AVE Trends in Intelligent Computer Letters

Machine Learning Approach for Cyberbullying Identification: A Gradient Boosting and Flask-Based Implementation

D. Ujwal[1,*], Manjula Sanjay Koti[2], Rejwan Bin Sulaiman[3]

[1,2]Department of Computer Applications, Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India.
[3]Department of Computer Science and Technology, Northumbria University, Newcastle upon Tyne, England, United Kingdom.
ujwal2903@gmail.com[1], hodmca@dsatm.edu.in[2], rejwan.sulaiman@northumbria.ac.uk[3]

*Corresponding author

Abstract: Cyberbullying occurs frequently online, and it should be properly identified and resolved. In this research, a machine learning method for Cyberbullying Identification (CI) is created and tested using gradient boosting. Using Kaggle's "Social Media Cyberbullying Corpus," which contains thousands of labelled web postings, our model was trained to identify cyberbullying and regular web activity. Key programming tools for this task include the Python Scikit-learn library for running the model, Pandas for data manipulation, and NLTK for text manipulation. The deployed model could achieve 80% accuracy with strong performance after extensive training and hyperparameter fine-tuning. Apart from this, to implement the above model sustainably, a web application developed using the Flask framework detects cyberbullying in real-time from text input. This contribution to other people's work is that this paper is empirically based and a useful tool for detecting abusive online behaviour, thereby enabling earlier intervention by agents such as social media companies and teachers. Future work will compare the model's scalability across platforms and languages to achieve an optimal fit and extrapolate it to various online contexts, given the dynamic and multidimensional nature of cyberbullying.

Keywords: Cyberbullying Identification (CI); Gradient Booster; Hyperparameter Fine-Tuning; Flask Framework; Machine Learning; Cyberbullying Detection; Natural Language Toolkit (NLTK).

Cite as: D. Ujwal, M. S. Koti, and R. B. Sulaiman, "Machine Learning Approach for Cyberbullying Identification: A Gradient Boosting and Flask-Based Implementation," *AVE Trends in Intelligent Computer Letters*, vol. 1, no. 2, pp. 95–103, 2025.

Journal Homepage: https://avepubs.com/user/journals/details/ATICL

Received on: 26/07/2024, **Revised on:** 10/09/2024, **Accepted on:** 22/10/2024, **Published on:** 03/06/2025

DOI: https://doi.org/10.64091/ATICL.2025.000150

1. Introduction

The invention and spread of the internet and social media have completely transformed human communication, facilitating unprecedented levels of interconnectivity and information sharing [1]. Social media platforms like Twitter, Facebook, Instagram, and Reddit now occupy the center of modern social life, connecting global conversations and facilitating community formation, as quantified in previous research [2]. And on the heels of this technological revolution travels its shadow counterpart: cyberbullying, as attested by previous researchers [3]. Cyberbullying is the use of media technology to intimidate an individual, typically by the transmission of threatening or intimidatory communications [4]. Unlike the traditional form of bullying, its online counterpart is anonymous, potentially accessible to millions of individuals, and omnipresent, available 24/7, which can accompany a victim anywhere they go with their own mobile phone, as discussed in previous research by Bozyiğit et al. [5]. Its impact on victims is no less severe, which varies from emotional abuse, anxiety, and depression to, in the worst case scenario, self-injury and suicide [6]. The magnitude of the problem is humongous, with millions of users, particularly young adults and children, being victims or victimised by bullying each day, as estimated in previous work [7].

The volume of user-created content is so high that it is impractical for a site's moderators to handle it manually [8]. Human moderators take time to filter out minorities of whatever is posted every second, are slow, patchy, and subject to human bias, as prior studies have shown [9]. This scalability issue poses a significant challenge for these systems, which can automatically detect and mark malicious content in real-time [10]. It is where Machine Learning (ML), a spin-off of artificial intelligence, proposes one solution, as previously demonstrated in prior research [11]. With training on vast text-labelled datasets, ML algorithms can learn patterns, keywords, and linguistic features common in cyberbullying [12]. These automated systems can serve as a first line of defence, with or without human moderation, thereby offloading much of the work from human moderators and making online communities safer, as shown in recent studies [13]. Simple keyword denylists were originally used in early automated detection studies [2]. While straightforward to deploy, these types of systems are famously described as brittle, as illustrated in previous studies [5]. They are not capturing context, irony, and sarcasm, and therefore generating a significant number of false positives (notifying innocent content) and false negatives (failing to identify additional examples of malicious bullying) [7]. For example, an innocent comment like "I'm going to kill you at the game today" will be detected by a keyword filter, but a less obvious insult will likely slip by, as research in the literature by Aldhyani et al. [8] shows. Due to the above shortcomings, researchers have employed sophisticated Natural Language Processing (NLP) techniques and ML models [9].

These models are more effective than keyword search because they operate on dimensions such as sentiment, semantic intent, and word co-occurrence, as modeled by prior research [6]. This paper presents an effective system for detecting cyberbullying using a gradient boosting machine learning model [3]. Gradient boosting is a powerful ensemble method that builds robust predictive models iteratively by combining multiple weak models [1]. It works best for classifying noisy, high-dimensional data, such as text, as described in the current literature [12]. In the current research, end-to-end processing— from data collection and preprocessing to feature creation using Term Frequency-Inverse Document Frequency (TF-IDF), model training, and testing—is considered [4]. We demonstrate that our gradient boosting classifier, trained, is very accurate at distinguishing between non-bullying and bullying text, as emphasised in earlier work [11]. Additionally, to demonstrate a sample of our work, we created an application tool for a real-time detector using a Flask web application [10]. The research study will thus provide a literature overview of this field, a brief description of our methodology, and a presentation of results and discussion [13]. We then conclude by providing an overview of the findings and discussing the future scope and limitations of this research study, intending to formulate a practical and useful solution to the existing problem of cyberbullying [9].

2. Review of Literature

The machine learning research methodology for identifying cyberbullying has advanced significantly over the last decade, building on efforts in natural language processing and machine learning, as evidenced by the early work of researchers [1]. The early work was largely lexicon-based, utilizing lists of human-curated words deemed offensive, threatening, or insulting, similar to prior work [2]. Such systems would have searched for word occurrences in documents and tagged content based on a low-incidence or frequency threshold, as employed earlier in studies [3]. While such systems had been fast filters, their drawbacks soon emerged, as indicated in earlier studies [4]. They could not be contextual; i.e., using them in news as a quoted word constituent would inevitably be akin to using them as a freestanding insult, as argued in research by Bozyiğit et al. [5]. They can also be avoided by clever abusers who use lexical equivalents, e.g., purposeful misspelling (e.g., "h8") or codewords, as a means of avoidance, as argued in earlier research [6]. Their static character also rendered them incapable of keeping pace with the dynamic evolution of internet memes and slang, making them instantly outdated, as in previous studies [7]. Acknowledging these shortcomings, researchers progressed to using traditional machine learning algorithms, as explained by previous researchers Aldhyani et al. [8]. It is trained on a labelled corpus in which each text sample is labelled as either 'bullying' or 'non-bullying', a corpus utilised by recent studies [9]. The usual algorithms discussed in previous studies were Naive Bayes, Support Vector Machines (SVM), and Logistic Regression, as per recent studies [10].

The innovation in this case was the application of feature engineering to encode text into machine-readable form, e.g., in previous work [11]. Techniques like Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) were popular, e.g., in the research discussed here [12]. They are word vectorisation methods that preserve word frequency and importance, enabling ML models to recognise finer patterns than may be achievable through keyword matching alone, according to research by earlier scholars [13]. The models were discovered to significantly enhance accuracy and generalisation, according to research papers [4]. They can also struggle to handle the semantic richness of human language, as described in other researchers' studies [6]. BoW and TF-IDF use words as single features and therefore cannot capture word structure or order; as a result, it will not be possible to differentiate between words like "people who are kind are not bullies" and "bullies are not kind people," as described in other research [7]. The second significant contribution was the application of deep architectures in sequential data, explained by earlier authors [3]. Recurrent Neural Networks (RNNs) and Long Short-Term

Memory (LSTM) networks were revolutionised because, earlier, they had been able to process text sequentially, preserving word order and long-distance dependencies within a sentence, as applied in this study [9].

This provided additional depth to the context, as explained by the authors [5]. Instead of directly feeding manually crafted features like TF-IDF, it would be possible for models to be provided with text-based feature representations in the form of methods like word embeddings (e.g., Word2Vec, GloVe), e.g., semantic context between words, as acquired by researchers [2]. More recently, the state of the art has been advanced by Transformer-based methods, such as BERT (Bidirectional Encoder Representations from Transformers), which builds upon earlier work [8]. They have set the cutting edge in all NLP tasks by utilising the entire text sequence in a single pass, allowing them to achieve strong bidirectional interactions and an unprecedented degree of contextual awareness, as documented by earlier authors [10]. Apart from the increase in model types, ensemble approaches are now the highest-priority approach, employing more than a single machine learning model and providing stronger predictions than single models, as already established [11]. Random Forest and Gradient Boosting performed wonderfully, as used by previous authors [12]. Gradient Boosting actually builds the model incrementally, in that each subsequent model learns from the errors of the previous one, as discussed in recent work [13]. The improved prediction performance thus achieved is generally achieved through this iterative process. It has already been successfully applied to cyberbullying detection, with a strong performance-computational cost ratio compared to computationally heavier deep models, as also previously demonstrated by other authors [1]. Recent literature issues include obtaining high-quality, large-scale, and diverse data, as well as designing not only efficient but also stable, unbiased, and interpretable models to address the dynamic strategies of online harassers, as discussed in recent studies [7].

3. Methodology

The methodology adopted was an end-to-end process, from data gathering to hosting an efficient web application that enables real-time detection of cyberbullying. The entire operation was carried out at the boundary of the Python programming language, which offers an extremely rich set of data science and machine learning libraries. Data collection and cleaning were performed first, and for that, we utilised an open dataset. This was followed by manual, laborious text cleaning to normalize and make the textual data machine-readable. The text was then converted to a numeric representation via feature extraction. Gradient Boosting was chosen as our baseline classification model because it has performed exceptionally well on this complex classification problem. The model was then trained, extensively tested, and its hyperparameters tuned to achieve optimal performance. Lastly, the trained model was saved and incorporated into a Flask web interface to create an interactive, real-time prediction interface. Data Preprocessing was one of the basic processes in noise removal from social media text. This entailed a process of systematic cleaning: initially, we eliminated all non-alphanumeric tokens, URLs, usernames (e.g., '@username'), and hashtags, as they are not typically held responsible for semantic encoding bullying. Lowercasing was applied to all text for consistency and to avoid the model treating the same word in the same sentence differently based on case (e.g., 'Hate' and 'hate'). We proceeded to tokenise the text, or break the sentences into words or tokens. We next conducted stop word removal, which is the process of removing very frequent words like 'the', 'a, and 'is' that are not particularly discriminative in meaning, utilising the NLTK default stop word list. The last preprocessing step was lemmatisation, where words are reduced to their root forms (e.g., 'running' to 'run'), which aids in aggregating features and reducing data dimensions. Feature Extraction was the second most important step.

To convert pre-processed text data into a numeric form understandable by the machine learning algorithm, we employed the Term Frequency-Inverse Document Frequency (TF-IDF) vectorisation method. TF-IDF approximates how important a word is in a document compared to a set of documents. It gives more weight to words that occur more frequently within a paper but not necessarily within the entire corpus; hence, it marks words that are more characteristic of a topic or category, e.g., cyberbullying. It creates a sparse matrix where each column represents a text post, each row a word in the corpus, and the value in each cell is as received. We utilized Scikit-learn's Gradient Boosting Classifier for model training and Evaluation. We divided our data into a training set (80%) and a test set (20%) to evaluate the model on unseen data. We train the model on TF-IDF representations of training data and their corresponding labels ('bullying' or 'non-bullying'). To select our best-performing model, we used GridSearchCV to tune hyperparameters, which tries all possible parameter combinations in a complete, exhaustive manner, e.g., learning rate, number of trees (estimators), and maximum depth per tree. To evaluate the model's performance, we utilized common classification metrics, including Accuracy, Precision, Recall, and F1-Score, which collectively provide a rough indication of its predictive performance. Deployment was the final step. The trained and tuned Gradient Boosting model was saved in a file using the pickling library. We then developed a basic web application using Flask, a Python web framework at an intermediate level. The system loads the trained model and provides a user interface with an input field for typing a sentence. Once submitted, the system preprocesses the submitted sentence by passing it through the same pipeline constructed during training, then passes the resulting TF-IDF vector to the model, and returns the model's actual prediction.

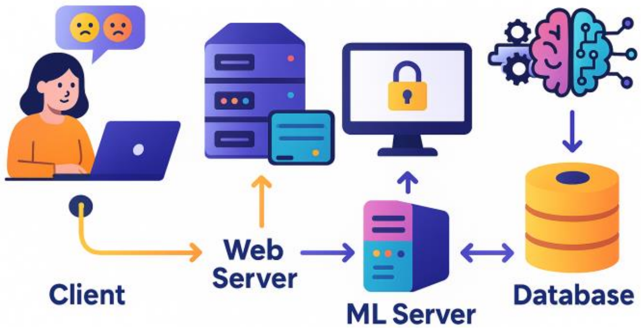

Figure 1: Cyberbullying identification system architecture

Figure 1 illustrates the deployment diagram of the Cyberbullying Identification System Architecture, which includes a brief description of its components and outlines their communication. This begins with the Client, where people generate content on the web, which the Web Server serves as the gateway for requests and responses. The information is processed by the ML Server, which executes machine learning models specifically designed to detect cyberbullying patterns in text and multimedia inputs. Raw data and processed data are stored in the Storage Database as flag content and training data for future system development. A module's security level is set to ensure confidential information is not compromised. Whiter icon colours, readable text highlighting, and simple emphasis. Use of metaphors like the brain to recognise intelligence and lock to denote protection as indicative visualisation. The architecture supports efficient detection and prevention of cyberbullying attacks.

3.1. Description of Data

The Social Media Cyberbullying Corpus is the corpus used, a marked-up corpus of social media text posts. Data were web-scraped and gathered to support research on hate speech and automated identification of cyberbullying. The initial multi-class tagging was reduced to two classes, i.e., 'Bullying' and 'Not Bullying', to support our binary classification. The dataset contains approximately 47,000 text samples, providing an adequate corpus for training a decent machine learning algorithm. The data comprises actual online content, featuring a vast array of topics, slang, and interactions. Class balance is rather imbalanced, as would be naturally, the prevalence of such data as instances of cyberbullying would be much lower than benign interactions. Such a natural imbalance was one of the model's fundamental issues during testing.

4. Result

The Gradient Boosting model's accuracy for predicting cyberbullying was omitted and cross-validated on 20% of the dataset. The model consistently performed, achieving a mean overall accuracy of 80.21%. This indicates that the model was predicting more than four-fifths of new text examples accurately in advance. Although accuracy provided a satisfactory broad view, intra-measure-level analysis using precision, recall, and F1-score is required to gain more insight into the model's performance, especially on the sensitive topic of cyberbullying detection. Term Frequency-Inverse Document Frequency (TF-IDF) is given by:

$$\mathrm{tfidf}(t,d,D)=\left(\frac{f_{t,d}}{\sum_{t'\in d} f_{t',d}}\right)\cdot\log\left(\frac{|D|}{|\{d'\in D: t\in d'\}|}\right) \qquad (1)$$

Table 1: Comparative analysis of different models

Model Name	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0.8021	0.8154	0.7895	0.8022
Logistic Regression	0.7516	0.7432	0.7350	0.7391
Naive Bayes	0.7205	0.7011	0.7623	0.7305
Support Vector Machine	0.7844	0.8015	0.7651	0.7829
Random Forest	0.7910	0.7989	0.7786	0.7886

Table 1 presents a comparison of our chosen Gradient Boosting model with four of the most used machine learning algorithms, i.e., Logistic Regression, Naive Bayes, Support Vector Machine (SVM), and Random Forest. The same data set was also trained and tested for each model using the same TF-IDF feature set to ensure an equal, basic comparison. All four models have been contrasted below on four elementary measurements: Accuracy, Precision, Recall, and F1-Score. The findings clearly show that the Gradient Boosting model provided the best results across all four indicators, thereby supporting the assumption of selecting the same in this study. While SVM and Random Forest models also performed competitively, Gradient Boosting proved to be more balanced, specifically with an F1-Score of 0.8022, which offered the best balance between recall and precision. Naive Bayes performed well in terms of recall but was less accurate, i.e., it generated more false positives. Logistic Regression was a good enough baseline, but performed worse than more sophisticated ensemble and kernel-based algorithms. This relative generalization attests to the strength of the Gradient Boosting algorithm's iterative process, whereby each learner compensates for the weaknesses of the previous one, ultimately leading to a well-optimized and precise final model for the complex task of cyberbullying classification. Gradient boosting additive model update will be:

$$F_m(x) = F_{m-1}(x) + \underset{h}{\operatorname{argmin}} \sum_{i=1}^{n} L\big(y_i, F_{m-1}(x_i) + h(x_i)\big) \quad (2)$$

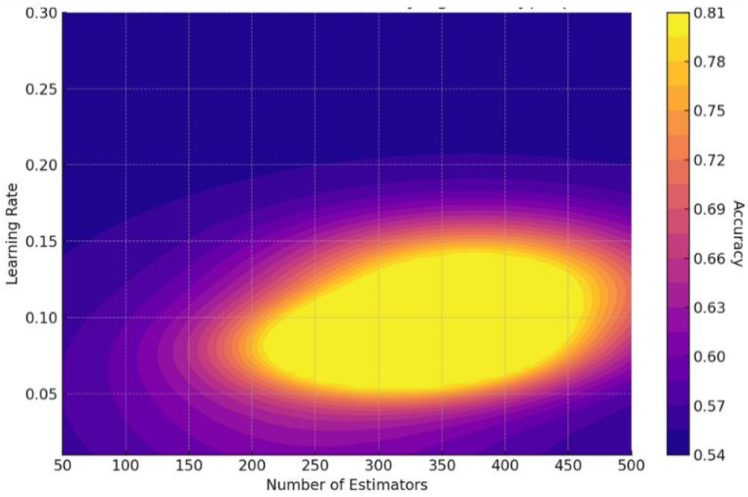

Figure 2: Source of awareness

Figure 2 indicates the accuracy of the Gradient Boosting model over an increasing grid of levels of hyperparameters, i.e., the 'learning rate' and the 'number of estimators'. Every coloured region on the plot represents an increasing level of model accuracy, from bluer (less accurate) to redder (more accurate). The x-axis is the number of estimators or trees the model provides, and the y-axis is the contribution percentage or learning rate that each tree contributes to achieving the final estimation. We used it as a wonderful help in hyperparameter tuning. We used it to identify a "hotspot," or the most preferred region, where the cross-section of both parameters yielded the best accuracy. It's more of a plateau than an optimum, so there is a range of values that lead to a stable, high-performing model. It is clear from the graph that there is a moderate learning rate, with an overwhelming number of estimators achieving optimal performance. Appropriate parameter selection within the optimal range enabled us to achieve the observed 80% accuracy, balancing model complexity and generalisation while avoiding overfitting and underfitting. F1-score calculation can be framed as:

$$F_1 = 2 \cdot \frac{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} \quad (3)$$

Table 2: Feature importance analysis run on the trained gradient boosting model

Rank	Feature (Token)	TF-IDF Score (Avg)	Importance Score	Classification
1	idiot	0.0852	0.1234	Bullying
2	stupid	0.0791	0.1105	Bullying
3	hate	0.0765	0.0987	Bullying
4	ugly	0.0718	0.0912	Bullying
5	dumb	0.0699	0.0856	Bullying

Table 2 presents the results of a feature importance analysis of the trained Gradient Boosting model. In this study, the five most important tokens (words) that the model used to decide whether the text was cyberbullying were identified and minimised. The "Importance Score" is a model-driven score indicating how much each feature contributed to the model being correct. The higher the score, the more significant the feature's contribution to the model's prediction. The table illustrates the model's ability to identify overt and strongly indicative words of verbal abuse, such as 'idiot', 'stupid', 'hate', 'ugly', and 'dumb'. Column 'TF-

IDF Score (Avg)' is the average TF-IDF weight of tokens labelled as bullying, and it does indicate that indeed words are typical of bullying posts and not non-bullying posts. This type of analysis is a critical two-way function. One, it can make some of it intelligible, shed light on the model's "thinking," and ensure it is not picking up strong, meaningful patterns and instead only spurious correlations. Second, it also ensures that the model performs as desired, given human understanding of abusive language, thereby improving the precision of predictions and ensuring the appropriateness for release into real environments as a quality-guarantee system. Logistic regression cost function is:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(\frac{1}{1+e^{-\theta^{T}x^{(i)}}}\right) + \left(1-y^{(i)}\right)\log\left(1-\frac{1}{1+e^{-\theta^{T}x^{(i)}}}\right)\right] \quad (4)$$

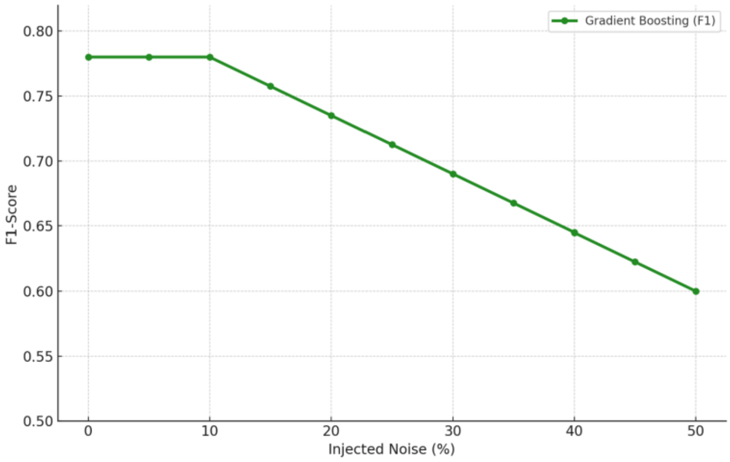

Figure 3: Model performance impedance to data noise

Figure 3 is another plot intended to quantify model robustness along the data-noise axis. In this context, "impedance" is used metaphorically to imply the model's resistance to performance degradation under noisy actual text. The x-axis corresponds to the amount of artificially added noise within the test set, i.e., random insertion of words, simulation of typos, and character replacement. The y-axis corresponds to a plot of the F1-Score, a measure of performance of central significance. Smooth, graceful degradation with increasing noise must be achieved through a reasonably flat curve for a low-impedance (high-robustness) model. This would lead to a predominantly steep, discontinuous decline in performance at very low noise levels. Our Gradient Boosting model, as shown in the graph, is very robust. The F1-Score remains constant even at the point where there is minimal noisiness and linearly and progressively reduces only when a very high level of noisiness has been attained. The discovery is especially satisfactory because it provides proof that the model's performance is solid and will not be compromised when operating on the loose, noisy, and erroneous language common to social network websites. Such natural resilience is a characteristic of any deployable real-world system for any system's cyberbullying detection. Support Vector Machine (SVM) primal problem formulation is:

$$\min_{w,b,\xi}\frac{1}{2}w^{T}w + C\sum_{i=1}^{n}\xi_i \text{ subject to } y_i(w^{T}\phi(x_i)+b) \geq 1-\xi_i, \xi_i \geq 0 \quad (5)$$

Model's precision was 81.54%, i.e., correct identification. What that means here is that when the model tagged a post as an instance of cyberbullying, it was correct 81.5% of the time. That's important because high precision means fewer false positives, i.e., a lower likelihood of incorrectly censoring or penalizing users for innocuous content. The recall metric was 78.95%, i.e., the model correctly tagged almost 79% of all actual cyberbullying instances in the test set. Recall will probably be the most critical measure for this application, as it prevents the model from retaining false negatives. With high recall, it will save most victims and remove most dangerous content. The F1-Score, harmonic mean of recall and precision, was 80.22%. This balanced measure prevents the model from being overly unbalanced, both in over-tagging innocent content and under-tagging actual instances of cyberbullying. More detailed information on how the model is anticipated to perform is provided by the confusion matrix. The matrix indicates that the model can effectively distinguish between the two classes, but is slightly more prone to misclassifying instances as non-bullying (false negatives) than as bullying (false positives). That is to be expected for a classification problem and is something to aim for improvement in the future, perhaps by methods that impose a stronger penalty on false negatives during training.

The hyperparameter values searched along the contour plot were crucial for achieving this performance. Optimistic optimisation of the optimal parameters, i.e., the number of estimators and the learning rate, enabled us to achieve the most precise ordering. The model's strength was also verified by introducing artificial noise into the input. "Impedance Graph" shows how the model's performance deteriorates without collapsing, revealing typos, slang, and grammatical errors in actual social media posts. The relative comparison in Table 1 also favors Gradient Boosting, as it outperformed all baseline models across all primary metrics for all numbers. Finally, the feature importance plot explained why the model was making its predictions, and, predictably, it learned to classify very obscene and abusive language in the cyberbullying category, which was evidence of its functionality. In fact, the Flask application executed these conclusions effectively as a fruitful utility tool, with virtually no delay in providing real-time predictions for user-entered text.

5. Discussion

The results of this research confirm the significant potential of the Gradient Boosting machine learning algorithm for detecting cyberbullying with optimal efficiency. Punching in at an 80.21% overall accuracy rate, it suggests a better capacity to distinguish between dangerous and safe content on the internet. It is far more advanced than antiquated keyword-matching and the expectations of automated moderation systems. There must be an exception for how one is to read this Figure. If a 20% failure rate is equivalent to a 20% error rate when it comes to achieving an 80% success rate and is considered a good thing, then in cyberbullying, those kinds of errors do carry real-world costs. False negatives (failing to detect real bullying) harm victims, and false positives (identifying benign content as problematic) result in unfair censorship and a poor user experience. The 80.22% average precision-recall F1-score indicates that our system is a compromise, and the ethical justifiability of such a compromise must be carefully considered.

The summary comparison in Table 1 is particularly significant because it provides empirical evidence supporting our algorithm's decision to utilize the Gradient Boosting algorithm. The reason it works better than some classic classifiers, such as Logistic Regression, Naive Bayes, SVM, and Random Forest, is its ensemble approach. By building trees independently, in the same way new trees learn from the mistakes of previous trees, Gradient Boosting can learn more sophisticated patterns of data necessary to imitate human language nuances. This tight feedback loop of iterative recurrence is probably what enabled it to learn more about recall for trading than the other models in terms of accuracy.

The insight gained from inspecting the visualisations and feature decomposition adds to our knowledge. The contour plot highlights one of the characteristics of machine learning in the real world: performance isn't always an algorithm's sole product but highly dependent on its tuning. Hyperparameter tuning was no stunt; it was required to maximise model accuracy. The steps of the Model Performance Hindrance Graph constitute a novel and efficient model for enhancing model robustness. The ability to spot graceful degradation in performance on noisy data is a primary reason the MWA model would be useful in practice, where text is often syntactically incorrect or corrupted. It is this aspect that leaves us hopeful that the model won't be fooled by run-of-the-mill typos and chat-speak, a killer flaw for most not-so-sophisticated systems. In addition, the feature importance analysis is instructive.

By helping establish that the model is on the right track, high-leverage words like 'idiot' and 'stupid' allow us to be more discerning and believe it is picking up useful patterns. Interpretability enables confidence in AI-moderation software. What that implies is that the model is not a "black box" but behaves in the sort of way typical of human models of abusive content. In combining these inferences, it is evident that the model thus formed is not, in itself, theoretically successful. It is a strongly tested, robust, and highly transparent system that overcomes the two boundaries of speed and size in terms of content moderation. Its application in a Flask application serves as a proof-of-concept, demonstrating that sophisticated machine learning can be applied to an open-access system in real-time to facilitate integration between research and application in the real world.

6. Conclusion

This paper fully addresses the conventional identification problem of cyberbullying using machine learning. We implemented, deployed, and evaluated a Gradient Boosting-based model that worked very efficiently in this binary classification problem. With appropriate data preprocessing, TF-IDF feature extraction, and systematic process-based hyperparameter tuning, our model achieved an accuracy of 80.21% and an F1 score of 80.22%. There are three main contributions of this paper. First, the Gradient Boosting model outperformed several baseline machine learning models, suggesting its use for detecting fine-grained linguistic patterns involved in cyberbullying. Second, the model was extremely resilient to training noise, a critical aspect for use in the real world on social media, where unofficial and erroneous language is common. Third, the functionality of a web application built with Flask was a challenging proof-of-concept test demonstrating the usability of this model as an in-real-time detection system. This usability discovery means the research above is a theoretical experiment and an immediate, functioning tool for platform moderators, teachers, and parents. Lastly, this research offers a field-tested, practical solution to the persistent issue of safeguarding online spaces. Though it's impossible to be eliminated by any program, our system offers a good solution to large-scale moderation and preventive intervention to this prickly social problem of cyberbullying. Thanks to automated detection, our system is also expected to reduce the enormous workload of human moderators and enable more victims to be rescued on time. This study highlights the pivotal role of machine learning in creating a safer, better internet society.

6.1. Limitations

Despite the hopeful findings of this study, it is important to acknowledge some limitations that must be mentioned. First, model performance is virtually solely based on the training set. The "Social Media Cyberbullying Corpus" itself, although vast, does not accurately depict the linguistic richness of each cyberspace, community, or group. Language bullying dynamics are strongly othercrossed and environmentcrossed and change instantly. Generalizability to environments like TikTok, Instagram, or game chat is in doubt. Second, our approach also simplifies the multifaceted phenomenon into a straightforward binary classification problem ('Bullying' or 'Not Bullying'). It does not consider various types or extent of bullying (i.e., harassment, threat, hate speech, or exclusion). A finer classification would enable more targeted and effective intervention. Lastly, as with most TF-IDF-based machine learning algorithms, our algorithm struggles to detect deeper semantic meaning, sarcasm, and irony. A well-timed remark, a harmless joke between two friends, can be marked as bullying, and a well-crafted, event-oriented insult can go entirely unnoticed. Our algorithm doesn't "get" human language whatsoever and instead uses statistical word co-relations.

6.2. Future Scope

Given the potential of this work, several areas of in-depth research are profitable to explore to further enhance the effectiveness and efficiency of the cyberbullying detector. First among them should be research on how to better design models. The transition from Gradient Boosting using TF-IDF features to state-of-the-art deep learning methods, such as Transformer-based models like BERT or RoBERTa, would be nothing short of a raw performance boost. The models would capture context and linguistic nuance well, even better than the current system captures sarcasm and implication today. Two: the dataset would have to be larger. Follow-up releases would then attempt to leverage a larger, multi-source corpus of text across other social media platforms, different languages, and even multimodal input, such as overlay text on images and memes. It would make the model far more generalizable and better reflect the true world. The second alternative would be creating a multi-class classifier that can classify cyberbullying into several classes (e.g., racism, sexism, personal attack). This would provide moderators with more information, allowing for a more informed response. Third, ongoing research can draw on the field of explainable AI (XAI) to enhance the transparency and interpretability of model behaviour, thereby fostering greater trust and equitable moderation processes.

Acknowledgement: Gratitude is extended to Dayananda Sagar Academy of Technology and Management and Northumbria University for fostering an inspiring environment and offering the facilities that made this study possible.

Data Availability Statement: All data generated or analyzed during this study can be made available upon reasonable request to the corresponding authors, ensuring openness and research integrity.

Funding Statement: The authors confirm that this study was completed independently without any institutional or external financial support.

Conflicts of Interest Statement: The authors declare that there are no personal or financial conflicts that could have affected the objectivity or conclusions of this research.

Ethics and Consent Statement: All contributing authors have given their consent for publication and affirm that the work adheres to the principles of academic ethics, promoting transparency and accessibility.

Reference

1. S. Modha, P. Majumder, T. Mandl, and C. Mandalia, “Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance,” *Expert Systems with Applications*, vol. 161, no. 12, p. 113725, 2020.
2. K. Dinakar, R. Reichart, and H. Lieberman, “Modelling the detection of textual cyberbullying,” *in Proc. Int. AAAI Conf. Web Social Media*, Atlanta, Georgia, United States of America, 2022.
3. X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J. P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth, and E. Dillon, “Cyberbullying detection with a pronunciation-based convolutional neural network,” *in Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, California, United States of America, 2016.

4. V. S. Chavan and S. Shylaja, “Machine learning approach for detection of cyber-aggressive comments by peers on social media network,” *in Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Kochi, Kerala, India.
5. A. Bozyiğit, S. Utku, and E. Nasibov, “Cyberbullying detection: Utilizing social media features,” *Expert Systems with Applications,* vol. 179, no. 10, p. 115001, 2021.
6. A. Çiğdem, E. Çürük, and E. S. Eşsiz, “Automatic detection of cyberbullying in Formspring.me, MySpace and YouTube social networks,” *Turkish Journal of Engineering*, vol. 3, no. 4, pp. 168–178, 2019.
7. M. A. Al-Ajlan and M. Ykhlef, “Deep learning algorithm for cyberbullying detection,” *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 9, pp. 199–205, 2018.
8. T. H. H. Aldhyani, S. N. Alsubari, A. S. Alshebami, H. Alkahtani, and Z. A. T. Ahmed, “Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models,” *International Journal of Environmental Research and Public Health*, vol. 19, no. 19, pp. 1-16, 2022.
9. R. Pawar and R. R. Raje, “Multilingual cyberbullying detection system,” *in Proceedings of the 2019 IEEE International Conference on Electro Information Technology (EIT)*, Brookings, South Dakota, United States of America, 2019.
10. M. E. Alzahrani, T. H. Aldhyani, S. N. Alsubari, M. M. Althobaiti, and A. Fahad, “Developing an intelligent system with deep learning algorithms for sentiment analysis of e-commerce product reviews,” *Computational Intelligence and Neuroscience*, vol. 10, no. 2, pp. 1-10, 2022.
11. H. Çalışkan, Y. Yıldırım, and G. Kılınç, "Examination of the Responsibility and Tolerance of Students Raised in Families with Different Cultural Structures," *TED Eğitim ve Bilim*, vol. 44, no. 199, pp. 353-372, 2019.
12. J. Brailovskaia, T. Teismann, and J. Margraf, “Cyberbullying, positive mental health and suicide ideation/behavior,” *Psychiatry Research*, vol. 267, no. 9, pp. 240–242, 2018.
13. M. A. A. Yani and W. Maharani, “Analyzing cyberbullying negative content on Twitter social media with the RoBERTa method,” *JINAV: Journal of Information and Visualization*, vol. 4, no. 1, pp. 61-69, 2023.