Leveraging A Hybrid Machine Learning Model for Enhanced Cyberbullying Detection

Fenny Syafariani[1*], Muhamad Safiih Lola[2], Sharifah Sakinah Syed Abd Mutalib[3], Wan Nuraini Fahana Wan Nasir[4], Abdul Aziz K. Abdul Hamid[5], Nurul Hila Zainuddin[6]
[1,2,3,4,5]Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, Malaysia
[6]Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Malaysia
[1]r.fenny.syafariani@email.unikom.ac.id, [2]safiihmd@umt.edu.my, [3]s.sakinah@umt.edu.my, [4]wannuraini.fahana@umt.edu.my, [5]abdulazizkah@umt.edu.my, [6]nurulhila@fsmt.upsi.edu.my
***Corresponding Author**

Article Info

Article history:

Submission November 12, 2024
Revised April 9, 2025
Accepted April 25, 2025
Published April 30, 2025

Keywords:

Machine Learning
Cyberbullying Detection
Hybrid SVM-LDA
Digital Technology

ABSTRACT

Cyberbullying is a form of bullying that occurs through digital technology on various social media platforms. This issue has become critical, particularly when it involves racial statements that can threaten community harmony. Many researchers worldwide are working on solutions for automatic hate speech and cyberaggression detection using different machine learning models. **This study aims to** introduce a novel hybrid method for detecting cyberbullying, utilizing a combination of Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA), collectively referred to as SVM-LDA. The methodology involves integrating **SVM and LDA** techniques. The models efficiency was assessed using various metrics, offering a comparative analysis of the hybrid model against individual machine learning models. **The results show that** the proposed hybrid model achieved 96.1% accuracy and outperformed single machine learning models on the Twitter dataset. The hybrid model also demonstrated robustness in handling imbalanced classes for cyberbullying detection. **The proposed** SVM-LDA hybrid approach shows significant potential in effectively detecting cyberbullying, even in cases of class imbalance. This model offers a more robust solution compared to traditional single machine learning models in detecting cyberaggression.

DOI: https://doi.org/10.34306/att.v7i2.536

1. INTRODUCTION

Recently, cyberbullying has become a major issue worldwide, especially among young people who frequently use the internet and social media platforms like Facebook, Twitter, and Instagram [1]. This phenomenon involves using electronic communication to insult, intimidate, or threaten others, causing significant psychological harm [2]. Cyberbullying attacks are typically carried out in various ways, such as spreading false information, sharing private content without permission, posting offensive comments, sending threats, or impersonating others. The internet's anonymity and wide reach help cyberbullies harass their victims, which can cause severe emotional distress and, in extreme cases, lead to suicide [3], [4], [5]. To address this growing problem, researchers explored methods to detect and prevent cyberbullying that also have strong relevance to the Sustainable Development Goals (SDGs), especially in creating a safer and more inclusive digital environment, which supports Goal 4 (Quality Education) and Goal 16 (Peace, Justice and Strong Institutions). A powerful approach is to leverage machine learning algorithms capable of analyzing vast datasets to uncover relationships and patterns [6]. However, detecting cyberbullying is still a challenging task, and relying on just one machine learning algorithm is not enough [7], [8].

In this research, we present an innovative hybrid model that combines SVM with LDA for detecting cyberbullying in online text. The LDA is a popular statistical tool that often outperforms more sophisticated modern machine learning techniques in several cases, such as remote sensing [9] and violence detecting [10]. Discriminative classifiers aim to create a decision boundary that best separates different classes. LDA is attractive because it has low model complexity and can capture key data characteristics (mean and covariance) from limited training data, then use these to estimate the decision boundary [11, 12]. However, it also often used as a feature reduction technique in the preprocessing step for classification and machine learning applications [13]. Additionally, SVMs function as discriminative classifiers by utilizing a local separation index, known as the margin [14].

2. LITERATURE REVIEW

Cyberbullying on social media platforms, especially Twitter (now rebranded as X), poses a significant issue due to its detrimental effects on individuals, particularly the youth who frequently use these platforms [15]. These platforms facilitate harassment, threats, and humiliation, which can lead to considerable emotional and psychological damage for the victims. Detecting cyberbullying is a challenging task that involves various considerations, including the language of online interactions, the identities of the message sender and recipient, and other factors. Relying on a single machine learning algorithm may not be adequate for detecting all forms of cyberbullying [16]. For instance, while some algorithms may excel at identifying specific types of cyberbullying, others might be more effective with different kinds. The literature presents a variety of methods based on machine learning, including Naïve Bayes (NB), Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM) [17], [18].

The NB Classifier is a probabilistic supervised learning method that mostly uses metrics from training data to determine how likely an item is to belong to a certain class. The NB classifier is commonly applied in areas such as text classification, sentiment analysis, spam filtering, and recommendation systems. It assumes that when conditioned on the target class, the features (or attributes) are independent. In other words, given the class variable, the value of one feature does not rely on the value of any other feature [19]. Additionally, Table 1 provides a comprehensive comparison of the techniques and evaluation metrics used in previous studies within this domain.

Table 1. Literature review

References	Proposed Model	Datasets	The Best Accuracy
[20]	Random Forest (RF), SVM, Decision Tree (DT), and NB	Obtained from Twitter	0.913
[21]	RF with Term Frequency - Inverse Document Frequency (TF-IDF), NB, SVM, Logistic Regression (LR), and XGBoost	Hate Speech Dataset obtained from an Association for Computational Linguistics, Github	0.830
[22]	Gated Recurrent Units (GRU), Convolutional Neural Network (CNN), and Hybrid CNN-GRU	Obtained from Twitter	0.790
[23]	NB, Multilayer Perceptron (MLP), SVM, and AdaBoost (AB)	Obtained from Twitter	0.834
[24]	SVM, RF, and Recurrent Neural Network (RNN)	Obtained from Twitter	0.946
[25]	XGBoost, SVM, LR, NB, Feed Forward Neural Network (FFNN)	Obtained from Reddit, YouTube, Twitter, and Wikipedia	0.920
[26]	NB, RF, DT, SVM, Deep Neural Network (DNN)	Obtained from Twitter	0.746

[27]	Long Short-Term Memory (LSTM), NB, DT, LR, SVM, RF, and Hybrid LSTM-CNN	Obtained from Twitter and Facebook	0.975
[28]	RF, SVM, NB, RNN, CNN, and Hybrid RF-CNN	Obtained from Twitter and Instagram	0.984

[20] presents an advanced machine learning system designed to automatically identify hate speech within Arabic social media platforms. This system captures various emotional types and employs diverse feature sets for analysis. Four machine learning algorithms SVM, NB, RF, and DTare applied, utilizing emotion-related, profile-related, and TF-IDF features. Among these, RF with profile-related and TF-IDF features has the highest accuracy of 91.3% among the tested models.

Similarly, [21] focuses on classifying both fake news and hate speech by extracting features from content labeled as real or fake news. This study employs the XGBoost, Naive Bayes (NB), and Logistic Regression (LR) algorithms with TF-IDF features. XGBoost achieves an accuracy of 83.0%, indicating that 17% of the data with hateful content is misclassified. Furthermore, the model attains a precision of 82.0%, meaning 18% of the hateful data is also misclassified. In [22], the issue of hate speech within the Saudi Twitter sphere is explored through various deep learning methods. Experiments are conducted on two datasets using, GRU, CNN, a hybrid CNN-GRU, and BERT.

Reference [23] explores the automatic detection of racism and hate speech in Indonesian tweets by employing various machine learning models. The models consist of Naive Bayes (NB), Support Vector Machine (SVM), AdaBoost (AB), and Multi-Layer Perceptron (MLP). To mitigate the issue of class imbalance, the study applies the Synthetic Minority Oversampling Technique (SMOTE), and experiments are conducted using features with and without SMOTE. The MLP model utilizing SMOTE features achieves an accuracy of 83.4%, while the AdaBoost and Naive Bayes models, using non-SMOTE data, attain an accuracy of 71.2%.

[24] focuses on identifying hate speech in social media information. In this study, audios were extracted from various videos which then were converted into text using a speech-to-text converter. The tests use Recurrent Neural Network (RNN), Random Forest (RF), Support Vector Machine (SVM), and Naive Bayes (NB) models. Two experimental settings are used: the first classifies movies as normal or hostile, while the second divides them into normal, racist, and sexist categories.

[25] introduces an innovative system designed to identify hate speech across various social media platforms, including Twitter, YouTube, Wikipedia, and Reddit. This system utilizes a comprehensive dataset where only 20% is labeled as hateful and 80% of the data is labeled as non-hateful. The study evaluates multiple machine learning algorithms such as FFNN, LR, XGBoost, NB, and SVM and finds that XGBoost achieves the highest accuracy of 92.0%. Similarly, Reference [26] examines hate speech related to Islam on social media. This research develops an automated tool capable of classifying content into strong Islamophobic, weak Islamophobic, and non-Islamophobic categories. Various machine learning algorithms namely RF, NB, DT, LR, DNN, and SVM are tested, with SVM achieving a 74.6% accuracy based on 10 fold cross-validation. A recent study discussed in Reference [27] applies a hybrid approach combining LSTM and CNN for text classification tasks. This research compares several machine learning algorithms, including NB, DT, LR, SVM, RF, and LSTM, and finds that the LSTM-CNN hybrid model surpasses all others with an impressive accuracy of 97.5%. Reference [28] proposes another hybrid model for detecting hate speech on social media, exploring various machine learning algorithms such as RF, SVM, NB, RNN, CNN, and a Hybrid RF-CNN model. The Hybrid RF-CNN model achieves the highest accuracy of 98.4%.

Based on the success of these hybrid models, this research uses the SVM-LDA hybrid approach to improve the detection of racist comments on Twitter. In contrast to hybrid approaches such as CNN-LSTM and RF-CNN, the SVM-LDA model combines two more statistical and classical-based techniques. SVM is an effective machine learning method for classification problems, especially in the context of imbalanced data, as it focuses on separating classes by a maximum margin. LDA is used to reduce the dimensionality of the data while retaining features that can distinguish classes.

3. RESEARCH METHOD

In this research, three baseline models have been investigated and examined for detecting cyberbullying, namely NB, SVM, and LDA. Then, we introduced a hybrid SVM-LDA model. The detailed explanation is given in the following subsections.

3.1. Naïve Bayes (NB)

The NB algorithm, a classification technique grounded in statistical and probabilistic principles, was introduced by the British scientist Thomas Bayes. As a machine learning model, it applies Bayes theorem to predict future outcomes by drawing on past data [29]. A key characteristic of the NB classifier is its strong yet simplistic assumption that each condition or event is independent of the others. The dataset has a label, class, or target as a reference [30]. In a NB classifier, learning is a process that calculates the stochastic value of a case. Below is the equation for the NB algorithm [31]:

$$P(H \mid x) = \frac{P(x \mid H) \cdot P(H)}{P(x)} \tag{1}$$

To explain the equation, the data point x belongs to an unknown class, with $P(x)$ representing the probability of x, $P(H)$ denotes the prior probability of the hypothesis H, while $P(x \mid H)$ refers to the likelihood of x given the hypothesis H. Additionally, $P(H \mid x)$ is the posterior probability of the hypothesis H based on condition x. For classification, certain rules are required to determine the appropriate group for further examination, as outlined below:

$$\textit{Posterior} = \frac{\textit{prior} \times \textit{likelihood}}{\textit{evidence}}. \tag{2}$$

To summarize, posterior is the probability of class appearance, prior is the class before sample introduction, likelihood is the occurrence of sample features in a class, and evidence is the worldwide emergence of sample characteristics. The NB algorithm consists of several stages: First, the number of classes or labels ($P(H)$) is counted, followed by calculating the number of cases for each class ($P(x \mid H)$). Next, all class variables are multiplied, and finally, the results are compared across classes. The NB classifier is designed to identify the class with the highest probability when assigning test data to the most suitable category. Each document is represented by a set of attributes, $x_1, x_2, \ldots, x_n$, where x_1 corresponds to the first word, and $x_1, x_2, \ldots, x_n$ represent Tweet categories. During classification, the algorithm seeks the category with the highest probability (V_{MAP}) for the documents being tested, as described by the following equation [32, 33].

$$V_{MAP} = \arg\max_{V_{jev}} \frac{P(x_1, x_2, \ldots, x_n \mid V_j) \cdot P(V_j)}{P(x_1, x_2, \ldots, x_n)}. \tag{3}$$

The value of P(x1, x2,. . . ,xn) is constant for all categories (Vj). Therefore, the equation is as follows:

$$V_{MAP} = \arg\max_{V_{jev}} P(x_1, x_2, \ldots, x_n \mid V_j)\, P(V_j). \tag{4}$$

The equation can be simplified into the following:

$$V_{MAP} = \arg\max_{v_{jev}} \left(\prod_{i=1}^{n} P(x_i \mid V_j) \right) P(V_j). \tag{5}$$

To describe, V_j is the tweet category, and j is $1, 2, \ldots, n$. In this research, j_1 is in the category of a tweet with negative sentiment, while j_2 is the category of a tweet with positive sentiment. Other than that, j_3 is in the neutral tweet category, and j_4 is a question sentiment tweet category with $P(x_i \mid V_j)$ = probability x_i in category V_j, and $P(V_j)$ = probability of V_j. $P(V_j)$ and $P(x_1 \mid V_j)$ are calculated on the training data where the equation is.

$$P(V_j) = \frac{|\text{doc } j|}{|\text{sample}|}, \tag{6}$$

$$P(x_i \mid V_j) = \frac{n_k + 1}{n + |\text{vocabulary}|}. \tag{7}$$

The total number of documents in all categories is denoted as $|sample|$, while $|doc_j|$ represents the document count for each specific category j. Additionally, n refers to the frequency of a word in each category, and n_k indicates how often a particular word appears. Finally, the total number of words across all categories is summed up as $|vocabulary|$.

3.2. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a deterministic binary classifier that operates on linear functions within a high-dimensional feature space. It can distinguish data by defining decision boundaries based on a subset of feature vectors [34]. The SVM framework relies on optimization algorithms and adheres to the principle of Structural Risk Minimization, which aims to identify the optimal hyperplane for separating two classes in the input data [35], [36].

$$\begin{aligned} &\min \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{s.t} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad \forall i = 1, \ldots, n. \end{aligned} \tag{8}$$

The optimization problem in the equation 8 can be solved using Quadratic Programming with Lagrange Multipliers, where $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ with only the α_i values corresponding to data points that meet the hyperplane equality constraint in equation 8 being non-zero. These α_i values are also known as support vectors.

3.3. Linear Discriminat Analysis (LDA)

The Linear Discriminant Analysis (LDA) classifier, frequently employed in supervised classification tasks, serves as a dimensionality reduction technique [37]. This method, widely applied in statistics and various other fields, identifies a linear combination of functions that distinguishes or separates objects or events across two or more classes. It is most commonly used for feature extraction in pattern classification problems. Simply put, dimensionality reduction techniques are crucial for machine learning applications as they reduce the dimensions (that is, variables) of a particular dataset while retaining most of the data.

The LDA method has been effectively utilized across numerous domains, including face recognition [38], [39], text categorization [40], and gene microarray analysis [41]. The classical LDA method aims to find an optimal transformation that reduces the distance within the same class while increasing the distance between different classes, leading to effective discrimination. Mathematically, this involves solving an optimization problem to determine the direction of w*Rd as follows:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} \tag{9}$$

Where the covariance between classes, S_b, and the within-class covariance, S_w, are defined as follows:

$$S_b = (m_1 - m_2)(m_1 - m_2)^T \tag{10}$$

$$S_w = \sum_{i\in\{1,2\}} \sum_{x} (x - m_i)^2 \tag{11}$$

Here, m_i represents the empirical class means of the mapped data. The matrix $S_w^{-1} S_b$ can be optimized through eigen decomposition to yield the discriminant function w^* [42]. The eigenvector associated with the largest eigenvalue determines w^*. After disregarding the scaling factor, w can be expressed as follows [43]:

$$\mathbf{w}^* = \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{12}$$

A common issue that arises is when S_w turns out to be a singular matrix. To address this weakness, one approach is to add a diagonal matrix (a small scalar value λ multiplied by the identity matrix) to the S_w matrix [44]. This allows us to obtain the discriminant function w^* as follows:

$$\mathbf{w}^* = (\mathbf{S}_w + \lambda\mathbf{I})^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{13}$$

3.4. Proposes SVM LDA Classifier(SVM-LDA)

In this section, we explained the SVM-LDA algorithm. We begin by discussing cases where the data is assumed to be linearly separable. Following that, we address scenarios in which the data cannot be separated linearly.

3.4.1. Linearly Separable Data Cases

The goal of SVM is to find a hyperplane $f(x) = w^T x + b$ that divides the data into two classes (e.g., cyberbullying and non-cyberbullying). The hyperplane is defined by the weights w and bias b, and it should separate the classes in a way that maximizes the margin (distance between the hyperplane and the nearest data points). The objective function of this model is as follows:

$$\min_{w \neq 0, b, \lambda} \frac{1}{2} w^T \left(\lambda S_w + I\right) w \text{s,t } y_i \left(w^T x_i + b\right) \geq 1 \quad \forall i = 1, \ldots, n \tag{14}$$

This equation represents the objective function for the SVM-LDA model, which aims to minimize the weighted sum of the data's covariance and identity matrix. Where S_w is the covariance matrix from Equation 10, and I is the identity matrix with dimension $p \times p$. From the equation above, we can derive:

$$w^T S_w w = \sum_{i=1,2} \sum_{x} \left(w^T (x - m_i)\right)^2 \tag{15}$$

This equation calculates the spread of data points within each class. Here, m_i represents the mean of class i, and x is a data point. The term $w^T (x - m_i)$ measures how far each data point is from its class's mean. Just like in SVM, Equation 13 can be solved using Quadratic Programming with Lagrange Multipliers, where $\Sigma = \lambda S_w + I$:

$$L(w, b, \alpha) = \frac{1}{2} \mathbf{w}^T \Sigma \mathbf{w} - \sum_{i=1}^{n} \alpha \left[y_i \left(\mathbf{w}^T x + b\right) - 1\right] \tag{16}$$

By taking the derivatives with respect to w and b, we obtain the corresponding dual form. This results in an alternative expression called the primal Lagrange, which integrates both SVM and LDA approaches:

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j x_i^T \Sigma^{-1} x_j \quad \text{s.t.} \quad \sum_{i=1}^{n} y_i \alpha_i = 0, \quad \alpha_i \geq 0, \quad i = 1, \ldots, n. \tag{17}$$

The hyperplane function for the SVM-LDA combination is given by:

$$f(x) = \sum_{i=1}^{n} y_i \alpha_i x_i^T \Sigma^{-1} x + b = 0.$$

The SVM-LDA formulation outlined in equation 13 is equivalent to the following formulation, and as previously discussed, it can be efficiently addressed using the general SVM approach. This alignment with the standard SVM method simplifies its application.

The SVM-LDA formulation outlined in equation 13 is equivalent to the following formulation, and as previously discussed, it can be efficiently addressed using the general SVM approach. This alignment with the standard SVM method simplifies its application.

$$\min_{w \neq 0, b, \lambda} \frac{1}{2} \|w^2\| \quad \text{s.t.} \quad y_i \left(w^T \hat{x}_i + b\right) \geq 1 \quad \forall i = 1, \ldots, n \tag{18}$$

Where

$$\hat{w} = \Sigma^{1/2} w \tag{19}$$

$$\hat{x}_i = \Sigma^{-1/2} x_i \quad \text{for } i = 1 \ldots n \tag{20}$$

And

$$\Sigma = \lambda S_w + I \tag{21}$$

By substituting Equations 18, 19, 20 into equation 17, Equation 13 is obtained.

3.4.2. Cases Where The Data Are Not Linear Separable

In general, it is rare to encounter separable cases. The common issue in classification problems is dealing with non-separable cases. For non-separable cases, the goal is to maximize the margin by minimizing classification errors, which is represented using slack variables and denoted as ξ_i, commonly referred to as the soft margin hyperplane. The optimization problem can be written as follows [45], [46]:

$$\begin{aligned} \min_{w\neq 0,b,\lambda,C>0} \quad & \frac{1}{2}w^T(\lambda S_w + I)w + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i\left(w^T x_i + b\right) \geq 1 - \xi_i \quad \forall i = 1,\dots,n \\ & \xi_i \geq 0 \quad i = 1,2,\dots,n \end{aligned} \tag{22}$$

In this context, C represents a positive regularization parameter, while ξ_i indicates the slack variable associated with data point i, corresponding to the training error. To tackle this problem, which is classified as quadratic programming, available SVM software can be utilized as shown in Section III-B. The goal of this formulation is to minimize the cumulative training error while maximizing the margin. C is the coefficient that determines the penalty for classification errors, and ξ_i is called the slack variable. Minimizing $C\sum_{i=1}^{n}\xi_i$ means reducing the error during the training process. The optimization problem in the equation can be solved using Quadratic Programming with Lagrange Multipliers, similar to how it was done in SVM models.

3.5. The Data Set

We utilized a dataset from [47], curated by Fatma Elsafoury, focusing on online bullying and toxicity. This dataset comprises information on various cyberbullying detection efforts, gathered from diverse sources. The dataset, which was acquired from various social media sources like YouTube, Kaggle, Wikipedia Talk pages, and Twitter, consists of texts categorized as either cyberbullying or non-cyberbullying. The dataset consists of a total of 13,471 instances, with a focus on racism-related content. These instances are categorized into various types of cyberbullying, including hate speech, aggression, insults, and toxicity. Specifically, the dataset is divided as follows: 45% of the data is related to hate speech, 30% to insults, 15% to aggression, and the remaining 10% to other forms of toxicity. Table 2 displays a sample of the data utilized.

Table 2. Sample text from the dataset

ID	User	Text	Category
5.77E+17	@AAlwuhaib1977	Muslim mob violence against Hindus in Bangladesh continues in 2014. #Islam http://t.co/C1JBWJwuRc	Hate Speech
5.59E+17	@aymannathem	As soon as ISIS chased all the minorities out of Mosul, the Sunni Arabs were happy to steal their property. So fuck them.	Hate Speech/Insult

3.6. Pre Processing Data

The collected data is still unstructured, with the contents of each sentence written in a non-standard language. This stage will clean the data by removing extraneous characters, converting all data to lowercase, tokenizing, removing stop words, punctuation, lemmatization, and stemming. Proper preprocessing and cleaning of the document are essential to ensure effective model training. There are as many as 1970 “Racism” labels and 11501 “Non-Racism” labels among the 13471 data. The following are two examples of data before and after preprocessing in Table 3. There are punctuation marks such as periods (.) and commas (,) as well as slang words and others. As a result, data cleaning is performed in such a way that noise-free data is obtained [48].

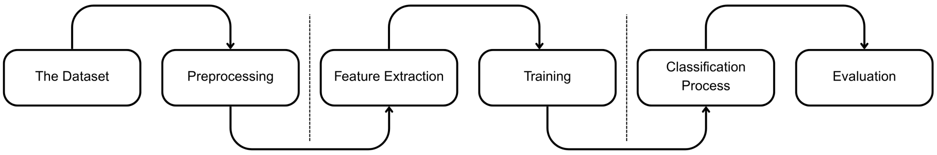


Figure 1. Calculation Result

At the Figure 1 preprocessing stage, several steps are undertaken to clean the data. Tokenization is the first step, where natural text is divided into tokens by removing white spaces, effectively breaking down sentences into individual words. Although this process appears simple, determining the appropriate tokens is quite complex.

Table 3. Sample text before and after the preprocessing.

ID	User	Preprocessing Text	Postprocessing Text
5.77E+17	@AAlwuhaib1977	Muslim mob violence against Hindus in Bangladesh continues in 2014. #Islam http://t.co/C1JBWJwuRc	Muslim mob violence hindu bangladesh continues islam
5.59E+17	@Alfonso_AraujoG	@ardiem1m @MaxBlumenthal It has nothing to do with their grandpas. It is inherited with their religion.	Nothing grandpa inherited religion

Lemmatization, however, takes into account the context of a word and reduces it to its base form. This process is essential for minimizing the number of unique word occurrences and ensuring that similar words are processed in their canonical form. Next, stop words are removed because they offer nothing to the machine learning model's training and merely increase complexity by expanding the feature space. Words like ”a”, ”am”, and ”an” are deleted to boost the model's learning efficiency. Case normalization is then applied to treat words with different cases in the same way, such as ”Racism” and ”racism”. The lemmatization process is more careful as it preserves the meaning of the word in the context of the sentence (Table 4).

Table 4. Results of TF-IDF feature extraction on sample tweets.

ID	Bangladesh	Continues	Grandpa	Hindu	Inherited	Islam	Mob	Muslim	Nothing	Religion	Violence
5.77E+17	0.3780	0.3780	0.0000	0.3780	0.0000	0.3780	0.3780	0.3780	0.0000	0.0000	0.3780
5.59E+17	0.0000	0.0000	0.5000	0.0000	0.5000	0.0000	0.0000	0.0000	0.5000	0.5000	0.0000

3.7. Evaluation

To evaluate their performance, metrics like accuracy, precision, recall, specificity, and the F1 score were used.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$Recall = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

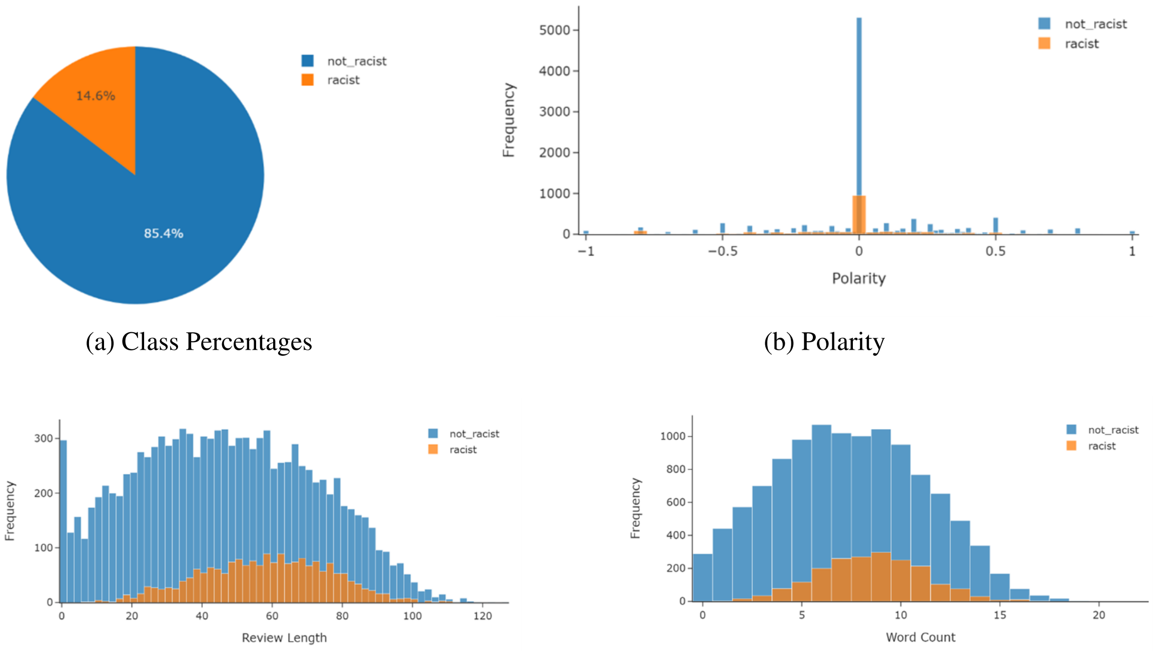
$$Specificity = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

$$F1\ Score = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

The performance of a classifier depends on how well it can correctly classify each instance in a dataset, which is evaluated by the ratio of correct predictions to the total number of predictions. Precision is an essential metric in machine learning, representing the ratio of true positive cases to the total instances predicted as positive by the classifier.

4. RESULT AND DISCUSSION

For providing the distribution of the dataset, with respect to each class, we present in Figure 2. Figure 2a shows that the pie chart visually represents the distribution of responses in two categories: non-racism and racism. Each category is represented by a segment of the pie chart. Most of the tweets belong to the non-racism tweets, the larger blue segment, approximately 85.4%, and 14.6% racism tweets, the smaller orange segment. This pie chart provides a quick snapshot of the distribution, highlighting the prevalence of non-racism responses.



(a) Class Percentages

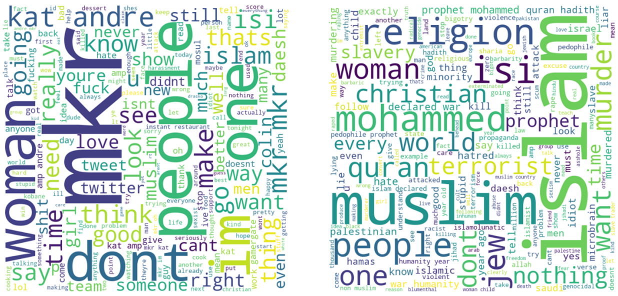
(b) Polarity

(c) Review Length

(d) Word Counts Distribution

Figure 2. Distribution of tweets sentiments in different classes

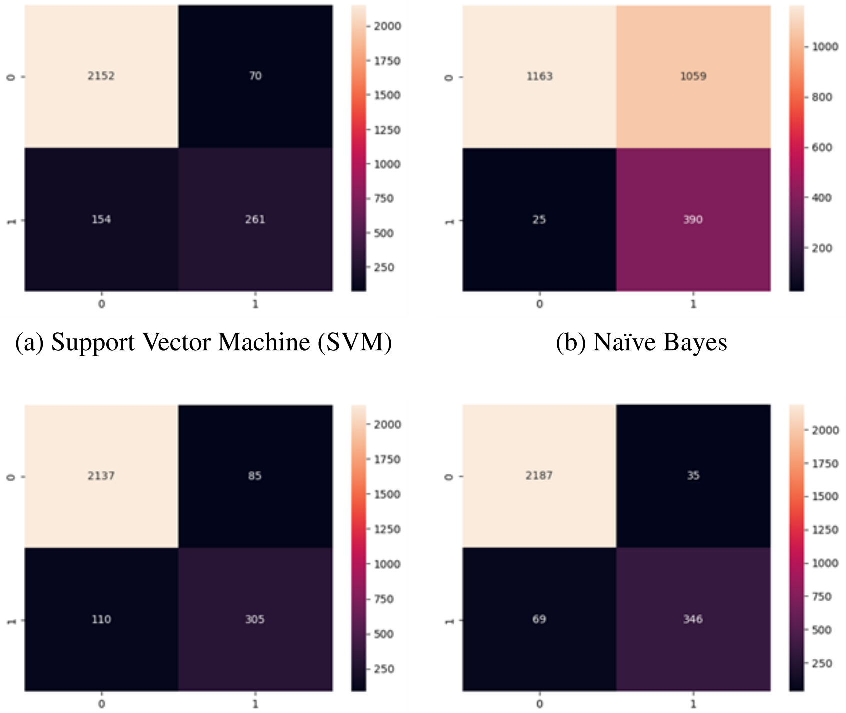
Figure 2b represents review polarity distribution from the tweets. The graph compares two sets of classes. The x-axis represents polarity of the tweets ranging from -1 to 1. The y-axis represents frequency, with values ranging from 0 to 5000. In non-racism Reviews, the majority of data points cluster around the center (near zero polarity). A significant spike in blue bars occurs at this central point, indicating a high frequency of non-racist content with neutral polarity [49, 50]. That Figure 2c histogram represents two categories of reviews, non-racism (in blue) and racism (in orange). The vertical axis shows the frequency (number of reviews), while the horizontal axis represents review length. Both categories exhibit a roughly bell-shaped distribution which is similar to a normal distribution. For review length range, most reviews fall within the range of approximately 10 to 80 units on the review length axis.



(a) Non Racist (b) Racist Class

Figure 3. Word clouds for (a) non racist, and (b) racist class

In Figure 2d, we present word counts distribution. From two categories, non-racism and racism, both categories exhibit a roughly bell-shaped distribution. The peak frequency for both non-racism and racism reviews occurs around a word count of 7-8. Specifically, more frequent (higher bars) in the range of 5 to 15 words in non-racism reviews, but most common word count is around 7-8. Moreover, word counts distribution for racism reviews has lower frequency overall, but it also peaks around 7-8 words, but with significantly fewer occurrences. Additionally, we provide Figure 3 that also shows the word frequency in the dataset through word-cloud.



(a) Support Vector Machine (SVM) (b) Naïve Bayes

(c) Linear Discriminat Analysis (d) Hybrid SVM-LDA

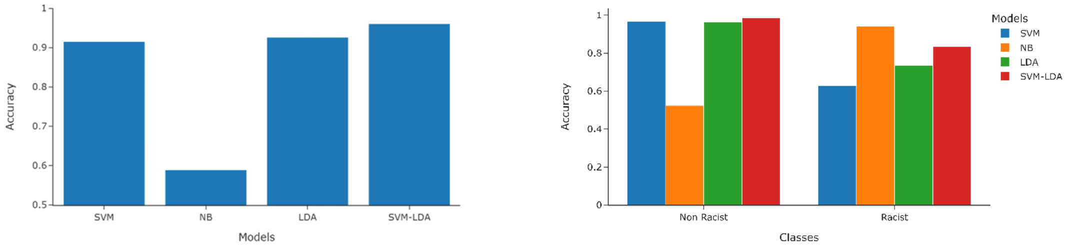
Figure 4. Confusion matrix belong to (a) SVM, (b) NB, (c) LDA and (d) Hybrid SVM-LDA model

Figure 4 presents confusion matrices denoted 0 as non-racism and 1 as racism that evaluate four models using metrics derived from a matrix encompassing four terms. True-positive (TP) refers to instances where offensive text is present in tweets, and the model accurately identifies it as such. False-positive (FP) describes situations where tweets do not contain offensive text, but the model incorrectly predicts them as offensive. False Positives (FP) and False Negatives (FN) in cyberbullying detection can result from factors such as ambiguous language, inadequate feature extraction, and data imbalance.

Table 5. Example of classification results

ID	Text	True Class	SVM	NB	LDA	SVM-LDA
5.75E+17	@MaxBlumenthal Yeap, there is only so much bandwidths for self genocidal Jews, and it's Blumenthal's turn to be the center of attention.	1	0	0	0	1
5.62E+17	RT @TRobinsonNewEra: http://t.co/SCPKHxreTP BREAKING NEWS: 25 muslim men charged with sexual offences against two children in Calderdal #ha...	1	0	0	0	1
5.63E+17	@obsurfer84 The story about her age came from both Aisha and Ursa. It can be found in both Bukhari and Muslim.	1	0	0	0	1
5.77E+17	@dankmtl Are you now going to play the ignorant argumentative asshole and pretend there is no Arabian peninsula?	1	0	0	0	1
5.76E+17	@pNibbler @AlterNet @MaxBlumenthal They want their own Islamic schools to prevent that kind of education.	0	1	1	1	0
5.78E+17	@halalflaws @biebervalue @greenlinerzjm Because what you think is Islam has no resemblance to the real Islam.	0	1	1	1	0
5.79E+17	@harmslesstree2 To suggest that Jews of Israel should subject their lives to the same barbarity that the Copts of Egypt live under is insane	0	1	1	1	0
5.52E+17	@ibnHlophe @eeviewonders @anjemchoudary Murdering Muslims every day is the only way ISIS can keep control.	0	1	1	1	0

Based on the four models evaluated, it is evident from model NB in Figure 4b that this model is unable to address the issue of imbalanced datasets, where the FN value is very low at 25, but the FP value is very high at 1059. This indicates that the NB model is not reliable for detecting cyberbullying, especially in cases of imbalanced classes. Unlike the new hybrid model proposed, Figure 4d, which shows smaller FN and FP values of 69 and 25 respectively, compared to individual models such as SVM with the FN value of 154 and the FP of 70 (Figure 4a), and LDA with the FN value of 110 and the FP of 85 (Figure 4c). This proves that the proposed hybrid model is much better at predicting cyberbullying, even though there are cases of class imbalance in its dataset. We show several datasets that can be well classified by the hybrid SVM-LDA model, but other models cannot, as displayed in Table 5.



(a) Overall Accuracy (b) Class-wise Accuracy

Figure 5. Results of (a) overall accuracy, and (b) class-wise accuracy values for each model

In Figure 5, we visualize the performance of our models, where the hybrid SVM-LDA model stands out with superior results. Achieving an accuracy of 0.961 in detecting cyberbullying, this hybrid model surpasses both the SVM and LDA models in terms of accuracy, precision, specificity, and F1-score. Table 5 further supports these findings by comparing the baseline and proposed hybrid models across several metrics, including accuracy, precision, sensitivity/recall, specificity, and F1-score on the Twitter dataset. Although the hybrid SVM-LDA model slightly underperforms in sensitivity, with a score of 0.834 compared to the NB model 0.940, it excels in the other indices. Overall, based on Table 6, the hybrid SVM-LDA model proves to be the most effective.

Table 6. Comparison of SVM, NB, LDA, and SVM-LDA

Criteria	**SVM**	**NB**	**LDA**	**SVM-LDA**
Accuracy	0.915	0.589	0.926	0.961
Precision	0.789	0.269	0.782	0.908
Sensitivity/Recall	0.629	0.940	0.735	0.834
Specificity	0.968	0.523	0.962	0.984
F1 Score	0.700	0.418	0.758	0.869
MAE	0.085	0.411	0.074	0.039
MSE	0.085	0.411	0.074	0.039
RMSE	0.291	0.641	0.272	0.199
MAPE	1.195E+14	1.809E+15	1.452E+14	5.977E+13
AUC	0.799	0.732	0.848	0.909

The NB model has an overall accuracy score of 0.589. This poor performance demonstrates the NB model inadequacy in predicting racism and non-racism behaviors. The relatively low specificity of 0.523 indicates that the NB model ability to predict the non-racism category is quite poor, whereas the model ability to predict the racism category is excellent with sensitivity score of 0.940. The SVM model results reveal an overall accuracy score of 0.915. This model predicts the non-racism category very well, as evidenced by the relatively high specificity of 0.968 with a sensitivity value of 0.629 which means that the model is also good at predicting the racism category. At the same time, the LDA model yields an overall accuracy of 0.926. The LDA model ability to predict the non-racism category is also very good, as evidenced by the high specificity of 0.962, while the sensitivity of model to predict the racism category is 0.735.

Figure 6 incorporate the Area Under the Curve (AUC) score as part of our evaluation. The AUC score is widely used for assessing binary classification tasks, such as the detection of cyberbullying on social media. This metric assesses a classifier overall performance by considering how well it balances the False Positive Rate (FPR) and the True Positive Rate (TPR) across various threshold values. In cyberbullying detection, the FPR indicates the frequency of non-bullying instances incorrectly labeled as bullying, while the TPR denotes the percentage of genuine bullying cases accurately detected. By providing a measure of the extent to which the model can distinguish between positive and negative classes, AUC provides greater insight than metrics such as accuracy, and enables a fairer assessment of the model performance in cyberbullying detection. The AUC value ranges from 0 to 1; an AUC of 1 denotes perfect classification where all genuine bullying instances

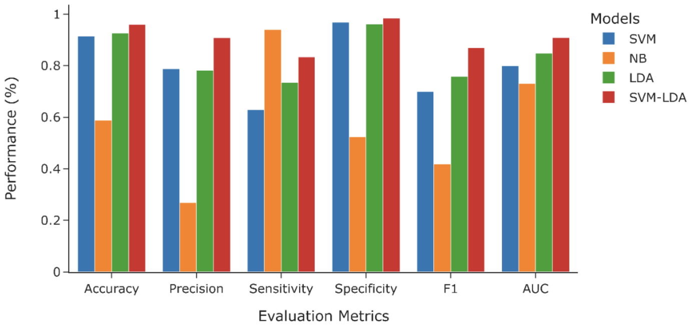


Figure 6. Evaluation metrics each model

are accurately detected and no non-bullying examples are misclassified as bullying whereas an AUC of 0.5 indicates a classifier that performs at the level of random chance. A higher AUC value indicates a more effective model in distinguishing between positive and negative samples. Table 5 displays the AUC values for all models in our investigation, demonstrating that our hybrid SVM-LDA model outperforms the others in detecting cyberbullying on the Twitter platform.

5. MANAGERIAL IMPLICATION

To summarize, this research compares algorithms in machine learning classification in assessing and detecting racism or non-racism tweets. The hybrid SVM-LDA model outperforms the NB, SVM and LDA models in terms of accuracy, precision, specificity, F1 score, and AUC metrics, particularly when there are imbalance cases in the datasets. The results also indicate the requirement for NER system that can generate training data automatically and utilizes machine-labeled data to reduce the cost of labeling and address class imbalance in an online context. Accordingly, this would result in an improvement in the efficiency of the NER system. Hopefully, this study can contribute to the field of machine learning classification for the NER by providing insights into the performance of different algorithms. This research introduces a hybrid SVM-LDA approach that presents distinct advantages compared to conventional cyberbullying detection methods. Although the model proposed in this research is focused on cyber bullying detection, the concepts and architecture used can be extended to various other applications in the field of Natural Language Processing (NLP).

6. CONCLUSION

Cyberbullying is becoming more prevalent on social media platforms like Twitter, making it crucial to automatically detect and stop it to prevent further spread. This research focuses on using sentiment analysis to identify both racist and non-racist content. To address this, we present an innovative hybrid method that combines Support Vector Machine (SVM) with Linear Discriminant Analysis (LDA) for detecting cyberbullying on Twitter. Our method harnesses the strengths of both SVM and LDA to extract pertinent features from text data. Extensive testing and assessment have shown that our approach effectively identifies cyberbullying content. By integrating SVM with LDA, our model proficiently analyzes and classifies textual data, offering better performance for cyberbullying detection. Our innovative SVM-LDA hybrid approach shows considerable potential for detecting cyberbullying even in the case of imbalanced datasets.

By combining these techniques, we have created a robust tool for identifying and addressing this pressing social issue. While our hybrid SVM-LDA model shows promising results, there are potential limitations to consider, such as handling false positives, which could lead to incorrect classifications of non-racist content as racist. It is crucial to continuously refine these models to minimize errors and ensure fairness, transparency, and accountability in their application. While our hybrid SVM-LDA model shows promising results, there are potential limitations to consider, such as handling false positives, which could lead to incorrect classifications of non-racist content as racist. Additionally, future research could adapt this model to handle multilingual datasets or explore its applicability in detecting other forms of online harassment, such as hate speech or gender-based discrimination.

7. DECLARATIONS

7.1. About Authors

Fenny Syafariani (FS) https://orcid.org/0009-0006-0905-623X

Muhamad Safiih Lola (MS) https://orcid.org/0000-0001-9287-7317

Sharifah Sakinah Syed Abd Mutalib (SS) https://orcid.org/0000-0002-3803-4578

Wan Nuraini Fahana Wan Nasir (WN) https://orcid.org/0000-0003-1564-5111

Abdul Aziz K. Abdul Hamid (AA) https://orcid.org/0000-0002-3075-7536

Nurul Hila Zainuddin (NH) https://orcid.org/0000-0001-9972-7573

7.2. Author Contributions

Conceptualization: FS, MS, SS, WN, AA and ID; Methodology: FS, MS and WN; Software: FS, AA and ID; Validation: AA, ID, WN; Formal Analysis: FS, MS and WN; Investigation: FS, MS and WN; Resources: SS, WN, AA and ID; Data Curation: WN, AA and ID; Writing Original Draft Preparation: FS; Writing Review and Editing: MS, SS, WN and ID; Visualization: FS, MS SS WN; All authors, FS, MS, SS, WN, AA and ID, have read and agreed to the published version of the manuscript.

7.3. Data Availability Statement

The datasets used to support the findings of this study are available from the direct link in the dataset citation.

7.4. Funding

This work was supported in part by Universiti Malaysia Terengganu (UMT) under Private Partnership Research Grant (PPRG) Vot No. 55373; We thank Universiti Malaysia Terengganu for providing funding support for this project (UMT/PPRG/2022/55373).

7.5. Declaration of Conflicting Interest

The authors declare that they have no conflicts of interest, known competing financial interests, or personal relationships that could have influenced the work reported in this paper.

REFERENCES

[1] E. Bozzola, G. Spina, R. Agostiniani, S. Barni, R. Russo, E. Scarpato, A. Di Mauro, A. Di Stefano, C. Caruso, and G. Corsello, "The use of social media in children and adolescents: Scoping review on the potential risks," *International Journal of Environmental Research and Public Health*, vol. 19, p. 9960, 2022.

[2] M. Vismara, N. Girone, D. Conti, G. Nicolini, and B. Dell'Osso, "The current status of cyberbullying research: A short review of the literature," *Current Opinion in Behavioral Sciences*, vol. 46, p. 101152, 2022.

[3] K. Subaramaniam, R. Kolandaisamy, A. Jalil, and I. Kolandaisamy, "Cyberbullying challenges on society: A review," *Journal of Positive School Psychology*, vol. 6, pp. 2174–2184, 2022.

[4] D. Kee, M. Al-Anesi, and S. Al-Anesi, "Cyberbullying on social media under the influence of covid-19," *Global Business and Organizational Excellence*, vol. 41, pp. 11–22, 2022.

[5] M. Arisanty and G. Wiradharma, "The motivation of flaming perpetrators as cyberbullying behavior in social media," *Jurnal Kajian Komunikasi*, vol. 10, pp. 215–227, 2022.

[6] J. Hair, J.F. and M. Sarstedt, "Data, measurement, and causal inferences in machine learning: Opportunities and challenges for marketing," *Journal of Marketing Theory and Practice*, vol. 29, pp. 65–77, 2021.

[7] A. Mazari, N. Boudoukhani, and A. Djeffal, "Bert-based ensemble learning for multi-aspect hate speech detection," *Cluster Computing*, pp. 1–15, 2023.

[8] United Nations, "Sustainable development goals," https://sdgs.un.org/goals, 2024.

[9] T. Suesse, A. Brenning, and V. Grupp, "Spatial linear discriminant analysis approaches for remote-sensing classification," *Spatial Statistics*, vol. 57, p. 100775, 2023.

[10] A. Pambudi, N. Lutfiani, M. Hardini, A. R. A. Zahra, and U. Rahardja, "The digital revolution of startup matchmaking: Ai and computer science synergies," in *2023 Eighth International Conference on Informatics and Computing (ICIC)*. IEEE, 2023, pp. 1–6.

[11] C. Lukita, M. Hardini, S. Pranata, D. Julianingsih, and N. P. L. Santoso, "Transformation of entrepreneurship and digital technology students in the era of revolution 4.0," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 5, no. 3, pp. 291–304, 2023.

[12] A. H. Sayed, "Linear discriminant analysis," in *Inference and Learning from Data: Learning*. Cambridge: Cambridge University Press, 2022, pp. 2357–2382.

[13] J. Zhou, Q. Zhang, S. Zeng, B. Zhang, and L. Fang, "Latent linear discriminant analysis for feature extraction via isometric structural learning," *Pattern Recognition*, vol. 149, p. 110218, 2024.

[14] J. A. Rad, K. Parand, and S. Chakraverty, *Learning with Fractional Orthogonal Kernel Classifiers in Support Vector Machines: Theory, Algorithms and Applications*. Springer, 2023.

[15] A. Hermawan, W. Sunaryo, and S. Hardhienata, "Optimal solution for ocb improvement through strengthening of servant leadership, creativity, and empowerment," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 5, no. 1Sp, pp. 11–21, 2023.

[16] S. Paul, S. Saha, and J. P. Singh, "Covid-19 and cyberbullying: Deep ensemble model to identify cyberbullying from code-switched languages during the pandemic," *Multimedia Tools and Applications*, vol. 82, pp. 8773–8789, 2023.

[17] A. Chhabra and D. K. Vishwakarma, "A literature survey on multimodal and multilingual automatic hate speech identification," *Multimedia Systems*, vol. 29, no. 3, pp. 1203–1230, 2023.

[18] E. Aldreabi and J. Blackburn, "Enhancing automated hate speech detection: Addressing islamophobia and freedom of speech in online discussions," in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2023, pp. 644–651.

[19] U. Rusilowati, U. Narimawati, Y. R. Wijayanti, U. Rahardja, and O. A. Al-Kamari, "Optimizing human resource planning through advanced management information systems: A technological approach," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 6, no. 1, pp. 72–83, 2024.

[20] I. Aljarah, M. Habib, N. Hijazi, H. Faris, R. Qaddoura, B. Hammo, M. Abushariah, and M. Alfawareh, "Intelligent detection of hate speech in arabic social network: A machine learning approach," *Journal of Information Science*, vol. 47, no. 3, 2020.

[21] S. Goswami, M. Hudnurkar, and S. Ambekar, "Fake news and hate speech detection with machine learning and nlp," *PalArch's Journal of Archaeology of Egypt / Egyptology*, vol. 17, no. 6, pp. 4309–4322, 2020.

[22] R. Alshalan and H. Al-Khalifa, "A deep learning approach for automatic hate speech detection in the saudi twittersphere," *Applied Sciences*, vol. 10, no. 23, p. 8614, Dec. 2020.

[23] T. Putri, S. Sriadhi, R. Sari, R. Rahmadani, and H. Hutahaean, "A comparison of classification algorithms for hate speech detection," in *IOP Conference Series: Materials Science and Engineering*, vol. 830, Apr. 2020, p. 032006.

[24] U. Bhandary, "Detection of hate speech in videos using machine learning," M.S. thesis, San Jose State University, San Jose, CA, USA, 2019.

[25] J. Salminen, M. Hopf, S. A. Chowdhury, S.-G. Jung, H. Almerekhi, and B. J. Jansen, "Developing an online hate classifier for multiple social media platforms," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, pp. 1–34, Dec. 2020.

[26] B. Vidgen and T. Yasseri, "Detecting weak and strong islamophobic hate speech on social media," *Journal of Information Technology & Politics*, vol. 17, no. 1, pp. 66–78, Jan. 2020.

[27] D. Sultan, M. Mendes, A. Kassenkhan, and O. Akylbekov, "Hybrid cnn-lstm network for cyberbullying detection on social networks using textual contents," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 9, 2023.

[28] N. H. T, P. M, S. E, S. S, B. D. P., G. N., and U. L. S., "Protect: a hybrid deep learning model for proactive detection of cyberbullying on social media," *Frontiers in Artificial Intelligence*, vol. 7, p. 1269366, 2024.

[29] I. S. Mambina, J. D. Ndibwile, and K. F. Michael, "Classifying swahili smishing attacks for mobile money users: A machine-learning approach," *IEEE Access*, vol. 10, pp. 83 061–83 074, 2022.

[30] S. Gibson, B. Issac, L. Zhang, and S. M. Jacob, "Detecting spam email with machine learning optimized with bio-inspired metaheuristic algorithms," *IEEE Access*, vol. 8, pp. 187 914–187 932, 2020.

[31] N. Lutfiani, D. A. Astrieta, V. Wildan, H. Sulistyaningrum, M. R. Anwar, and E. D. Astuti, "Emotional

well-being and psychological support in infertility a multi-modal ai approach," *International Journal of Cyber and IT Service Management*, vol. 5, no. 1, pp. 81–92, 2025.

[32] T. Kim and J.-S. Lee, "Exponential loss minimization for learning weighted naive bayes classifiers," *IEEE Access*, vol. 10, pp. 22 724–22 736, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9722892

[33] D. Dinarwati, M. G. Ilham, and F. Rahardja, "Cybersecurity risk assessment framework for blockchain-based financial technology applications," *ADI Journal on Recent Innovation*, vol. 6, no. 2, pp. 168–179, 2025.

[34] L. Zou, X. Luo, Y. Zhang, X. Yang, and X. Wang, "Hc-dttsvm: A network intrusion detection method based on decision tree twin support vector machine and hierarchical clustering," *IEEE Access*, vol. 11, pp. 21 404–21 416, 2023.

[35] Q. Aini, P. Purwanti, R. N. Muti, E. Fletcher *et al.*, "Developing sustainable technology through ethical ai governance models in business environments," *ADI Journal on Recent Innovation*, vol. 6, no. 2, pp. 145–156, 2025.

[36] R. A. Sunarjo, M. H. R. Chakim, S. Maulana, and G. Fitriani, "Management of educational institutions through information systems for enhanced efficiency and decision-making," *International Transactions on Education Technology (ITEE)*, vol. 3, no. 1, pp. 47–61, 2024.

[37] D. Chopra and R. Khurana, *Introduction to machine learning with Python*. Bentham Science Publishers, 2023.

[38] A. Menon, "Overview of face recognition methodologies: A literature review," *ScienceOpen Preprints*, 2023.

[39] M. I. Afjal, M. N. I. Mondal, and M. A. Mamun, "Segmentation-based linear discriminant analysis with information theoretic feature selection for hyperspectral image classification," *International Journal of Remote Sensing*, vol. 44, no. 11, pp. 3412–3455, 2023.

[40] M. Hardini, Q. Aini, U. Rahardja, R. D. Izzaty, and A. Faturahman, "Ontology of education using blockchain: Time based protocol," in *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*. IEEE, 2020, pp. 1–5.

[41] F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: A review," *Bioengineering*, vol. 10, no. 2, p. 173, 2023.

[42] G. U. Höglinger, C. H. Adler, D. Berg, C. Klein, T. F. Outeiro, W. Poewe, R. Postuma, A. J. Stoessl, and A. E. Lang, "A biological classification of parkinson's disease: the synneurge research diagnostic criteria," *The Lancet Neurology*, vol. 23, no. 2, pp. 191–204, 2024.

[43] J. Liu, W. Xu, F. Zhang, and H. Lian, "Properties of standard and sketched kernel fisher discriminant," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10 596–10 602, 2023.

[44] U. Rahardja, C. T. Sigalingging, P. O. H. Putra, A. Nizar Hidayanto, and K. Phusavat, "The impact of mobile payment application design and performance attributes on consumer emotions and continuance intention," *Sage Open*, vol. 13, no. 1, p. 21582440231151919, 2023.

[45] A. Thielmann, C. Weisser, A. Krenz, and B. Säfken, "Unsupervised document classification integrating web scraping, one-class svm and lda topic modelling," *Journal of Applied Statistics*, vol. 50, no. 3, pp. 574–591, 2023.

[46] F. J. Tehrani, B. Nasihatkon, K. Al-Qawasmi, M. R. Al-Mousa, and R. Boostani, "An efficient classifier: Kernel svm-lda," in *2022 International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI)*, Zarqa, Jordan, 2022, pp. 1–4.

[47] H. Aljalaoud, K. Dashtipour, and A. AI_Dubai, "Arabic cyberbullying detection: A comprehensive review of datasets and methodologies," *IEEE Access*, 2025.

[48] M. Hardini, H. Hetilaniar, S. E. E. Girsang, S. N. W. Putra, and I. N. Hikam, "Advancing higher education: Longitudinal study on ai integration and its impact on learning," *International Journal of Cyber and IT Service Management*, vol. 5, no. 1, pp. 23–30, 2025.

[49] R. Aprianto, E. P. Lestari, E. Fletcher *et al.*, "Harnessing artificial intelligence in higher education: Balancing innovation and ethical challenges," *International Transactions on Education Technology (ITEE)*, vol. 3, no. 1, pp. 84–93, 2024.

[50] A. C. Pramono and W. Prahiawan, "Effect of training on employee performance with competence and commitment as intervening," *Aptisi Transactions on Management*, vol. 6, no. 2, pp. 142–150, 2022.