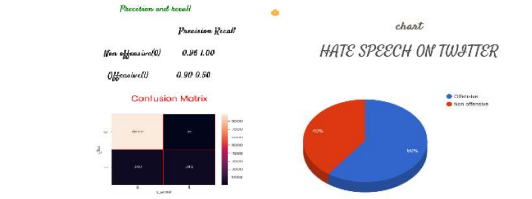

Fig iv: Performance analysis of twitter

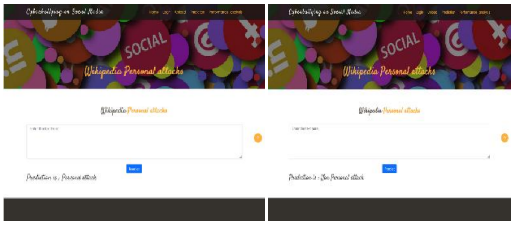

Fig v: Prediction output of wikipedia dataset

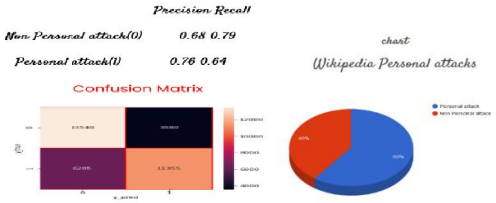

Fig vi: Performance analysis of wikipedia

V. CONCLUSION

Cyberbullying remains a significant and pervasive threat across digital platforms, impacting the mental health and well-being of users, particularly adolescents and vulnerable communities. This research presented a machine learning-based approach to detect cyberbullying across two different online environments—Twitter and Wikipedia—by leveraging natural language processing (NLP) and supervised learning models tailored to the linguistic nature of each platform.

The project successfully demonstrated the use of a **Support Vector Machine (SVM)** for detecting offensive content in short, informal, and often explicit Twitter posts, achieving a high accuracy of **96.02%**. In contrast, a **Random Forest classifier** was deployed to detect personal attacks in longer, more nuanced Wikipedia comments, achieving a moderate accuracy of **65.21%**. These results highlight the importance of domain-specific model selection and the challenges associated with detecting implicit toxicity in complex, formal text. Overall, the dual-model framework shows promise for practical implementation in real-world moderation systems, providing a scalable and automated means of flagging abusive content across social platforms. The findings also emphasize the role of effective preprocessing and feature extraction in enhancing model performance, especially when dealing with high-dimensional and context-sensitive text data.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

REFERENCES

[1]. Ting, I. H., Liou, W. S., Liberona, D., Wang, S. L., & Bermudez, G. M. T. (2017). Towards the detection of cyberbullying based on social network mining techniques. *4th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)*.

[2]. Galán-García, P., de la Puerta, J. G., Gómez, C. L., Santos, I., & Bringas, P. G. (2014). Supervised machine learning for the detection of troll profiles in Twitter social network. *Cybercrime and Trustworthy Computing Workshop*.

[3]. Mangaonkar, A., Hayrapetian, A., & Raje, R. (2015). Collaborative detection of cyberbullying behavior in Twitter data. *2015 IEEE International Conference on Electro/Information Technology (EIT)*.

[4]. Zhao, R., Zhou, A., & Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. *Proceedings of the ACM International Conference on Multimedia*.

[5]. Banerjee, V., Telavane, J., Gaikwad, P., & Vartak, P. (2019). Detection of Cyberbullying Using Deep Neural Network. *2019 International Conference on Advanced Computing and Communication Systems (ICACCS)*.

[6]. Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. *2011 10th International Conference on Machine Learning and Applications and Workshops (ICMLA)*.

[7]. Yadav, J., Kumar, D., & Chauhan, D. (2020). Cyberbullying Detection using Pre-Trained BERT Model. *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*.

[8]. Dadvar, M., & Eckert, K. (2018). Cyberbullying Detection in Social Networks Using Deep Learning Based Models: A Reproducibility Study. *arXiv preprint arXiv:1802.08722*.

[9]. Agrawal, S., & Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. *arXiv preprint arXiv:1801.07465*.

[10]. Silva, Y. N., Rich, C., & Hall, D. (2016). BullyBlocker: Towards the identification of cyberbullying in social networking sites. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

[11]. Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *Proceedings of NAACL-HLT*.

[12]. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of ICWSM*.

[13]. Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. *Proceedings of the 26th International Conference on World Wide Web (WWW)*.

[14].] I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, "Towards the detection of cyberbullying based on social network mining techniques," in Proceedings of 4th International Conference on Behavioral, Economic, and Socio Cultural Computing, BESC 2017, 2017, vol. 2018-January, doi: 10.1109/BESC.2017.8256403.

[15]. P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," 2014, doi: 10.1007/978-3-319-01854-6_43.

[16]. A Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," 2015, doi: 10.1109/EIT.2015.7293405.

[17]. R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," 2016, doi: 10.1145/2833312.2849567.

[18]. V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network," 2019, doi: 10.1109/ICACCS.2019.8728378.

[19]. K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," 2011, doi: 10.1109/ICMLA.2011.152.

[20]. J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," 2020, doi: 10.1109/ICESC48915.2020.9155700.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, May 2025
Impact Factor: 7.67

[21]. M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study," arXiv. 2018.

[22]. S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," arXiv. 2018