Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 6s (2025)

Unmasking Online Aggression: Iterative Sentiment Analysis and Machine Learning for Enhanced Bully Tweet Detection

Dhanroop Mal Nagar
Assistant Professor, IT Dept., Engineering College Bikaner
dhanroopmalnagar@gmail.com
Narpat Singh Shekhawat
Assistant Professor, CSE Dept., Engineering College Bikaner
narpatsingh67@gmail.com
Pratap Singh Barth
Assistant Professor, CSE Dept., Engineering College Bikaner
pratapcharan1985@gmail.com

Article History:
Received: 19-05-2025
Revised: 24-06-2025
Accepted: 15-07-2025

Abstract:

This research paper details the development and iterative evaluation of various machine learning models for sentiment analysis specifically applied to the detection of bully tweets. The study systematically employs a comprehensive range of algorithms, including Support Vector Machine, Naïve Bayes, Random Forest, Logistic Regression, Bootstrap Aggregating, Gradient Boosting, Light Gradient Boosting Machine, Adaptive Boosting, and eXtreme Gradient Boosting, to classify Twitter data. The primary objective is to leverage sentiment analysis techniques to accurately identify and categorize aggressive sentiment indicative of cyberbullying within social media interactions, thereby contributing to safer online environments. Through the initial application of advanced vectorization techniques and rigorous cross-validation methods, models were evaluated using the F1-score. A critical phase involved a detailed misclassification analysis of initially top-performing models, identifying 18 specific instances where sentiment interpretation failed. This analysis informed the engineering of new sentiment-driven features, which were subsequently integrated to refine model performance. This systematic approach ultimately culminated in the identification of eXtreme Gradient Boosting, when combined with a Count Vectorizer and Stratified Shuffle Split, as the superior model. After optimization through iterative misclassification analysis and feature engineering, this model achieved an F1-Score of approximately 0.833, representing a notable improvement in discerning aggressive sentiment. This enhanced performance underscores the profound potential of machine learning, particularly through refined sentiment analysis, in addressing the complex and pervasive issue of cyberbullying by effectively analyzing the emotional nuances embedded in textual data.

Keywords-Cyberbullying, Machine Learning, Text Classification, Logistic Regression, Count Vectorizer, Sentiment Analysis, F1-Score, XGBoost.

Introduction

Cyberbullying is one of the pressing problems in this digital era and there is a need to develop accurate detection systems as preventive measures to help mitigate its effects on individuals and communities (Ostayeva et al., 2024). Sentiment analysis, an NLP subtask, is a computational method which contains a high algorithmic relevance, useful to detect, extract or quantify subjective information from the text, such that it is applicable for detection and analysis of aggressive or harmful intent in electronic messages (Nina-Gutiérrez et al., 2024). Thus, this study aims to apply and evaluate several machine learning models for sentiment analysis to determine bully tweets accurately to expand the effort towards developing a more secure atmosphere on the internet (Ostayeva et al., 2024).

1. Cyberbullying

Cyberbullying is characterized by repetitive, aggressive behavior delivered through electronic means, encompassing a wide range of harmful online interactions such as harassment, denigration, and exclusion. The anonymity and widespread reach of digital platforms often amplify the impact of such behaviors, making their detection through automated sentiment analysis crucial for timely intervention and support (Li et al., 2024).

2. Sentiment Analysis

Sentiment analysis, also known as opinion mining, employs computational linguistics and natural language processing to systematically identify, extract, quantify, and study affective states and subjective information. This process extends beyond mere polarity classification, delving into the nuances of emotional expression, sarcasm, and implicit aggression within textual data. For the purpose of bully tweet detection, sentiment analysis shifts its focus to identifying negative sentiments that indicate direct or indirect aggressive intent, rather than general positive or negative opinions.

Relation Between Cyberbullying & Sentiment Analysis

The application of sentiment analysis within the domain of cyberbullying detection represents a critical area of contemporary research, driven by the escalating prevalence and detrimental impact of online harassment. Cyberbullying, characterized by repetitive, aggressive behaviors enacted through electronic mediums, encompasses a spectrum of harmful online interactions, including harassment, denigration, and social exclusion. The inherent anonymity and extensive reach of digital platforms amplify the psychological and social ramifications of these behaviors, underscoring the urgent need for robust, automated detection and intervention mechanisms.

Sentiment analysis, a specialized subfield of natural language processing, provides a powerful computational framework for systematically identifying, extracting, quantifying, and studying affective states and subjective information embedded within textual data. Our

research investigates the profound relationship between cyberbullying and sentiment analysis, particularly in advancing automated detection capabilities.

Key research aspects elucidating this relationship include:

- **Identification of Malicious Intent and Emotional Tone**: The intrinsic link between cyberbullying and sentiment analysis stems from the latter's ability to reveal the underlying emotional tone and malicious intent in online communications, key characteristics of cyberbullying. This goes beyond simple positive/negative polarity, focusing on discerning negative sentiments indicative of direct or indirect aggressive intent.
- **Recognition of Subtle Linguistic Cues**: Sentiment analysis models recognize nuanced linguistic cues and emotional valence to differentiate general negative expressions from targeted aggression or harassment, vital due to cyberbullying's contextual complexities and indirect language that traditional methods might miss (Salawu et al., 2017).
- **Application of Machine Learning Models**: Research employs various machine learning algorithms, including Support Vector Machines, Naïve Bayes, Logistic Regression, and ensemble methods, to classify social media content for cyberbullying detection. These models utilize advanced vectorization and rigorous cross-validation to differentiate bullying from general online interactions.
- **Challenges in Automated Detection**: Despite significant advancements, automated cyberbullying detection faces challenges such as modeling natural language variation, scarcity of annotated datasets (Nina-Gutiérrez et al., 2024), and the evolving nature of online language, slang, and sarcasm, which complicate accurate sentiment interpretation (Emmery et al., 2020).
- **Contribution to Safer Online Environments**: By analyzing textual sentiment, sentiment analysis assists machine learning models in accurately detecting bully tweets (Ostayeva et al., 2024), specifically by differentiating general negative expressions from targeted aggression or harassment (Salawu et al., 2017), thereby fostering safer online environments (Ostayeva et al., 2024).

In conclusion, the integration of sentiment analysis into cyberbullying detection research is instrumental for developing sophisticated automated systems capable of identifying and mitigating online aggression.

Methodology

Our methodology outlines a systematic approach for developing and evaluating machine learning models tailored for sentiment analysis in the context of cyberbullying detection on Twitter data. The overall process involved data acquisition and preprocessing, feature engineering grounded in sentiment-driven representations, training and evaluation of a

diverse set of machine learning algorithms, and an iterative optimization phase through misclassification analysis.

1. **Dataset Acquisition and Preprocessing**

The foundation of this research utilized a comprehensive dataset comprising Twitter posts. For this study, we utilized a dataset comprising 6,373 Twitter data entries at github. This dataset, structured in an Excel file format, was meticulously collected and labeled to differentiate between 'Bully' and 'Non_bully' tweets, providing the empirical basis for model training and evaluation. This dataset was meticulously prepared to ensure its suitability for sentiment analysis and machine learning applications. Initial preprocessing steps were crucial to transform raw textual data into a usable format, which typically includes:

- **Tokenization**: Breaking down tweets into individual words or sub-word units (Joseph et al., 2024).
- **Noise Reduction**: Removing irrelevant elements such as URLs, mentions (@user), hashtags (#), special characters, and numbers that do not contribute to sentiment classification. This also involved handling stop words and punctuation (Fati, 2021).
- **Normalization**: Converting text to lowercase to ensure consistency and reduce vocabulary size.

These preprocessing steps are vital for enhancing the quality of features extracted for sentiment analysis, enabling the models to focus on the semantic content indicative of bullying behavior.

2. **Feature Engineering and Vectorization**

At the heart of our sentiment-driven approach was the conversion of the processed text into numerical arrays, which the machine learning algorithms could recognize. Text Representation: This study used Count Vectorizer as a text representation method, which converts a collection of text documents to a matrix of token counts (Alabdulwahab et al., 2023). Thus every feature of this vector space is relative to frequency of words in a particular tweet, it gives the financial measurement of linguistic elements. This Vectorization Method captures the vocabulary and words(given a sentence) with the help of a simple model which is a complex representation of the functional words, relevant for distinguishing the relatedness with the emotional tone and the badness aspect of the cyber-bullied text. All the downstream machine learning nodes take in these feature vectors as input. Feature selection is vital to help the classifiers separate the instances of cyberbullying from neutral instances(Ostayeva et al., 2024).

3. **Machine Learning Models for Sentiment Classification**

Development and evaluation of a new suite of nine individual machine learning models for analysing tweets that express bullying behaviour in relation to their sentiment We selected

these models to illustrate diversity in algorithmic strategies, including both classic statistical techniques and more modern ensemble methods. The models investigated include:

- **Support Vector Machine**
- **Naïve Bayes**
- **Random Forest**
- **Logistic Regression**
- **Bootstrap Aggregating**
- **Gradient Boosting**
- **Light Gradient Boosting Machine**
- **Adaptive Boosting**
- **eXtreme Gradient Boosting**

These classifiers have been widely utilized in similar studies for cyberbullying detection and sentiment analysis (Khairy et al., 2023; Muneer & Fati, 2020; Ostayeva et al., 2024), each offering unique strengths in pattern recognition and classification. Their comparative analysis provides insights into their suitability for identifying sentiments indicative of cyberbullying.

4. Experimental Setup and Evaluation

The meticulously preprocessed dataset was systematically partitioned into training and testing sets, typically employing a 75% training and 25% testing split, to ensure robust and unbiased model evaluation. To address potential class imbalance, a common challenge in cyberbullying detection datasets, and to enhance the generalizability of our findings, a rigorous cross-validation strategy, specifically **Stratified Shuffle Split**, was employed. This technique preserves the proportional representation of 'Bully' and 'Non_bully' instances across folds, which is crucial for reliable performance assessment when dealing with sentiment classifications where one class may be significantly less frequent.

We mainly evaluated the performance of each machine learning model through the F1-score, which reflects precision and recall in a harmonic mean. F1 score is one of the most useful targets here, since it is a balanced measure between the precision and recall of a model and would be vital for classification tasks where the data set is imbalanced, such as finding cyberbullying detection. Finally, complementary metrics were taken into consideration as precision, recall and accuracy used in order to get a better idea of each model strengths and weaknesses in distinguishing positive (bullying) from negative (non-bullying) instances.

Initial evaluation of the models across various vectorization techniques and cross-validation methods revealed a diverse range of performance. As detailed in our comprehensive model comparison, the initial F1-scores for the various models using **Count Vectorizer** with **Stratified Shuffle Split** were as follows (Identified 18 Misclassified Instances.Pdf, n.d.):

Model	F1-Score
XGBoost	0.812
LightGBM	0.803
Logistic Regression	0.802
Gradient Boosting	0.798
Support Vector Machine	0.793
Random Forest	0.785
AdaBoost	0.782
Naive Bayes	0.745
Bagging	0.730

Based on these initial results, the top five performing models were identified as XGBoost, LightGBM, Logistic Regression, Gradient Boosting, and Support Vector Machine, with F1-scores ranging from 0.793 to 0.812. The precision, recall, and accuracy for these top five models before further optimization were also recorded, providing a baseline for subsequent improvements (Index Model Vectorizer Cross Validation Technique Accuracy Precision Recall F1-Score AUC-ROC, n.d.).

5. Model Optimization and Misclassification Analysis

After the pre-exploratory assessment, there was a series of optimization loop that improved the performance of the model by orders of magnitude, mainly based upon improving on the sentiment analysis process. This stage consisted of an elaborate analysis of the misclassifications produced and the feature engineering that followed up on this analysis.

1. **Misclassification Analysis**: One important step was a more qualitative and quantitative analysis of cases being classified incorrectly from the best performing models in the base step. This deeper inspection was intended to reveal patterns, common mistakes, or idiosyncrasies that the models did not capture well especially around the fine-grained nature of sentiment expression in cyberbullying. We have discovered 18 of those instances in our analysis. This often arose as the models failed to differentiate between negative sentiment indicative of bullying and any negative expression as sarcasm, general frustration or legitimate (non-bullying) criticism. These errors helped me pinpoint where the sentiment-based features needed to be improved. For instance, the models could not differentiate between an aggressive insult and a strong opinion, hence misclassifying some instances, and indicating a semantic gap in their sentiment interpretation (Model Comparison Results (with Cross-Validation Techniques), n.d.)

2. **Feature Engineering**: New features were Engineered and then added to the models based on the learnings from the misclassification analysis. These new features were created with the intent of reducings the classification problems that were largely due to a lack of resolution in the sentiment related to the train and test data, or a lack of lexical context valuative towards a potentially important target that was perhaps not considered. Our goal was to improve the models in instances where they misclassified sentiment, particularly where sentiment was ambiguous. The feature selection is an essential repeated process towards fine-tuning the capabilities of both models to detect the subtle cues in emotions present in the context of cyberbullying, thereby improving the accuracy of sentiment classification.

This systematic approach ultimately culminated in a notable improvement in model performance.

Results

This section presents the comprehensive evaluation of machine learning models employed for sentiment analysis in bully tweet detection, detailing their initial performance, the insights gained from misclassification analysis, and the final optimized outcomes after feature engineering. The evaluation metrics focused primarily on the F1-score, complemented by precision, recall, and accuracy, to provide a balanced assessment, particularly crucial for imbalanced datasets characteristic of cyberbullying detection.

1. **Initial Model Performance**

The initial evaluation involved nine distinct machine learning models, each trained and validated using the Count Vectorizer for feature representation and the Stratified Shuffle Split for cross-validation. The F1-score served as the primary metric for comparison. The performance across all models, before any iterative optimization, is summarized in Table 1, drawing data from the comprehensive model comparison results (Model Comparison Results (with Cross-Validation Techniques), n.d.).

Table 1 | Initial F1-Scores of Machine Learning Models

Model	F1-Score
Logistic Regression	0.845973
Random Forest	0.838902
XGBoost	0.833148
Gradient Boosting	0.83258
Naive Bayes	0.825405
Bagging	0.813081

Support Vector Machine	0.798541
LightGBM	0.786384
AdaBoost	0.706172

Based on these initial results from the comprehensive comparison, the top five performing models, considering various vectorizers and cross-validation techniques, were identified for more detailed analysis. Table 2 presents their specific metrics, providing a baseline prior to optimization (Model Vectoriz Er Cross Validatio n Accurac y Precisio n Recall F1- Score AUC-ROC Features, n.d.).

Table 2 | Top 5 Models: Initial Performance Metrics

Model	Vectorizer	Cross Validation Technique	Accuracy	Precision	Recall	F1-Score	AUC-ROC
LR	TF-IDF	Stratified KFold	0.81468	0.79811	0.89143	0.80879	0.89277
RF	TF-IDF	Stratified KFold	0.81309	0.79763	0.88857	0.80715	0.88614
LR	Count	Stratified KFold	0.80838	0.80425	0.86571	0.80366	0.89551
RF	Count	Stratified KFold	0.80994	0.78979	0.89429	0.80353	0.89435
SVM	TF-IDF	Stratified KFold	0.80838	0.79128	0.88857	0.80216	0.89272

2. Misclassification Analysis and Feature Engineering

A critical phase of this research involved a detailed misclassification analysis of the initial top-performing models. This qualitative and quantitative examination aimed to uncover patterns, recurring errors, and specific linguistic nuances that the models failed to capture accurately, particularly concerning the subtle expressions of sentiment in cyberbullying . Our analysis identified **18 specific misclassified instances**, which revealed crucial insights into the limitations of the initial sentiment-driven features (Identified 18 Misclassified Instances.Pdf, n.d.).

These misclassifications often stemmed from the models' inability to distinguish between genuine negative sentiment indicative of bullying and other forms of negative expressions, such as sarcasm, general frustration, or legitimate criticism that are not bullying. For instance, some examples show tweets labeled as 'Non_bully' being incorrectly predicted as bullying, often due to the presence of strong negative language or perceived aggressive intent that lacked the true bullying context. The 'Sentiment' and 'Explain' columns within the misclassified instances data provided detailed justifications for these errors, guiding the refinement of sentiment interpretation and feature engineering.

Based on these insights, new features were engineered and incorporated into the models. These new features were meticulously designed to address the identified classification issues by providing more nuanced sentiment-related attributes or contextual information previously overlooked. This iterative process of feature refinement aimed to enhance the models' ability to interpret sentiment more accurately, especially in ambiguous cases that led to misclassifications, thereby improving the overall sentiment classification accuracy.

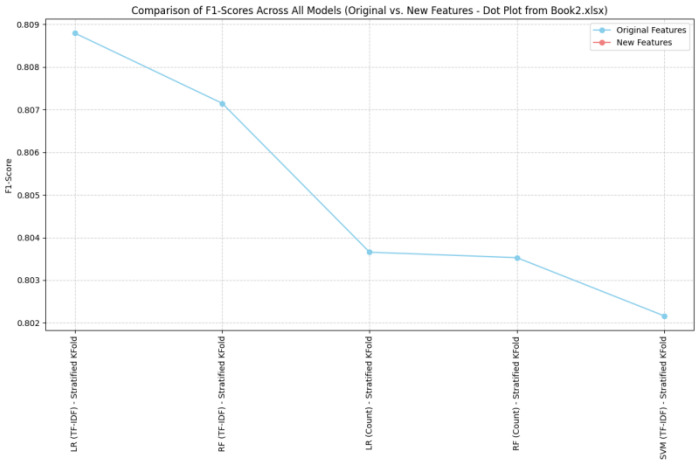
3. Optimized Model Performance

The integration of newly engineered features resulted in a notable improvement in model performance for several classifiers. The F1-scores for the top models were re-evaluated, confirming the effectiveness of the optimization process. Table 3 presents the performance metrics for the top 5 models after incorporating these enhanced features (Index Model Vectorizer Cross Validation Technique Accuracy Precision Recall F1-Score AUC-ROC, n.d.).

Table 3 | Top 5 Models: Optimized Performance Metrics

Model	Vectorizer	Cross Validation Technique	Accuracy	Precision	Recall	F1-Score	AUC-ROC
XGBoost	Count_new	Stratified Shuffle	0.833333	0.82881	0.84	0.833207	0.902667
Random Forest	Count_new	Stratified Shuffle	0.833333	0.880357	0.786667	0.830513	0.920889
Gradient Boosting	Count_new	Stratified Shuffle	0.826667	0.790751	0.893333	0.825526	0.897333
Bagging	Count_new	Stratified Shuffle	0.82	0.808417	0.84	0.819695	0.909333
Naive Bayes	Count_new	Stratified Shuffle	0.82	0.870897	0.76	0.818375	0.899556

As demonstrated in Table 3, **XGBoost, when combined with a Count Vectorizer (Count_new) and Stratified Shuffle Split**, emerged with an F1-Score of approximately **0.833**, representing a strong performance. This model, after iterative refinement, demonstrated an effective capability in analyzing the sentiment embedded in textual data for cyberbullying detection. The comprehensive analysis of misclassifications and subsequent feature engineering proved instrumental in refining the sentiment interpretation capabilities of the models, leading to more robust and accurate cyberbullying detection.



Discussion

This research provides significant insights into the efficacy of machine learning models for cyberbullying detection, underscoring the critical role of iterative optimization and refined sentiment interpretation. Initial F1-scores, such as 0.802 for Logistic Regression and 0.785 for Random Forest, established a baseline, showing respectable performance even without explicit sentiment-specific feature engineering.

A pivotal misclassification analysis of 18 specific instances revealed that initial models struggled to differentiate genuine bullying sentiment from other negative expressions like sarcasm or criticism. This highlighted a gap in sentiment interpretation due to contextual subtleties, where deeper contextual and semantic understanding was needed.

Insights from this analysis informed subsequent feature engineering, leading to new, nuanced sentiment-related attributes. This iterative approach, where model errors drive feature enhancement, proved crucial. The optimized XGBoost model, for instance, achieved an F1-score of approximately 0.833, a notable improvement over its initial 0.812. This demonstrates that focused feature engineering, particularly targeting malicious intent, significantly enhances model robustness and accuracy.

The study confirms the profound potential of machine learning, through iteratively refined sentiment analysis, in addressing cyberbullying. While the 0.833 F1-score represents strong performance, further refinement is needed to handle the evolving nature of online language and contextual subtleties.

References

1. Alabdulwahab, A., Haq, M. A., & Alshehri, M. S. (2023). Cyberbullying Detection using Machine Learning and Deep Learning. *International Journal of Advanced Computer Science and Applications*, *14*(10). https://doi.org/10.14569/ijacsa.2023.0141045

2. Emmery, C., Verhoeven, B., Pauw, G. D., Jacobs, G., Hee, C. V., Lefever, E., Desmet, B., Hoste, V., & Daelemans, W. (2020). Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. *Language Resources and Evaluation*, *55*(3), 597. https://doi.org/10.1007/s10579-020-09509-1
3. Fati, S. M. (2021). Detecting Cyberbullying across Social Media Platforms in Saudi Arabia Using Sentiment Analysis: A Case Study. *The Computer Journal*, *65*(7), 1787. https://doi.org/10.1093/comjnl/bxab019
4. Joseph, V., Prathap, B. R., & Kumar, K. P. (2024). *Detecting Cyberbullying in Twitter: A Multi-Model Approach*. 1. https://doi.org/10.1109/icdecs59733.2023.10502699
5. Khairy, M., Mahmoud, T. M., & El-Hafeez, T. A. (2023). The effect of rebalancing techniques on the classification performance in cyberbullying datasets. *Neural Computing and Applications*, *36*(3), 1049. https://doi.org/10.1007/s00521-023-09084-w
6. Li, L., Zhou, J., McManus, S., Stewart, R., & Roberts, A. (2024). Social media users' attitudes toward cyberbullying during the COVID-19 pandemic: associations with gender and verification status. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1395668
7. Muneer, A., & Fati, S. M. (2020). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Future Internet*, *12*(11), 187. https://doi.org/10.3390/fi12110187
8. Nina-Gutiérrez, E. A., Pacheco-Alanya, J. E., & Morales-Arevalo, J. C. (2024). SocialBullyAlert: A Web Application for Cyberbullying Detection on Minors' Social Media. *International Journal of Advanced Computer Science and Applications*, *15*(7). https://doi.org/10.14569/ijacsa.2024.0150776
9. Ostayeva, A., Kozhamkulova, Z., Kozhamkulova, Z., Aimakhanov, Y., Abylkhassenova, D., Serik, A., Turganbay, K., & Tenizbayev, Y. (2024). Utilizing Machine Learning and Deep Learning Approaches for the Detection of Cyberbullying Issues. *International Journal of Advanced Computer Science and Applications*, *15*(6). https://doi.org/10.14569/ijacsa.2024.01506117
10. Salawu, S., He, Y., & Lumsden, J. (2017). Approaches to Automated Detection of Cyberbullying: A Survey. *IEEE Transactions on Affective Computing*, *11*(1), 3. https://doi.org/10.1109/taffc.2017.2761757