 ISSN 2579-7069

A Text Classification Approach for Detecting Cyberbullying Risk on Twitter Using Support Vector Machine with Naive Bayes and Random Forest Comparison

Sri Yarsasi[1,*], Angga Iskoko[2]

[1,2]*Magister of Computer Science, Amikom Purwokerto University*

(Received March 11, 2025; Revised July 9, 2025; Accepted October 1, 2025; Available online December 27, 2025)

Abstract

The rapid development of social media as a means of digital interaction also presents serious challenges in the form of the spread of negative content, including cyberbullying. Cyberbullying is a form of verbal violence committed online and has a significant impact on mental health, especially in adolescents. This research aims to develop a text classification model to detect the risk of cyberbullying using the Support Vector Machine (SVM) algorithm. The data used comes from a collection of cyberbullying-themed tweets. The research stages include text preprocessing (normalization, cleaning, tokenization, stopword removal, and stemming), feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF), data division into training and testing sets, and model training using linear kernel of SVM. The model was evaluated using accuracy, precision, recall, and F1-score metrics. The results show that this approach is able to identify risky comments quite accurately, with optimal performance on the linear kernel. This research contributes to the development of automated detection systems to create a safer and healthier digital ecosystem, and supports preventive efforts in mitigating cyberbullying online.

Keywords: Cyberbullying Detection, Twitter, Support Vector Machine, TF-IDF, Text Classification, Machine Learning

1. Introduction

The development of information and communication technology has significantly transformed human interaction patterns, especially within digital environments [1]. Social media has emerged as a dominant platform for expressing opinions, disseminating information, and establishing interpersonal relationships in virtual spaces [2]. Despite these advantages, social media is widely recognized as an enabler for the propagation of harmful behaviors, including hate speech and cyberbullying [3], which continue to escalate alongside increased global connectivity.

Cyberbullying refers to a form of psychological aggression delivered through digital communication channels, often anonymously and repeatedly, and frequently targeting vulnerable user groups such as adolescents [4], [5]. Numerous studies have shown that cyberbullying can cause emotional distress, reduced self-esteem, social withdrawal, and in severe cases suicidal tendencies [6], [7]. Accordingly, the rapid identification of harmful online content has become a critical focus for researchers and digital platform providers. Automated detection systems based on machine learning have been increasingly adopted to support early identification and prevention efforts [8], [9].

This research applies data mining and machine learning methodologies specifically the Support Vector Machine (SVM) algorithm for classifying text with potential cyberbullying risk. The preprocessing pipeline employs normalization, tokenization, stopword removal, and stemming to reduce linguistic noise and to enhance computational efficiency [10]. Feature extraction uses the Term Frequency–Inverse Document Frequency (TF-IDF) model, which has proven effective in capturing key lexical indicators within text-based classification tasks [11]. The primary objectives of this research are as follows:

***Corresponding author: Sri Yarsasi (24MA41D007@students.amikompurwokerto.ac.id)**
DOI: https://doi.org/10.47738/ijiis.v8i4.290

a) To develop an automated text classification model capable of detecting cyberbullying indicators in user-generated social media content by recognizing abusive and harmful linguistic patterns.

b) To implement a comprehensive preprocessing and TF-IDF vectorization workflow for converting unstructured text into meaningful numerical features suited for machine learning pipelines.

c) To evaluate the capability of the SVM algorithm to identify explicit and implicit indicators of abusive discourse and to assess its robustness in handling short, informal online texts.

d) To measure the performance of the proposed model using standard classification metrics including accuracy, precision, recall, and F1-score, ensuring objective quantification of model reliability.

e) To contribute to broader cyberbullying mitigation strategies by providing a scalable computational approach that may support automated monitoring or real-time content filtering efforts across digital platforms.

This research is expected to deliver contributions from both practical and academic perspectives. Practically, the model developed in this study may assist social media administrators, educational institutions, and child protection agencies in early detection and mitigation of harmful digital interactions [11]. Automated systems capable of identifying toxic content can significantly support digital safety initiatives and reduce the psychological risks faced by adolescents and other vulnerable groups.

From an academic viewpoint, this research enriches the field of Natural Language Processing (NLP) and artificial intelligence by demonstrating the integrated use of TF-IDF and SVM in addressing socially relevant text classification problems [12]. The findings contribute additional empirical evidence on the effectiveness of classical machine learning techniques for detecting nuanced and context-dependent online aggression. Furthermore, this work provides a foundation for the development of systems capable of interpreting online hostility with increased sensitivity to linguistic and emotional dimensions, thereby encouraging more responsible digital citizenship [13], [14].

2. Literature Review

Cyberbullying detection using machine learning has grown substantially in response to heightened concerns regarding digital safety and online harassment. As cyberbullying frequently manifests through informal language, sarcasm, implicit aggression, and coded expressions, manual inspection becomes impractical at scale, and automated systems have become necessary [1], [3].

Early studies such as Dinakar et al. [2] applied machine learning techniques including Naive Bayes and SVM to classify hate speech on Twitter, demonstrating that SVM yields superior performance on sparse and high-dimensional textual features. Abusaqer and Saquer [4] explored bullying categories such as race, religion, and sexual orientation on YouTube comments, highlighting the importance of granular data labeling to boost model accuracy. Yorozu et al. [6] improved TF-IDF-based detection by incorporating user-level behavioral attributes, showing that contextualized features enhance performance in identifying harmful comments.

TF-IDF remains one of the most popular feature extraction methods in cyberbullying and hate speech detection due to its ability to highlight important lexical patterns while filtering irrelevant noise [7], [11]. On the algorithmic side, SVM has consistently demonstrated strong performance for text classification tasks due to its robustness against high-dimensional vectors and its ability to delineate complex decision boundaries [8], [12]. Rao et al. [10] compared SVM with Decision Tree and k-NN models, revealing SVM's superior precision, particularly for microtext formats such as tweets.

Text preprocessing also plays a vital role in classification outcomes. Mishra et al. [13] emphasized that normalization, stopword removal, and stemming significantly reduce lexical variability and enhance model interpretability. Without effective preprocessing, the model risks being misled by noise and irrelevant linguistic elements. Despite the advancements, challenges remain especially in detecting sarcasm, which often obscures explicit bullying cues, and in resolving class imbalance where non-bullying comments dominate datasets [14], [15].

To address these limitations, this study extends prior work by implementing an SVM-based classification model using real-world social media datasets, supported by a rigorous evaluation using accuracy, precision, recall, and F1-score.

Through this approach, the research aims to produce stronger generalization and enhanced detection capability for cyberbullying risk identification [16], [17], [18], [19], [20].

3. Methodology

The methodology adopted in this study comprises a series of well-defined and systematic stages designed to accurately classify cyberbullying risk based on textual content derived from social media, specifically Twitter. Each phase in the research pipeline from data acquisition and preprocessing to model training and evaluation was carefully structured to ensure the reliability and validity of the classification outcomes. An overview of the methodological workflow employed in this study is presented in Figure 1.

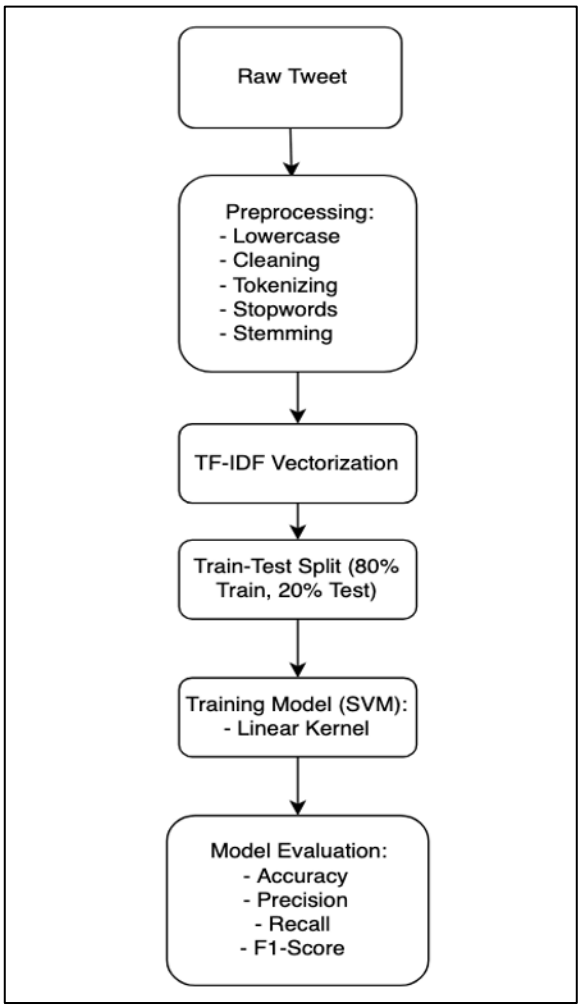

Figure 1. Flowchart of the Classification Process

3.1. Dataset

The dataset utilized in this study is the publicly available Cyberbullying Tweets Dataset, which contains a substantial collection of annotated textual data (tweets) specifically curated for the task of cyberbullying detection. Each instance in the dataset comprises a tweet's textual content along with a corresponding class label that indicates the presence or absence of cyberbullying behavior. The labels are structured to reflect various categories of online abuse, allowing for multi-class classification. This dataset was selected due to its relevance to real-world social media contexts and its widespread usage in prior research, thereby serving as a robust foundation for both model training and evaluation. The tweets are categorized into six distinct classes based on the nature of the bullying, as summarized in Figure 2. These categories include gender-based bullying, religion-based bullying, age-based bullying, ethnicity-based bullying, other forms of cyberbullying, and not_cyberbullying (i.e., non-abusive content).

 ISSN 2579-7069

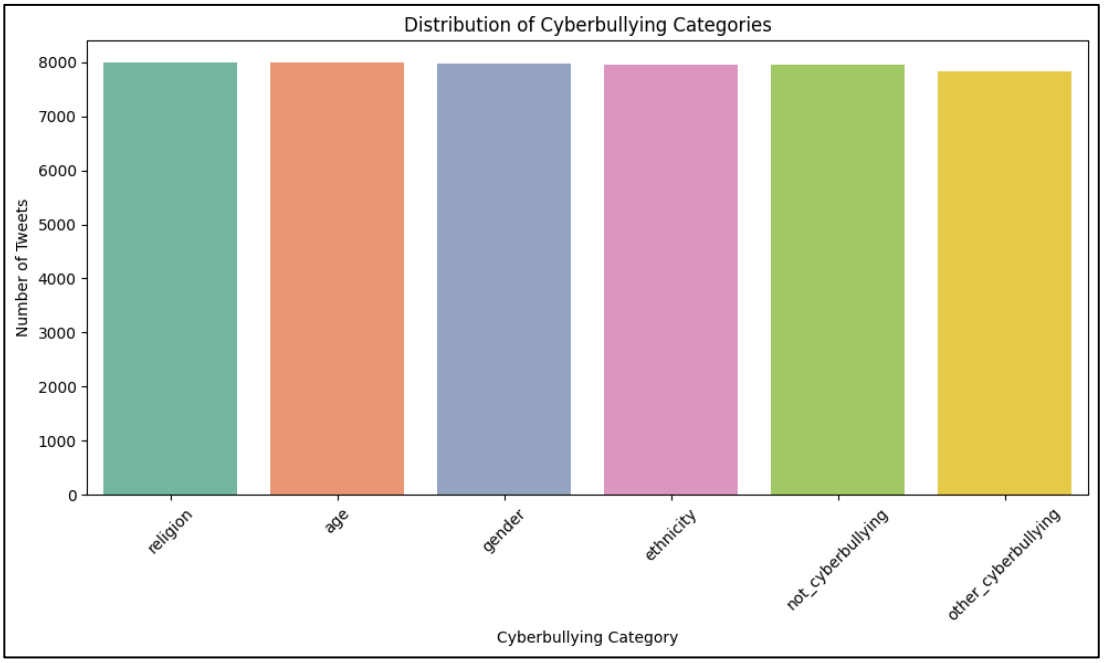

Figure 2. Dataset Visualization

The dataset comprises approximately 30,000 Twitter entries stored in Comma-Separated Values (CSV) format. Each record within the dataset contains the full textual content of a tweet along with a corresponding label that classifies the type of cyberbullying behavior represented. These labels facilitate supervised learning by serving as ground truth annotations for model training and evaluation purposes.

Table 1. Distribution of Tweets by Cyberbullying Category

Category	*Number of Tweets*
Age	1.586
Ethnicity	1.603
Gender	1.531
Not Cyberbullying	1.624
Other Cyberbullying	1.612
Religion	1.603

Table 1 illustrates that the dataset exhibits a relatively balanced distribution across the various cyberbullying categories, with a slight predominance observed in the 'not_cyberbullying' class. Recognizing this distributional characteristic is crucial, as class imbalance can significantly affect the performance of classification models. In particular, models trained on imbalanced datasets tend to be biased toward majority classes, potentially leading to underperformance in detecting minority class instances. To mitigate this issue, appropriate techniques such as class weight adjustment, oversampling of underrepresented classes, or synthetic data generation (e.g., SMOTE) may be employed to enhance model generalization and fairness across all categories.

3.2. Data Preprocessing

This stage is intended to systematically clean and prepare raw textual data for effective processing by machine learning algorithms. The preprocessing phase plays a crucial role in minimizing noise and standardizing input data, thereby enhancing the quality and consistency of feature extraction. The following sequential steps were implemented during the text preprocessing pipeline, as detailed in Table 2.

 ISSN 2579-7069

Table 2. Text Preprocessing Steps

Step	*Description*
Cleaning	All special characters, numeric digits, punctuation marks, URLs, and extraneous symbols were removed to reduce textual noise.
Lowercasing	All text was converted to lowercase to eliminate case sensitivity, ensuring uniform treatment of words.
Tokenization	Sentences were segmented into individual tokens or words to facilitate lexical-level analysis.
Stopword Removal	Common words that carry little to no discriminative meaning in the context of classification (e.g., "and," "the," "from") were excluded.
Stemming	Words were reduced to their root or base form (e.g., "playing" → "play") to unify semantic variations of the same word.

An illustrative example of this process is as follows:

Original tweet: "This is STUPID!!! #hate" || After preprocessing: "stupid hate"

These preprocessing steps were implemented using the Natural Language Toolkit (NLTK) and the regular expression (re) module in Python, which provided flexible and efficient tools for text normalization and linguistic transformation.

3.3. Equations

Following the text preprocessing stage, the cleaned textual data was transformed into a numerical representation using the TF-IDF technique. This method is widely adopted in natural language processing and text mining tasks due to its effectiveness in capturing the relative importance of words within a document while accounting for their frequency across the entire corpus.

The TF-IDF scheme operates based on two key components:

a) Term Frequency (TF): Measures how frequently a term t appears within a specific document d. It reflects the local importance of a word in that particular document and is often computed as the raw count or normalized frequency.
b) Inverse Document Frequency (IDF): Evaluates how unique or rare a term is across all documents in the corpus. Words that occur in many documents (e.g., common words) are assigned lower weights, while those appearing in fewer documents receive higher importance scores.

The TF-IDF score for a term t in document d is calculated as:

$$\text{TF-IDF}(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right) \quad (1)$$

where:
N : denotes the total number of documents in the corpus
$DF(t)$: represents the number of documents in which the term t appears
This transformation results in a sparse, high-dimensional vector representation of the corpus, where each document is encoded as a vector of TF-IDF weights corresponding to the vocabulary terms. These vectors are then used as input features for the classification model.

3.4. Classification Models

The classification model employed in this study is the SVM, a well-established supervised learning algorithm known for its effectiveness in handling high-dimensional and sparse data such as textual information. SVM operates by identifying the optimal hyperplane that maximally separates data points from different classes. In the context of this research, the objective of the SVM is to distinguish between various categories of cyberbullying based on linguistic patterns derived from Twitter posts.

The training and evaluation process of the model involved the following steps:

a) Data Partitioning: The dataset was partitioned into two subsets using an 80:20 ratio, where 80% of the data was allocated for model training and the remaining 20% for testing.
b) Model Training: The SVM classifier was trained using the training set, which enabled the model to learn the underlying patterns and associations between text features (TF-IDF vectors) and the corresponding cyberbullying categories.
c) Model Testing: The trained model was subsequently evaluated using the unseen test set to assess its generalization capability and overall classification performance.

For illustrative purposes, assuming a dataset containing 30,000 samples, approximately 24,000 instances were used for training, while the remaining 6,000 samples served as the testing set.

This standard train-test split technique ensures that the model's performance is evaluated on data it has not encountered during training, thereby providing a more realistic measure of its predictive accuracy and robustness.

3.5. Classification Models

The classification model was constructed using the SVM algorithm, a widely used supervised learning method known for its robust performance in handling high-dimensional data. SVM operates by identifying an optimal hyperplane that best separates data points belonging to different classes within the feature space. This optimal separation is achieved by maximizing the margin between support vectors, which are the closest data points to the decision boundary from each class.

SVM supports various kernel functions that enable it to model both linear and non-linear decision boundaries depending on the nature of the data. The kernel function implicitly maps the input features into a higher-dimensional space without explicitly computing the transformation, thus allowing the algorithm to find a suitable hyperplane even in complex scenarios. Commonly used kernels include the linear kernel, polynomial kernel, radial basis function (RBF), and sigmoid kernel.

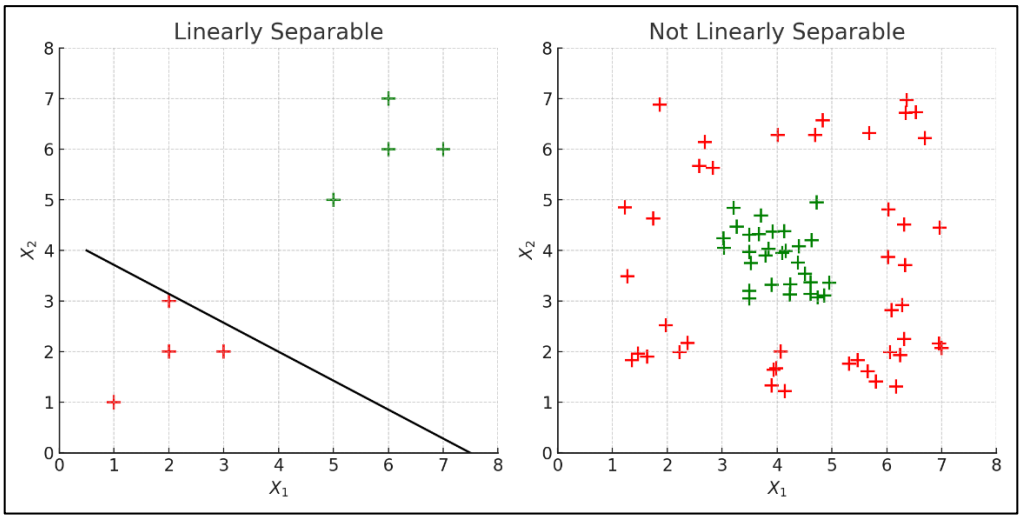

Figure 3. Illustration of Linearly Separable and Non-Linearly Separable Data

A visual illustration of the difference between linear and non-linear SVM kernels is presented in Figure 3, which demonstrates how the choice of kernel affects the shape of the decision boundary and the model's ability to classify data accurately in non-linearly separable contexts.

a) Kernel Concept in SVM
SVM employs a kernel function to transform input data into a higher-dimensional feature space, allowing for the separation of data that is not linearly separable in its original space. This transformation enables SVM to construct an optimal hyperplane that can distinguish between different classes, even when the decision boundary in the original space is complex or non-linear.

In general, kernel functions serve as mathematical tools that implicitly compute the inner product of data points in a transformed feature space without explicitly performing the transformation a technique known as the "kernel trick." This approach significantly reduces computational complexity while enabling SVM to model complex decision boundaries effectively.

There are various types of kernel functions commonly used in SVM implementations. Among them, the two most fundamental types are:

1. Linear Kernel:
   This kernel is appropriate when the data is linearly separable, meaning that a straight line (in two dimensions) or a hyperplane (in higher dimensions) can be used to classify the data points. The linear kernel is defined by the inner product of two input vectors and is expressed by the following equation:
   $$K(x, y) = x \cdot y \tag{2}$$
   The linear kernel is particularly well-suited for datasets that are linearly separable or approximately so, especially when dealing with high-dimensional feature spaces. This condition commonly arises in text mining and natural language processing tasks, where feature representations such as TF-IDF can result in thousands of sparse dimensions. Due to its simplicity, the linear kernel offers fast computation and low memory usage, making it highly efficient for large-scale classification problems. Moreover, its deterministic nature reduces the risk of overfitting, which is often encountered when working with complex or non-linear kernels on high-dimensional data.
2. Non-Linear Kernel
   Non-linear kernels are employed when the data is not linearly separable in its original feature space. These kernels transform the data into a higher-dimensional space where a linear hyperplane can be used to separate the classes. By applying such transformations, non-linear kernels enable the SVM to model complex and non-linear decision boundaries that would otherwise be impossible to represent in the original space.
   Two commonly used non-linear kernels include:
   - Polynomial Kernel:
     $$K(x, y) = (x \cdot y + c)^d \tag{3}$$
   - Radial Basis Function (RBF) or Gaussian Kernel:
     $$K(x, y) = \exp\left(-\gamma \|x - y\|^2\right) \tag{4}$$

   Non-linear kernels offer significant flexibility and are capable of capturing intricate relationships between input features. However, this increased modeling capacity comes at the cost of higher computational complexity and a greater risk of overfitting, particularly when applied to large and sparse datasets. Therefore, careful tuning of hyperparameters and validation techniques is required to ensure generalization and avoid excessive fitting to the training data.

b) Linear Kernel Selection

In this study, the linear kernel was selected as the preferred configuration for the SVM classifier based on several critical considerations. First, the use of TF-IDF for feature representation inherently produces high-dimensional and sparse vectors, often consisting of thousands of features. In such cases, textual data tends to exhibit linear separability due to the sparsity and structure of word distributions across documents.

Second, the linear kernel offers significant computational efficiency compared to non-linear alternatives. Its simplicity enables faster training and prediction times, which is particularly advantageous when dealing with large-scale datasets or real-time applications. Moreover, the linear kernel involves fewer hyperparameters, reducing the complexity of model tuning and mitigating the risk of overfitting.

Third, from a generalization standpoint, the linear kernel tends to be more robust on high-dimensional data, as it avoids mapping input features into excessively complex spaces that could lead to poor model generalization on unseen data.

 ISSN 2579-7069

By employing the linear kernel, the SVM constructs an optimal separating hyperplane directly within the TF-IDF feature space. This approach has proven sufficient for distinguishing between "cyberbullying" and "non-cyberbullying" categories, as it effectively leverages the discriminative power of the high-dimensional text features while maintaining computational tractability and model stability.

3.6. Model Evaluation

The classification model was rigorously evaluated using several standard performance metrics derived from the confusion matrix, which provides a detailed summary of prediction results by comparing actual labels with predicted outcomes. The following metrics were employed, as presented in Table 3.

Table 3. Evaluation Metrics for Classification Performance

Metric	*Description*
Accuracy	Measures the overall correctness of the model by calculating the proportion of true results (both true positives and true negatives) among the total number of predictions.
Precision	Indicates the proportion of correctly predicted positive instances (e.g., cyberbullying) relative to all instances predicted as positive. High precision reflects the model's ability to minimize false positives.
Recall (Sensitivity)	Measures the model's effectiveness in identifying all actual positive cases by quantifying the proportion of true positives detected out of all real cyberbullying instances.
F1-Score	Represents the harmonic mean of precision and recall, offering a balanced assessment of a model's performance when both false positives and false negatives are of concern.

3.7. Research Limitations

Despite demonstrating promising results, this study is subject to several limitations that must be acknowledged. First, the dataset used in this research was obtained solely from the Cyberbullying Tweets Dataset available on Kaggle. While comprehensive, this dataset may not fully represent the linguistic diversity, cultural variation, or behavioral nuances observed across different social media platforms and demographic groups. As a result, the generalizability of the findings may be constrained.

Second, the model relies exclusively on text-based features generated using the TF-IDF representation, without incorporating additional contextual or metadata features such as timestamp information, user interactions, emotional tone, or network-level attributes. Such contextual information could provide deeper insight into the intent and impact of cyberbullying expressions and potentially enhance detection performance.

Third, although the SVM algorithm was selected for its robustness and ability to handle high-dimensional data effectively, the study does not include direct comparative analysis with more advanced classification algorithms such as Random Forests or deep learning models (e.g., LSTM, BERT). Including such comparisons could yield valuable insights into the relative strengths and weaknesses of each approach in the domain of cyberbullying detection.

Finally, the dataset used exhibits class imbalance, with certain categories such as not_cyberbullying containing significantly more instances than others. This imbalance can bias the classifier toward the majority class, thereby reducing sensitivity to minority classes such as gender-based or religion-based cyberbullying. While the model performs well in aggregate metrics, its performance on underrepresented classes should be interpreted with caution. Future work may consider addressing this issue using resampling techniques (e.g., SMOTE), class weighting, or cost-sensitive learning.

In light of these limitations, further research is recommended to enhance the robustness and fairness of cyberbullying detection models. This may include the integration of multimodal data, exploration of more sophisticated learning algorithms, and validation across multiple social media platforms to ensure greater applicability and real-world impact.

4. Results and Discussion

4.1. Training Results and Model Evaluation

This study implemented and comparatively evaluated three machine learning algorithms SVM, Naive Bayes, and Random Forest to classify cyberbullying risk in social media text data. Each algorithm was trained using the same preprocessed dataset and feature representations, with TF-IDF applied as the vectorization technique to convert textual inputs into numerical format suitable for machine learning.

The experimental results obtained from the three models are summarized in Table 4, which highlights their comparative performance across all evaluation metrics. These results provide insight into the relative strengths and weaknesses of each algorithm in handling cyberbullying detection within short-text, high-dimensional data environments such as Twitter.

Table 4. Performance Comparison of Classification Algorithms for Cyberbullying Detection

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
SVM	0.8231	0.8236	0.8231	0.8229
Naive Bayes	0.7694	0.7592	0.7649	0.7562
Random Forest	0.8153	0.8185	0.8153	0.8154

From the evaluation results presented in Table 4, it is evident that the SVM algorithm outperforms the other two classifiers, achieving the highest accuracy and F1-score. These outcomes suggest that SVM is particularly adept at capturing complex linguistic patterns embedded in social media comments. Its capability to handle high-dimensional, sparse feature spaces such as those generated by TF-IDF representations makes it well-suited for modeling the nuanced textual expressions commonly found in cyberbullying-related content.

4.2. Visualization of Classification Results

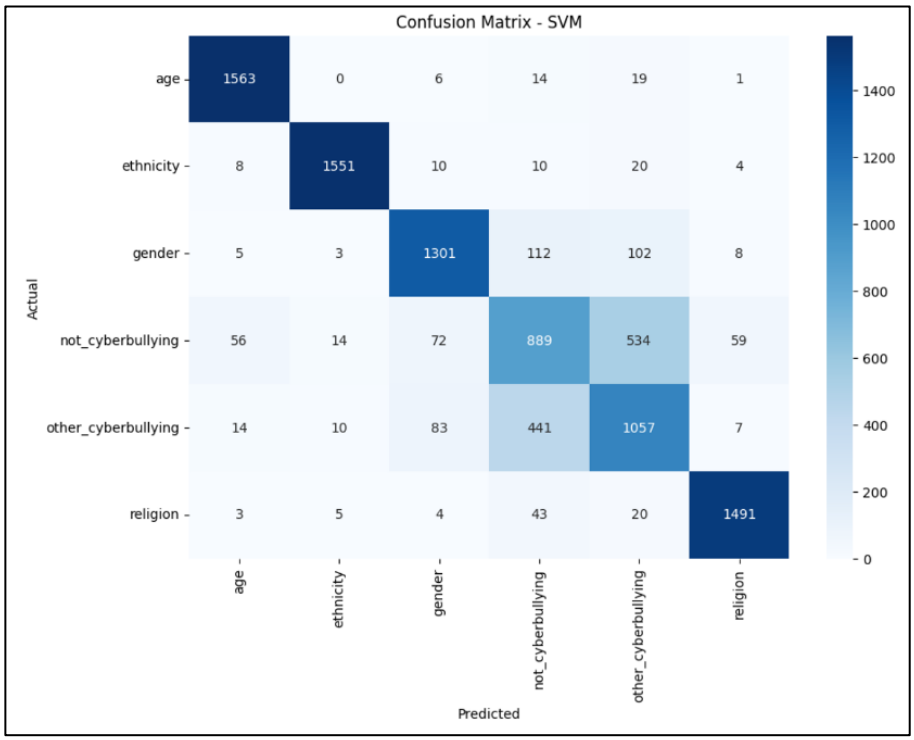

Figure 4. Confusion Matrix of SVM Classification Results for Cyberbullying Detection

Figure 4 presents the confusion matrix of the SVM model across six cyberbullying categories, providing a detailed view of per-class performance and misclassification patterns. Overall, the model performs strongly, especially in the age, ethnicity, and religion categories. For instance, it correctly classifies 1,563 of 1,603 age-related samples and 1,551

ISSN 2579-7069

of 1,603 ethnicity-related samples, yielding accuracies above 95%. This indicates that explicit, well-structured linguistic cues such as clear slurs or direct references are effectively captured by the model.

In contrast, substantial misclassification occurs between the not_cyberbullying and other_cyberbullying categories, with 534 instances of not_cyberbullying predicted as other_cyberbullying and 441 in the opposite direction. These errors reflect strong semantic overlap and ambiguity between the two classes, often driven by figurative or emotionally charged language (e.g., sarcasm, joking insults, or ironic remarks) that may sound negative but does not always constitute bullying. Without richer contextual information such as conversation history, tone, or user intent the model is more vulnerable to both false positives and false negatives in these borderline cases.

This challenge is further intensified by class imbalance, as the not_cyberbullying category contains significantly more examples than most other classes, encouraging bias toward majority patterns. The combination of subtle linguistic cues and skewed class distribution highlights the limitations of the current feature set and model. Addressing these issues will likely require more expressive text representations, the inclusion of contextual metadata, and potentially the use of deep learning architectures that can better infer meaning in nuanced, context-dependent social media content.

4.3. Analysis and Interpretation of Results

The evaluation results yield several key insights that enhance the understanding of how different machine learning algorithms behave when applied to the task of text classification in the domain of cyberbullying detection. These insights emphasize not only differences in model performance, but also the underlying factors that influence their predictive capabilities.

The SVM algorithm consistently delivered the most robust and superior performance among the three models evaluated. Its effectiveness can be attributed to its capacity to handle high-dimensional feature spaces, such as those produced by TF-IDF representations commonly used in textual data. SVM leverages its kernel-based mechanism to map input data into a higher-dimensional space, allowing it to construct an optimal hyperplane that maximally separates data points across distinct classes. The model demonstrated a strong balance between precision and recall across nearly all categories, indicating not only its accuracy in detecting explicit cases of cyberbullying but also its sensitivity in identifying subtle or implicit forms of verbal aggression that are often embedded in user-generated content on social media platforms.

In contrast, the Naive Bayes classifier exhibited generally lower performance, with particularly weak recall in the not_cyberbullying category, achieving only 43%. This outcome suggests a significant tendency of the model to misclassify neutral comments as instances of cyberbullying, leading to a higher rate of false positives. The performance limitations of Naive Bayes may stem from its foundational assumption of conditional independence between features an assumption that does not align well with the intricate and context-dependent nature of social media language. Given the rich semantic content and informal linguistic patterns typical of user comments, the model struggled to accommodate diverse sentence structures and variations in word usage, further reducing its classification accuracy in real-world settings.

The Random Forest algorithm, based on ensemble learning techniques, achieved accuracy levels comparable to those of SVM. However, it exhibited inconsistencies in classifying comments that were truly neutral versus those that were ambiguous or contextually complex. This is evidenced by fluctuating values of precision and recall, particularly in the not_cyberbullying category. While Random Forest is advantageous in capturing non-linear interactions between features, it appears to overfit literal word patterns in short-text data such as tweets, making it less effective in detecting indirect language elements such as sarcasm, irony, or implied hostility common features of cyberbullying discourse.

Taken together, these findings reinforce the importance of selecting classification algorithms based on the nature of the data, feature representation strategy, and context of application. For text-based cyberbullying detection, the combination of SVM and TF-IDF emerges as a highly effective approach, demonstrating a strong capacity to identify a broad range of cyberbullying expressions with both efficiency and accuracy. This suggests that future research and practical implementations in this domain should consider leveraging SVM, particularly when dealing with sparse, high-dimensional textual data and nuanced linguistic signals.

4.4. Impact of Data Imbalance

The distribution of data across class labels in the dataset was notably imbalanced, as illustrated in Figure 2. This class imbalance poses a significant challenge to the performance and fairness of the classification model. In particular, the not_cyberbullying category comprises a disproportionately larger number of instances compared to other classes such as gender, religion, and other_cyberbullying. This skewed distribution introduces bias during the training process, wherein the model tends to prioritize learning from the majority class while underrepresenting minority classes an issue commonly referred to as the class imbalance problem in machine learning literature.

As a consequence, the classifier demonstrates reduced sensitivity to minority categories, which is especially problematic in cyberbullying detection tasks where nuanced and context-dependent expressions such as sarcasm, indirect insults, or emotionally charged but non-abusive language are prevalent. Comments related to minority classes often exhibit complex linguistic features that are subtle in tone and intent, making them inherently more difficult to identify using models that rely primarily on frequency-based feature representations like TF-IDF.

For example, in the predictions generated by the SVM model, a frequent misclassification pattern emerged wherein not_cyberbullying instances were incorrectly labeled as other_cyberbullying, and vice versa. This confusion highlights the existence of semantic overlap and ambiguous linguistic boundaries between the two categories. Neutral statements that contain emotionally negative undertones or ambiguous language are particularly susceptible to misclassification, especially in the absence of broader conversational context or emotional indicators. This further underscore the limitations of traditional text representation methods in capturing the intricate social cues embedded in online discourse.

The observed classification errors also suggest that the model lacks sufficient contextual understanding to differentiate between benign expressions and subtle forms of harassment. Without mechanisms to interpret intent, irony, or sarcasm, the model may inaccurately infer maliciousness based solely on surface-level word patterns.

To address these challenges, several strategies can be explored in future work. Techniques such as resampling including oversampling minority classes or undersampling majority classes can help balance the training data. Additionally, class weight adjustment during model training can penalize misclassifications of underrepresented classes more heavily. More advanced approaches such as Synthetic Minority Oversampling Technique (SMOTE) or ADASYN can generate synthetic examples to enhance representation of minority classes. These strategies aim to promote fairness and ensure that the model not only achieves high overall accuracy, but also maintains consistent and equitable performance across all cyberbullying categories.

4.5. Alignment with Prior Studies

The findings of this study are consistent with previous research in the field of cyberbullying detection using machine learning. Kusuma and Nugroho [8] emphasized the superior performance of SVM in handling short-text data such as tweets. Their study demonstrated that SVMs are particularly effective in managing high-dimensional feature spaces derived from TF-IDF representations an observation that aligns with the results of the current research. SVM consistently achieved higher classification stability and precision compared to traditional classifiers like Naive Bayes and Decision Trees, both of which were more susceptible to noise and ambiguity in short, informal texts.

Similarly, Dinakar et al. [2] highlighted the importance of granular and specific labeling in enhancing the performance of cyberbullying classifiers. Their study found that segmenting labels by categories such as race, religion, and sexual orientation enabled the model to better capture the distinctive linguistic features associated with each type of bullying. In contrast, aggregating all forms of bullying into a single class reduced the model's ability to recognize specific patterns. This insight directly supports the methodology employed in the present study, which retained six distinct cyberbullying labels to preserve linguistic granularity and semantic clarity.

By adopting a multi-label classification structure and leveraging a robust algorithm like SVM, this study has demonstrated that accurate and context-sensitive cyberbullying detection is not solely contingent upon algorithm selection, but also relies heavily on thoughtful label design and feature representation strategies. The integration of

 ISSN 2579-7069

these components enhances the model's ability to discern subtle distinctions in online communication, thereby contributing to more effective and socially responsible detection frameworks.

5. Conclusion

This study developed a text-based classification system for detecting cyberbullying on social media using SVM with TF-IDF feature representation. The model effectively handled the high-dimensional, sparse, and noisy nature of Twitter-style text and consistently outperformed Naive Bayes and Random Forest, achieving accuracy, precision, recall, and F1-score of around 82.3%. SVM performed particularly well in categories with clear and explicit linguistic cues, such as age, ethnicity, and religion, where it accurately identified content containing overt verbal aggression.

However, the model faced difficulties in distinguishing between the not_cyberbullying and other_cyberbullying categories, where precision and recall were noticeably lower. These challenges are linked to semantic ambiguity, implicit hostility, sarcasm, and emotionally charged but non-explicit language, as well as imbalanced class distribution that biases learning toward majority classes. The reliance on TF-IDF, which primarily captures frequency-based lexical information, also limits the system's ability to fully represent contextual meaning, emotional tone, and pragmatic intent in social media discourse.

Future work should focus on expanding and diversifying the dataset across platforms, incorporating richer contextual features (e.g., timestamps, user interaction networks, sentiment signals), and adopting deep learning models such as LSTM and transformer-based architectures like BERT to better capture subtle and context-dependent forms of cyberbullying. In parallel, techniques for handling class imbalance such as oversampling, undersampling, cost-sensitive learning, or ensemble-based approaches are needed to improve fairness and robustness across all categories. With these enhancements, cyberbullying detection systems can become more accurate, context-aware, and impactful in promoting safer and more inclusive digital environments.

6. Declarations

6.1. Author Contributions

Author Contributions: Conceptualization, S.Y. and A.I; Methodology, S.Y. and A.I.; Software, A.I.; Validation, S.Y.; Formal Analysis, S.Y.; Investigation, A.I.; Resources, S.Y.; Data Curation, A.I.; Writing Original Draft Preparation, S.Y.; Writing Review and Editing, S.Y. and A.I.; Visualization, A.I. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Espinosa Anke, T. Declerck, D. Gromann, Z. Zhang, and L. Luo, "Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter," *Semant. Web*, vol. 10, no. 5, pp. 925–945, 2018, doi: 10.3233/SW-180338.

[2] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," *in Proc. Int. AAAI Conf. Web social media*, vol. 5, no. 3, pp. 11–17, 2021, doi: 10.1609/icwsm.v5i3.14209.

[3] P. Yi and A. Zubiaga, "ID-XCB: Data-independent Debiasing for Fair and Accurate Transformer-based Cyberbullying Detection," *in Proc. Int. AAAI Conf. Web social media*, vol. 19, no. 1, pp. 2143–2154, 2025, doi: 10.1609/icwsm.v19i1.35924.

[4] M. Abusaqer and J. Saquer, "A Comparative Analysis of Transformer and Traditional ML Models for Cyberbullying Detection on Twitter (now X)," *in Proc. IEEE 49th Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, Toronto, ON, Canada, 2025, pp. 1607–1612, doi: 10.1109/COMPSAC65507.2025.00216.

[5] P. Ferreira, N. Pereira, H. Rosa, S. Oliveira, L. Coheur, and S. Francisco, "Towards Cyberbullying Detection: Building, Benchmarking and Longitudinal Analysis of Aggressiveness and Conflicts/Attacks Datasets from Twitter," *IEEE Trans. Affective Comput.*, vol. 16, no. 3, pp. 1473–1487, Jul.–Sep. 2025, doi: 10.1109/TAFFC.2024.3518587.

[6] T. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron Spectroscopy Studies on Magneto-Optical Media and Plastic Substrate Interface," *IEEE Transl. J. Magn. Jpn.,* vol. 2, no. 8, pp. 740–741, Aug. 1987, doi: 10.1109/TJMJ.1987.4549593.

[7] F. Akar, "Performance Analysis of NLP-Based Machine Learning Algorithms in Cyberbullying Detection," *Erzincan Univ. J. Sci. Tech*nol., vol. 17, no. 2, pp. 445–459, 2024, doi: 10.18185/erzifbed.1474112.

[8] B. I. Kusuma and A. Nugroho, "Cyberbullying Detection on Twitter Uses the Support Vector Machine Method," *J. Tek. Inform. (JUTIF),* vol. 5, no. 1, pp. 11–17, 2024.

[9] F. W. Alsaade and M. S. Alzahrani, "Transformer Learning-Based Neural Network Algorithms for Identification and Detection of Electronic Bullying In Social Media," *Demonstr. Math*., vol. 57, no. 1, pp. 20230118, 2024, doi: 10.1515/dema-2023-0118.

[10] M. J. Rao, P. Prasanthi, B. Ramakrishna, K. G. D. Prasad, and M. Ramanaiah, "Implementing A Support Vector Machine Algorithm on Social Media Platforms to Detect and Restrict Cyberbullying Conversations," in Recent Advancements in Product Design and Manufacturing Systems (IPDIMS 2023), B. B. V. L. Deepak, M. R. Bahubalendruni, D. Parhi, and B. B. Biswal, Eds. Singapore: *Springer*, 2025, doi: 10.1007/978-981-97-6732-8_36.

[11] A. A. Mathpati, H. M. Sadriwala, and S. Shinde, "Dynamics of Cyberbullying on Twitter: ML Detection Models and The Catalytic Role of Tweets Engagement Metrics," *in Proc. 14th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, Noida, India, 2024, pp. 526–531, doi: 10.1109/Confluence60223.2024.10463218.

[12] I. A. Asqolani and E. B. Setiawan, "Hybrid Deep Learning Approach and Word2Vec Feature Expansion for Cyberbullying Detection on Indonesian Twitter," *Indones. J. Inf. Sci.,* vol. 28, no. 4, pp. 123–136, 2023.

[13] A. Mishra, S. Sinha, and C. P. George, "Shielding Against Online Harm: A Survey on Text Analysis to Prevent Cyberbullying," *Eng. Appl. Artif. Intell.,* vol. 133, pt. D, p. 108241, 2024, doi: 10.1016/j.engappai.2024.108241.

[14] T. Mahmud, M. Ptaszynski, and F. Masui, "Exhaustive Study Into Machine Learning And Deep Learning Methods For Multilingual Cyberbullying Detection In Bangla And Chittagonian Texts," *Electronics*, vol. 13, no. 9, p. 1677, 2024, doi: 10.3390/electronics13091677.

[15] L. Komati and K. Y. Reddy, "Cyberbullying Detection on Social Media: Leveraging TF-IDF and LSTM for Robust Classification," *J. Data Acquis. Process*., vol. 40, no. 1, pp. 56–71, 2025.

[16] S. Chen, J. Wang, and K. He, "Chinese cyberbullying detection using XLNet and deep Bi-LSTM hybrid model," *Information*, vol. 15, no. 2, p. 93, 2024.

[17] R. Joshi and A. Gupta, "Performance comparison of simple Transformer and Res-CNN-BiLSTM for cyberbullying classification," *arXiv preprint*, arXiv:2206.02206, 2022.

[18] A. G. Philipo, D. S. Sarwatt, J. Ding, M. Daneshmand, and H. Ning, "Assessing text classification methods for cyberbullying detection on social media platforms," *arXiv preprint*, arXiv:2412.19928, 2024.

[19] H.-Y. Chen and C.-T. Li, "HENIN: Learning heterogeneous neural interaction networks for explainable cyberbullying detection on social media," *arXiv preprint*, arXiv:2010.04576, 2020.

[20] M. S. Akter, H. Shahriar, and A. Cuzzocrea, "A trustable LSTM-Autoencoder network for cyberbullying detection on social media using synthetic data," *arXiv preprint*, arXiv:2308.09722, 2023.