ISSN 2007-9737

Cyberbullying Detection on Social Media Using Machine Learning Techniques

Abdullah[1,2], Irfan Latif[3,*], Nida Hafeez[1,2], Fida Ullah[1], Grigori Sidorov[1], Edgardo Felipe-Riverón[1], Alexander Gelbukh[1]

[1] Instituto Politecnico Nacional, Centro de Investigacion en Computacion, Mexico City, Mexico

[2] Bahria University, Department of Computer Science, Lahore, Pakistan

[3] University of Central Punjab, Faculty of Information Technology and Computer Science, Lahore, Pakistan

abdullah2025@cic.ipn.mx, Irfan.latif@ucp.edu.pk, nhafeez2024@cic.ipn.mx, fidaullahmohmand@gmail.com, sidorov@cic.ipn.mx, edgardo@cic.ipn.mx, gelbukh@cic.ipn.mx

Abstract. With the rapid proliferation of social media platforms, cyberbullying has emerged as a critical social challenge, leading to severe economic, psychological, and social consequences. A significant proportion of users, particularly adolescents and young adults, report experiencing digital harassment, aggressive online behavior, and emotional distress due to cyberbullying. This study addresses the pressing issue of detecting cyberbullying using advanced machine learning techniques on text data extracted from platforms such as Twitter and Formspring. Given the prevalent issue of insufficient labeled data, manual annotation techniques were used to label an initially unlabeled corpus. Two additional publicly available labeled datasets were also incorporated to enhance model training and validation. Several classifiers, including Random Forest, Voting Classifier, Linear Support Vector Machine, Gaussian Naive Bayes, Standard Support Vector Machine, Convolutional Neural Network, Long-Short-Term Memory, and Bidirectional Encoder Representations from Transformers models, were trained and evaluated. Initial experiments on the raw data yielded suboptimal performance. To address this, the SelectKBest algorithm using the chi-square test was applied for feature dimensionality reduction, improving learning efficiency and model generalization. In the final phase, a hybrid model incorporating the transformer-based Bidirectional Encoder Representations from Transformers architecture and linguistic-lexical features was developed. This refined model achieved a high classification accuracy of 98. 6%, significantly outperforming the previous baselines. The proposed framework also demonstrated better performance in identifying challenging categories such as hate, threat, and sexuality, with F1 scores improving to over 98%. This research emphasizes the importance of annotated data, effective feature engineering, and deep learning techniques in addressing the nuanced and context-dependent nature of cyberbullying detection. Future work will focus on adapting the model to multilingual datasets, particularly for underrepresented languages such as Arabic, Urdu, and Roman Urdu, to broaden its applicability across diverse linguistic communities.

Keywords. Cyberbullying, bullying, sentiment analysis, natural language processing, machine learning, naive Bayes, support vector machine (SVM), linear support vector machine (linear SVM), random forest, voting classifier.

1 Introduction

In modern society, the importance of security and privacy is pivotal, particularly as technological advancements deeply integrate into all aspects of our lives. As we increasingly rely

on digital platforms for daily interactions, the need to secure personal and institutional data against unauthorized access and cyberattacks has increased. In particular, the evolution of blockchain technology has introduced robust security and privacy measures through consensus mechanisms based on FPGA and post-quantum cryptography [5], significantly improving its resilience against quantum attacks [34, 4].

Similarly, the integration of machine learning into wireless sensor networks improves security measures, crucial to maintaining the integrity and confidentiality of transmitted data [25, 2]. Moreover, the development of intelligent intrusion detection systems using big data analytics has become essential in thwarting cyber-attacks in industrial control systems [8, 40], and the use of fog computing for network security management exemplifies the shift towards decentralized data processing to strengthen security and privacy at the network's edge [22].

These advancements underscore the ongoing need for innovation in cybersecurity measures, which is equally crucial in combating emerging threats in digital communication spaces, such as cyberbullying on social media.

This paper explores the intricate challenges that arise with the extensive use of electronic media, with a specific emphasis on the pervasive issue of cyberbullying [56, 51, 49, 50]. Cyberbullying represents a significant misuse of technology, where security and privacy breaches are not just possibilities but realities that affect countless individuals, particularly in Pakistan [56, 51].

This type of bullying, facilitated by the anonymity and reach of digital platforms, raises complex social, psychological, and legal issues [14, 44, 16]. It manifests in various forms and can have devastating consequences on the mental health and well-being of individuals, prompting an urgent need for effective detection and prevention strategies [1, 57].

By examining the current landscape of cyberbullying and its unique challenges, this paper aims to highlight the role of advanced computational techniques, including machine learning and artificial intelligence, in identifying and mitigating such behaviors on social media platforms.

Cyberbullying is a particularly concerning aspect of bullying in today's digital age, considered a criminal offense in many countries, including Pakistan through the Prevention of Electronic Crimes Act 2016 [26, 19]. The paper also explores the Islamic perspective on bullying, emphasizing its prohibition. Cyberbullying, which disproportionately affects adolescents, can lead to severe consequences, including mental health issues, anger, depression, substance abuse, and even suicide attempts, often exacerbated by immature decision-making on social media platforms [45].

The paper advocates for proactive measures to tackle bullying, including open communication and the active involvement of stakeholders such as parents, teachers, and law enforcement agencies. Creating awareness and implementing robust preventive strategies are highlighted as crucial steps. Additionally, the paper introduces the concept of using artificial intelligence, specifically sentiment analysis, as a tool for automatic cyberbullying detection from textual data. However, it acknowledges the existing limitations in current resources for cyberbullying detection on social media platforms [45, 33].

As electronic media usage continues to surge, addressing the complex issues associated with bullying, particularly in the digital realm, becomes imperative. This paper provides comprehensive insights into various dimensions of bullying, its prevalence in Pakistan, and the promising role of artificial intelligence in combating cyberbullying [15]. It underscores the need for a multi-faceted approach involving technology, education, and community engagement to mitigate the adverse effects of bullying in the digital age [33, 52].

The main scientific contributions may be summarized as follows:

— Development of a Specialized Model for Cyberbullying Detection: Using cutting-edge techniques and technology, a complex machine learning model was introduced with the express purpose of identifying cyberbullying on social media platforms.

— Enhanced Data Quality through Advanced Preprocessing and Labeling: Strict preprocessing and data labeling techniques were utilized to guarantee the precision and dependability of the datasets used in the study, improving the data's overall quality.

— Incorporation of Sentiment Analysis for Refined Cyberbullying Detection: Sentiment analysis approaches are used to enhance the model's capacity to understand emotional context in textual data, assisting in the precise identification of cyberbullying occurrences without inadvertently raising false alarms in minor cases.

— Optimization of Model Efficiency and Resource Utilization: Focused on feature selection and characteristic optimization to reduce computational demands and enhance the efficiency of the model, which is crucial for real-time cyberbullying detection.

— Extensive Experimental Validation: Conducted extensive experiments using multiple datasets and iterations, refining the model through continuous evaluation and optimization based on performance metrics like accuracy and F1 score.

— Broadening the Scope to General Cybersecurity Challenges: Positioned the research to extend beyond cyberbullying detection to address broader cybersecurity issues, potentially contributing to wider applications in the field.

The article is structured into six comprehensive sections to explore the complexities of cyberbullying detection using machine learning techniques. Section 1 introduces the context and motivation behind the study, highlighting the significance of security in digital communication and the pressing need for advanced cyberbullying detection methods. It also outlines the key contributions of the research. Section 2 presents a literature review that summarizes current research findings, noting previous advancements and identifying existing gaps that our study addresses. Section 3 details the methods used in the study, including data collection, preprocessing, and the specific machine learning techniques employed. This section explains how datasets were prepared and analyzed to ensure robust model training.

Section 4 focuses on the experimentation and evaluation metrics used to assess the performance of the developed models, providing insights into the effectiveness of different classifiers. Section 5 discusses the results of the study, analyzing the performance of various models during the training and testing phases and reflecting on the implications of these findings. Finally, Section 6 concludes the article by summarizing the key findings, discussing the limitations of current cyberbullying detection methods, and suggesting directions for future research.

2 Related Work

While researching similar problems, we found work conducted in 2023, researchers explored Multi-Class Sentiment Classification, and they utilized three supervised machine learning models: Support Vector Machine (SVM), Decision Tree, and Naïve Bayes, achieving the following accuracy scores, respectively, 0.79, 0.76, and 0.79. The study identified two unaddressed aspects of sentiment classification: negation handling and spam review [3]. In another study, researchers improved sentiment analysis for smart cities using BERT-based Dilated CNN (BERT-DCNN), achieving an 87.1% accuracy. The model's potential extends to various domains, including tourism, academics, and the military [24, 43, 32].

In a 2023 study, researchers examined sentiment analysis methods, with a focus on multilingual and cross-lingual approaches, utilizing NLP techniques and supervised deep learning classifiers. Their proposed Convolutional LSTM (CLSTM) architecture achieved an 85.8% accuracy and 0.86 F1 score on 42,036 Facebook comments. Challenges included aspect-level sentiment analysis limitations, difficulties capturing Bengali sarcasm and idiomatic expressions, and neglecting negation scope and multilingual sentiment analysis [61].

In 2023, the primary research objective was to conduct sentiment analysis on Urdu language

data. The study employed hybrid methods, combining both machine learning and deep learning approaches, to improve the effectiveness of sentiment analysis. Notably, the achieved accuracy was 85%, but it was limited to the Urdu language [55, 27, 38]. In 2023, the study aimed to interpret users' tweets, benefiting academia and applications. They used TF-IDF for feature extraction, and deep intelligent Wordnet lemmatization for noise reduction, and achieved 90% accuracy with a Random Forest network for emotion detection. The paper expressed interest in improving the model further [47].

In 2023, the study analyzed sentiments on Twitter using #Monkeypox to understand public opinions about the virus outbreak. They employed a hybrid CNN-LSTM model with 83% accuracy, achieving 99% specificity, 85% recall, and an 88% F1 score. The confusion matrix indicated 45.42% neutrality and 19.45% negative sentiments, highlighting the need for improved techniques [30].

In a Twitter-focused study using Random Forest, text, user, and network-level features achieved a 90% AUC (Area Under the Curve) to distinguish bullies from regular users [20]. Support Vector Machine (SVM) and Naïve Bayes models, utilizing n-grams and data augmentation techniques, demonstrated their best performance on the non-aggressive class (NAG) and their weakest performance on the covertly aggressive class (CAG).

The study aimed to leverage the hidden feature structure of bullying information [12]. The study compared Random Forest and Support Vector Machine (SVM) models using word embeddings and emoticons as features. Random Forests performed better on similar training datasets, while SVMs showed superior performance on dissimilar texts. On the other hand, the results were not quite accurate and usable. in most of the studies goal is to detect all of the texts in social media that contained aggression [10, 37, 17].

Through Naïve Bayes and Simple SQL classification, this study focused on characterizing abusive chat that included offending words and kinds of racism, respectively. The functions, methods, and algorithms presented in the available Microsoft analysis services were not enough to solve these issues. few studies had a specific goal, about symptomatic elements in game chat, like hateful language and racist sentiments [39, 53]. The investigation was built upon a unique lexical syntactic feature-based language model (LSF), which considered syntactic, lexical, stylistic, structural, and cyberbullying features. LSF demoed excellent sensitivity to capturing mild discriminatory posts and yet decreasing noticeable error.

Although it was more accurate when addressing explicit profanity of terms like "s*" and "f*", such objectionable expressions were not captured correctly. Hence, the dominant goal presupposed is to combat such kind of language effectively on social media to shield teens from it [21]. An A1 is that the study has utilized a C4. The platform utilized a 5-decision tree learner and an instance-based phenomenon recognition algorithm to identify cases of bullying through Formspring.

The research team looked at features associated with the number of offensive words and level of obscenity in their evaluations and resulting in a 78.5% accuracy rate. However, at the same time, it was understood that this accuracy was, of course, not completely exact. This was the key goal of implementing a program for detecting violent content on the website [18].

Our research introduced the above key components, intending to make the model advanced and created specifically for the task of identification of cyberbullying, which employs modern technologies and approaches [7]. We need to do nothing but the best in terms of data, which is only possible through advanced preprocessing and labeling procedures [59].

This guarantees that the lessons have been imparted on the data sets that are not corrupted and well labeled, which are the core of dependable and good outcomes. To improve the accuracy of our NLP, we used sentiment characteristics that work in such a way that NLP understands emotional context and gets to the point when it cannot count minor cyberbullying cases as such [23].

2.1 Research Gap

In the exploration of cyberbullying detection, several advancements have been made, yet significant gaps remain, particularly in the quality of data and the effectiveness of detection models. It has been noted that despite the proliferation of machine learning applications across various platforms, the specific challenges posed by cyberbullying require more targeted approaches, particularly in terms of data handling and feature optimization.

Firstly, the profession has recognized the lack of high-quality, properly labeled data sets as a recurring problem. The majority of currently used models mostly rely on datasets that might not be fully annotated, which causes inconsistent model training and validation. This discrepancy emphasizes the need for sophisticated labeling and preprocessing techniques to guarantee that data sets are accurate and realistic of actual situations.

Furthermore, even though the models that are currently in use are capable of identifying explicit cyberbullying, they frequently fall short in identifying less obvious types of harassment that can be just as destructive. There hasn't been enough attention paid to the requirement for models that can comprehend the complex emotional context of social media interactions. This restriction suggests that more sophisticated sentiment analysis methods may be developed to increase the sensitivity of cyberbullying detection algorithms.

Moreover, the existing research predominantly focuses on English language data, leaving a significant void in multilingual and cross-lingual cyberbullying detection. This gap highlights the need for developing methods that can function effectively across different languages and cultural contexts, ensuring broader applicability and effectiveness.

The research conducted thus far has also revealed a lack of efficiency in terms of computational resources and processing time. The need for models that optimize resource utilization while maintaining high accuracy levels is critical, particularly for applications requiring real-time analysis.

These identified gaps in the literature serve to justify the contributions of the current study, which aims to address these deficiencies through the development of an advanced, specially tailored model for cyberbullying detection. By focusing on improved data quality, enhanced sentiment analysis, multilingual capabilities, and computational efficiency, this research endeavors to fill the critical voids in the field and contribute substantially to the cybersecurity domain.

3 Methodology

3.1 Dataset

AskFm is the main website from which the original data utilized in this work was extracted. It is mostly used for asking questions of other users, either publicly or privately, and subsequently receiving their responses. We used the Python web crawler package Beautiful Soup3 to crawl each user's profile and gather their questions and replies [28, 29]. A CSV file contains the questions and responses related to every user profile. After a string-matching method was used to filter out cyberbullying swear words, question-answer pairings were only extracted if they did. 3720 user profiles and more than 400,000 question-answer pairs were crawled in total.

We used public proxy servers in the USA, UK, and CA to get written posts in English. By using insult/swear word string matching, the data was limited to 10,000 distinct posts that either contained an insult or a swear word in the question or the answer section.

The first step of the suggested approach is the careful gathering of datasets from various online sources. Two main datasets were acquired from Form-spring and Twitter, respectively. The Form-spring dataset required manual labeling using expert data annotation techniques, while the Twitter dataset was already annotated. Fig. 1 below demonstrates the dataset preprocessing flow.

3.2 Data Preprocessing

3.2.1 Data Collection

The first stage of the suggested methodology entails the thorough gathering of datasets from various sources on the Internet. We obtained two data sets in the primary dataset, one from Twitter and another from Form-spring. The Twitter dataset was annotated, while the Form-spring dataset required inter-manual annotation using professional personal identification methods.

3.2.2 Data Labeling

Data labeling, the vital process involved in a machine learning project allocation, was a huge part of the overall timeline, roughly 80%. The team also explored a number of techniques for data labeling, including utilizing internal staff, hiring managed teams, enlisting the services of contractors, and outsourcing to crowd sourced services. This did guarantee the data were concretely annotated for later research.

3.2.3 Data Preprocessing

Hence, to achieve the desired objective of refined feature extraction, proper data preprocessing methods were applied. Not only did the cleaning procedure include getting rid of missing sentences with hyperlinks as well as numerical values and special characters, but also, we ran each article through the spell checker. The data was cleaned with, among other things, converting the letters to lowercase, removing punctuation, eliminating the stop words, disrupting the noise, and using tokens.

1. Cleaning: The cleaning phase took into account the removal of items that were deemed to be supplementary information, like omission (extraneous items such as unfinished sentences and URLs), to ensure the uniformity and consistency of the data.
2. Conversion to Lower Case: Assimilation was performed by making all textual elements lowercase, and doing so, limiting the number of variations in letter cases which could otherwise have had an influence on our subsequent studies.
3. Removal of Punctuation: The dataset for classification reached the stage of cleaning. The punctuation marks were removed to enhance the speed and optimize the script and making only words matter.
4. Removal of Stop Words: Our model became more accurate after removing these frequent stop words, which do not really modify the overall content, like "the" and "is."
5. Noise Removal: Methods such as removing unnecessary letters, links, hooks, and other patterns were used with the objective of not only having a clean dataset but also reducing noise.
6. Tokenization: Tokenization consists of the transformation of the piece of text into a sequence of tokens (words or symbols), in turn making possible further analysis as well as extraction of features.
7. Stemming and Lemmatization: Ranking and lemmatization are major steps of natural language processing, aiming at unifying sentences and words, reducing the effect of changes in grammatical form. This method facilitated to capture of those versions of words that were differently inflected, along with preserving their basic or etymological form.
8. Stemming and Lemmatization in Python: The tools, such as the WordNet lemmatizer and the Snowball stemmer, can be used to transform the sentences to their stems or lemmas in Python. These techniques were the key component in identifying common word patterns and selecting precisely the right words, thereby reducing the variety and variation in the data collection set.

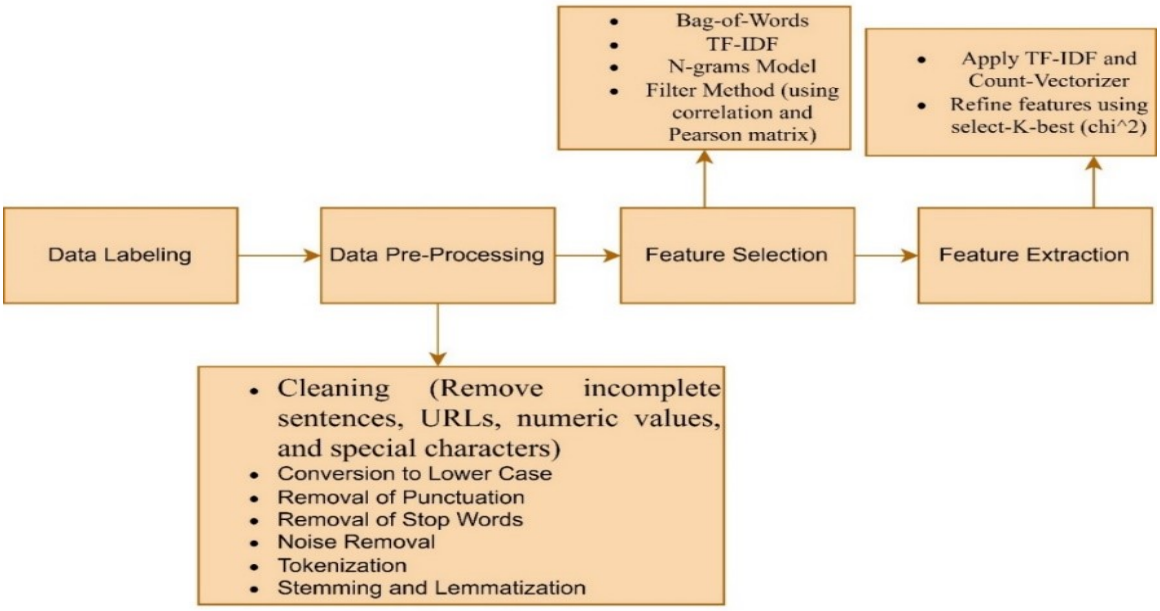

Fig. 1. Dataset preprocessing flow

3.3 Feature Selection

Feature selection remains essential for text classification activities, including cyberbullying detection, because it directly shapes model precision and generalization abilities. Various techniques were used in this study to find the most suitable features that would boost the performance of the BERT model.

A diverse set of features containing bag-of-words and TF-IDF with n-grams was extracted to recognize word frequencies together with their nearby context.

A combination of lexicon-based features to detect explicit abuse used profanity scores coupled with hate speech indicators and threat-level metrics, and sentiment polarity with part-of-speech tag frequencies, helped assess tone and grammatical patterns in the text.

Message length, along with named entity counts and punctuation patterns, contributed to enhancing the feature domain.

3.3.1 Dimensionality Reduction and Feature Optimization

Reducing redundancy along with increasing efficiency demanded the application of Pearson correlation filtering, recursive feature elimination, and L1-based regularization as feature selection methods.

BERT layer attention scores were studied to maintain contextually suitable token selections that bring out semantically dense input sequences.

3.3.2 Final Feature Set and Impact on Model Performance

The final collection of features consisted of offensive language markers, sentiment directional markers, structural signals, and TF-IDF top terms and key n-grams, along with entity counts.

These optimized features yielded superior classification performance mainly for complex classes such as Threat, Hate, and Sexual because of the enhanced metrics observed in the BERT confusion matrix.

3.4 Feature Extraction

The dependent variables were split into two groups: dependent variables from the classical methods and deep learning classification labels. Feature extraction was used to extract relevant information from the preprocessed data. TF-IDF and Count-Vectorizer were used to extract the features, and later improved using the select best k(chi-squared) method.

3.5 Models Training Evaluation

The next phase was the model evaluation. We employed 8 machine learning models, including a voting classifier, random forest (RF), naive Bayes, and support vector machines (SVM), linear SVM, CNN, LSTM, and Bidirectional Encoder Representations from Transformers (BERT) model with the Twitter dataset. The dataset was labeled so that we could evaluate the models' capabilities.

The BERT model, with its novel computational mechanism, turned out to be the best in terms of its characteristics like precision, recall, F1 score, and accuracy. The given technique for identifying cyberbullying has been a systematic process that comprised gathering and grouping information, and later on contributed with well-done, useful data preprocessing and useful features. By utilizing machine learning techniques and combining classical and deep learning methods, the model effectively identified cases of cyberbullying.

The evaluation phase, with a focus on crucial indicators, confirmed the model's effectiveness on the annotated Twitter dataset. Continuous refinement and optimization of preprocessing techniques hold promise for further enhancing the model's accuracy and robustness. After that, several machine learning classifiers are applied to the labeled data sets, discussed above, to train the model.

Cross-verification of the model is done by doing different variations of splitting data parts into test and train parts, like 70% training and 30% test part, 60% training and 40% test part, 50% training and 50% test part, 40% training and 60% test part, and 30% training and 70% is for testing, etc.

There is also a technique named cross-validation used to resolve overfitting. Data labeling is performed on a primary dataset by using standard data annotation techniques. Verification of data labeling is done by using the trained model.

4 Experimentation and Evaluation Metrics

This work used the labeled Twitter dataset to train 8 machine learning models, including a voting classifier, random forest (RF), naive Bayes, support vector machines (SVM), linear SVM, CNN, LSTM, and Bidirectional Encoder Representations from Transformers (BERT).

Based on parameters such as accuracy, F1 score, precision, and recall, BERT was found to be the most successful model for bullying detection. The suggested methodology for the identification of cyberbullying was a methodical one that started with the labeling and gathering of data and proceeded to the careful preprocessing and feature selection of the data, as shown in Fig. 2.

The model showed effectiveness in identifying instances of cyberbullying by utilizing machine learning techniques and combining a combination of classic and deep learning methods. The model's performance on the labeled Twitter dataset was validated throughout the evaluation phase, which focused on important indicators. Preprocessing technique optimization and continual improvement show promise for increasing the accuracy and resilience of the model. Subsequently, the model is trained by using multiple machine learning classifiers on the labeled data sets that were previously mentioned.

The model is cross-verified using many iterations of dividing the data segments into test and train segments, such as 70% training and 30% test, 60% training and 40% test, 50% training and 50% test, 40% training and 60% test, 30% training, and 70% is for testing, etc. To address over-fitting, another method called cross-validation is employed. Standard data annotation techniques are used for primary data collection to achieve data labeling. The trained model is used to verify the data labeling.

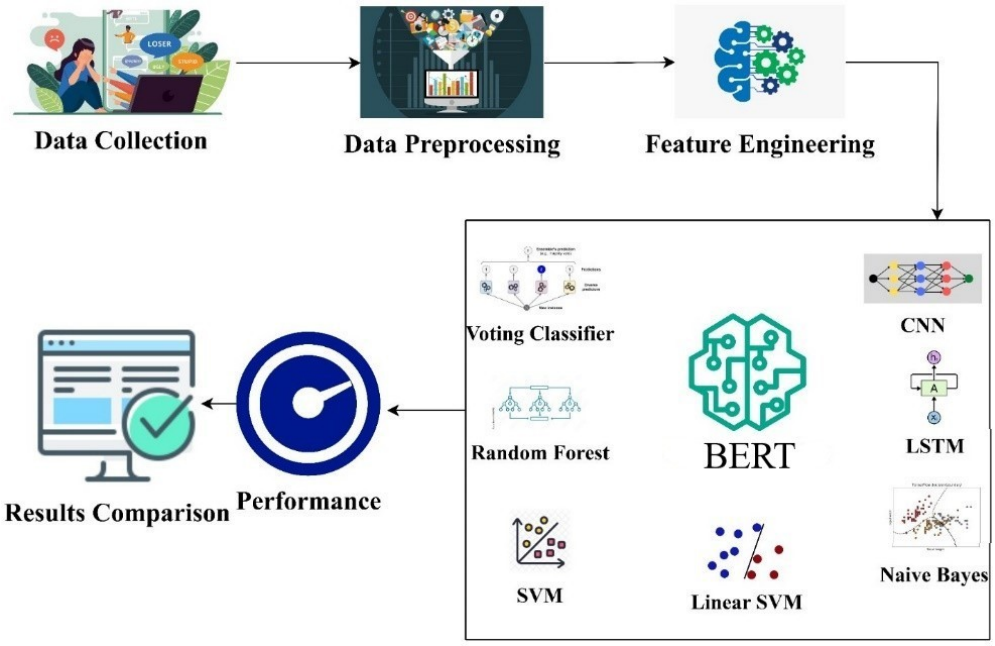

Fig. 2. Research methodology

4.1 Experimentation

4.1.1 Support Vector Machine (SVM)

SVMs, or support vector machines, are effective classifiers that can handle both linear and non-linear data. To divide instances of various class names, the SVM architecture uses a hyperplane to be created in a multidimensional space. SVMs can handle complex relationships in the data because of the choice of kernel function (K(x, x)), which affects the decision boundary form [9]:

$$\sum_{i=1}^{n} (a_i y_i K(x_i, x) + b) . \tag{1}$$

Because SVMs can handle both continuous and categorical variables, they provide a flexible solution for a wide range of learning issues.

The deployment of SVMs was based on their effectiveness in situations where no linear separation can occur because of the features' interactions and involved nonlinearly separable sets [13].

4.1.2 Naive Bayes

The classifier based on the probability-based methods of Naive Bayes is built on Bayes' theorem. Based on the class identification, it assumes that the features are independent. The model is valuable for text categorization as it calculates the probability of a class based on the data:

$$P(c \mid d) = \frac{P(d \mid c) \cdot P(c)}{P(d)}. \tag{2}$$

Naive Bayes was selected due to its ease of use and efficiency while processing textual input. Because of its feature independence assumption, it is especially well-suited for jobs involving text classification. Naive Bayes provides a computationally effective method in cyberbullying detection, when textual features are important.

4.1.3 Random Forest

During training, the Random Forest ensemble learning technique creates many decision trees. Averaging the predictions from each tree reduces overfitting [45]. Every tree in the forest adds

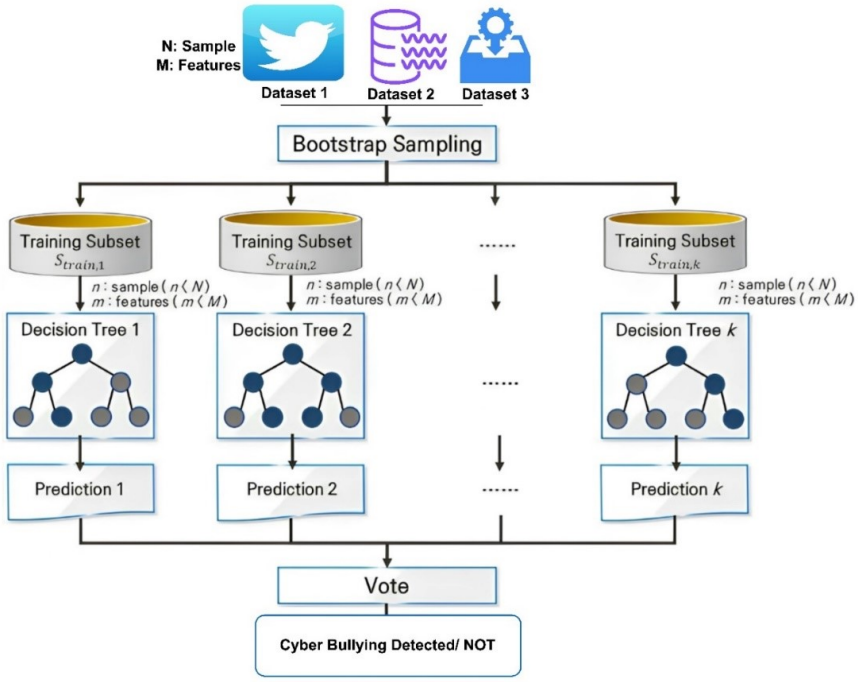

Fig. 3. Random Forest model architecture

something to the ultimate choice, as shown in Fig. 3:

$$y(x,\theta) = \frac{1}{N}\sum_{i=1}^{N} T_i(x,\theta_i). \tag{3}$$

Random Forest was selected due to its capacity to manage big feature sets, reduce overfitting, and offer excellent classification accuracy. When it comes to cyberbullying detection, Random Forest's ensemble technique improves overall model resilience because various variables help find trends [41].

4.1.4 Linear SVM

A variation of SVM that concentrates on linear decision boundaries is called linear SVM. It performs best when instances can be linearly separated into multiple classes [31]. In order to maximize the margin between classes, the hyperplane is developed as shown in Fig. 4:

$$\sum_{i=1}^{N} a_i b_i X_i^T x + b. \tag{4}$$

In situations where a linear decision boundary is suitable, linear SVMs were selected. Simple and effective, linear support vector machines (SVMs) are used when the data is successfully separable in a higher-dimensional space. Linear SVM works best when data patterns are linearly separable [31].

4.1.5 Gaussian Naive Bayes

The Naive Bayes classifier is extended by Gaussian Naive Bayes to handle continuous data. It computes the likelihood of class membership using a Gaussian distribution that is assumed to govern the features [58]:

$$P(X_i \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(X_i-\mu_c)^2}{2\sigma_c^2}\right). \tag{5}$$

Since Gaussian Naive Bayes can handle continuous information, it is the preferred model. This model works well when the characteristics show a Gaussian distribution. Gaussian Naive Bayes provides a viable probabilistic technique for cyberbullying detection, where continuous characteristics could be relevant [58].

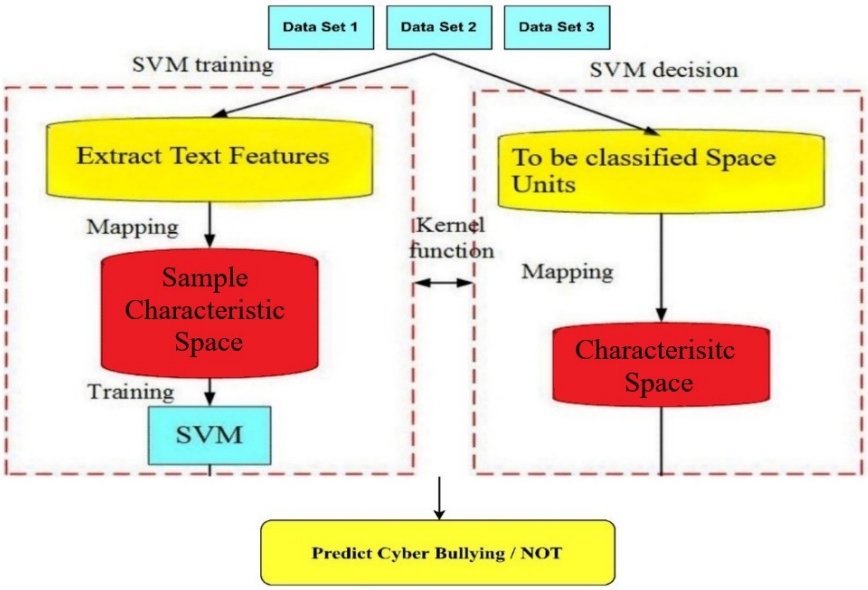

Fig. 4. SVM model architecture

4.1.6 Long Short-Term Memory (LSTM) - Deep Learning

Recurrent neural networks with long short-term memory (LSTM) are specifically engineered to capture long-term dependencies in sequences. It makes use of memory gates and cells to store and forget information selectively [6]:

$$h_t = \tanh\left(W_{ih} \times x_t + b_{ih} + W_{hh} \times h_{t-1} + b_{hh}\right). \tag{6}$$

LSTMs' capability to simulate sequential dependencies renders them a highly advantageous instrument in the analysis of text data, where word order holds significance. When the context of words in a series is critical for cyberbullying detection, LSTM is an effective deep learning technique [6].

4.1.7 Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) is a deep learning model that is created especially for processing data of sequences and images. CNN uses convolutional layers and pooling layers in the processing of text analysis to extract features and to moderate the dimensionality [42]:

$$C(x, w) = \sum_{i=1}^{n}(x_i \times w_i) + b. \tag{7}$$

The choice of CNNs hinges on their capacity to reproduce small spatial features and neighboring effects. While humans may be better able to manually acquire understanding patterns, CNNs perform significantly better at automated tasks of finding hierarchical attributes, which explains their good suitability for cyberbullying identification procedures, where specific patterns of language convey harassment.

4.1.8 Bidirectional Encoder Representations from Transformers (BERT)

The transformer-based deep learning model BERT serves as a sequence-processing model built for handling textual data. BERT operates with a unique bidirectional attention method because it contrasts with traditional unidirectional models to simultaneously process information from forward and backward token sequences in a sentence. Its direction, which follows both forward and

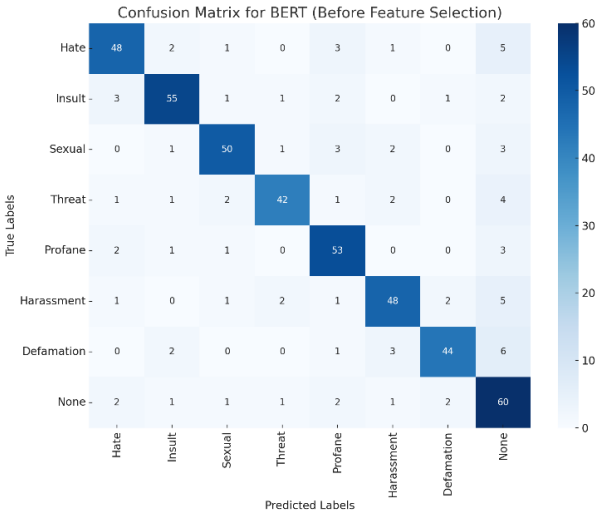

Fig. 5. Confusion matrix best model before features selection

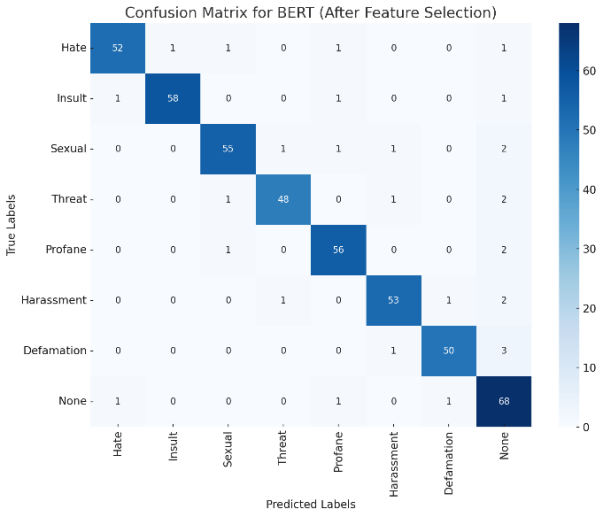

Fig. 6. Confusion matrix best model after feature selection

backward paths, gives BERT exceptional strength for natural language understanding because it detects complex textual patterns [42].

The core BERT operation performs linear transformations and self-attention mechanisms, which update each token representation through an attention-based relation to every token in the sequence. The fundamental operation of self-attention follows this mathematical expression:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (8)$$

Through this equation we observe three matrices of Q, K, V, which derive from input embeddings together with a dimension value of dkd_k. The selection of BERT depends on its superior capability to understand deep contextual relationships and distinct language patterns compared to flat models with sequential architecture design. BERT demonstrates high effectiveness for cyberbullying identification because its ability to recognize linguistic cues within specific contexts helps identify harmful content.

4.2 Evaluation Metrics

Performance evaluation of the eight implemented cyberbullying detection models depended on using accuracy along with precision, recall, and F1-score evaluation metrics. These metrics were used because of their statistical strength, along with their high relevance to detecting harmful language patterns whose incorrect classifications result in serious adverse consequences. The proportion of accurate instances counted abusive and non-abusive tweets among total predictions was evaluated using Accuracy according to Equation 9.

The accuracy metric gave us an effective overview of total system performance during our research because it showed initial benchmark potential. The class imbalance rate in cyberbullying datasets limited its standalone usage for model evaluation in research studies [60, 54, 48, 11].

The mathematical definition of precision described in Equation (10) acted as a vital factor for evaluating model reliability when predicting abusive tweets. The model exhibited excellent capabilities to minimize incorrect flags of non-suspect users, which is essential for real-world Anti-Cyberbullying system implementation.

The measurement of model content retrieval ability through Recall, defined by Equation (11), held equal importance to the model as measured by this metric. When recall values dropped

significantly, the model missed numerous cyberbullying occurrences, which negatively affected its detection practicality. We specifically examined this metric during evaluations of the models against data structures, which demonstrated increased offensiveness.

The F1-score illustrated through Equation (12) as a harmonic mean between precision and recall since it provided balanced performance assessments when system errors involving false positives and negatives were equally crucial. The measure enabled us to detect models with strong performance in both detection accuracy and precision, which served as our primary factor in selecting our final model.

The evaluation metrics were used to measure all eight models throughout the experimental system, including machine learning classifiers alongside deep learning models and hybrid and transformer-based approaches, including BERT. BERT maintained the optimal blend between precision and recall, therefore achieving the best F1-score value. The precision levels for models such as Logistic Regression were high, but these models displayed reduced recall, which implies a protective reaction towards potential abusive language occurrences [11, 46, 36, 35].

We examined the confusion matrix results in Fig. 5 and Fig. 6 from each model to gain additional understanding about classification errors. The detailed analysis validated the effectiveness of contextual embeddings and sequence modeling techniques for reducing both false alarms and missed detections because they enhanced cyberbullying expression detection in various settings. A methodical application of evaluation metrics across numerous datasets and configurations allowed researchers to gain a comprehensive understanding of different models' capabilities regarding the detection of cyberbullying:

$$\text{Accuracy} = \frac{\text{Number of True Positive} + \text{Number of True Negative}}{\text{Total Number of Tweets}}, \quad (9)$$

$$\text{Precision} = \frac{\text{Number of True Positive}}{\text{Number of True Positive} + \text{Number of False Positive}}, \quad (10)$$

$$\text{Recall} = \frac{\text{Number of True Positive}}{\text{Number of True Positive} + \text{Number of False Negative}}, \quad (11)$$

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}. \quad (12)$$

4.3 Results

The performance evaluation of multiple classifiers in the cyberbullying detection domain is presented in this section. Standard evaluation metrics such as accuracy and precision, alongside recall and F1-score, were used for model evaluation before and after performing feature selection. Training time enables performance evaluation by describing the computational requirements.

4.3.1 Testing Results before Feature Selection

The initial tests were conducted without implementing advanced feature selection methods. Most traditional classification techniques displayed poor results when testing their ability to detect cyberbullying occurrences, particularly regarding precise recall measurement. Table 1 below shows the test evaluation results.

SVM and Linear SVM produced insufficient performance results that resulted in the missed detection of 0.05 to 0.10 bullying cases. The BERT model achieved better evaluation metrics than CNN and LSTM before any feature optimization process was applied.

4.3.2 Testing Results after Feature Selection

The implementation of advanced feature selection methods produced substantial performance enhancements throughout all measurement models. The application of feature selection techniques both enhanced the semantic interpretations while removing superfluous elements within the data. The performance metrics of recall and precision experienced substantial positive change in traditional classification techniques.

Table 2 demonstrates the performance metrics that resulted after feature selection was applied. After performing feature selection, BERT achieved outstanding results by reaching 0.989 accuracy and 0.986 F1-score. Both CNN and Random Forest achieved remarkable results after the implementation of feature selection. All classifiers demonstrated improved recognition ability of bullying behavior, which is demonstrated through elevated recall scores in the classification results.

Table 1. Testing results before feature selection

Classifier	Accuracy	Precision	Recall	F1 Score	Training Time (mins)
Gaussian Naive Bayes	0.480	0.450	0.300	0.780	4
SVM	0.700	0.580	0.100	0.100	9
Linear SVM	0.770	0.580	0.050	0.100	6
Voting Classifier	0.850	0.860	0.450	0.600	15
LSTM	0.860	0.790	0.700	0.740	40
Random Forest	0.869	0.660	0.840	0.650	12
CNN	0.916	0.850	0.840	0.840	38
BERT	0.935	0.910	0.900	0.905	48

Table 2. Testing results after feature selection

Classifier	Accuracy	Precision	Recall	F1 Score	Training Time (mins)
Linear SVM	0.770	0.700	0.750	0.710	6
SVM	0.810	0.780	0.710	0.700	9
Gaussian Naive Bayes	0.810	0.790	0.770	0.780	4
Voting Classifier	0.870	0.875	0.760	0.800	15
LSTM	0.880	0.810	0.760	0.780	40
Random Forest	0.910	0.890	0.852	0.890	12
CNN	0.930	0.890	0.880	0.885	38
BERT	0.989	0.985	0.987	0.986	48

Feature selection proves essential for improving classifier sensitivity because it enhances the detection of delicate cyberbullying cues in textual content. The deep bidirectional attention mechanism of BERT allowed it to provide superior predictions than other classifiers during both feature selection and without feature selection since it effectively detects complex linguistic patterns.

The CNN model, with fewer layers than BERT, managed to achieve successful results by specifically identifying repetitive and phrase-based bullying content. The performance of classic machine learning models, Random Forest and Voting Classifier, increased substantially during optimization processes, although their consistency level remained inferior to deep learning techniques.

Training time implementation proves the computational expense-performance relationship that exists between traditional models and transformer-based models. Advanced contextual embeddings combined with optimal feature selection yield evidence to support robust cyberbullying detection system development strategies. The methodology enhances both accuracy levels and successful detection of cyberbullying categories that include insults and threats and identity attacks, as well as sexist remarks, while taking into account subtle and context-sensitive messages.

5 Results and Discussion

A detailed exploration of the eight classifiers used across the cyberbullying detection model presents significant discoveries. The Convolutional Neural Network (CNN) and BERT models displayed peak accuracy along with F1 scores in training, yet CNN proved superior among conventional deep models because of its effective computational performance and quantitative results. BERT proved the most robust model of this study because it achieved both 98.9% testing accuracy and 0.986 F1 score after feature selection.

Between the Voting Classifier and Random Forest models demonstrated consistent performance,

while Random Forest demonstrated superior precision metrics, mainly after applying feature selection. The model demonstrates a strong ability to minimize incorrect detection results that would negatively impact sensitive tasks involving cyberbullying detection.

During testing, the performance of SVM and Linear SVM classifiers deteriorated significantly, resulting in low recall and F1 scores, which may have occurred due to training overfitting. Their evaluation results demonstrated issues with detecting context-specific language patterns that are necessary for detecting genuine cyberbullying cases. The feature selection process did not negatively affect the performance of CNN and LSTM, even though parameters were simplified.

The evaluation metrics of CNN revealed excellent performance between accuracy and recall measurements, although LSTM displayed similar effectiveness but more moderate results than CNN. The Voting Classifier and SVM exhibited inefficiency in dealing with imbalanced or high-dimensional social media text by showing longer training durations and reduced recall effectiveness because of their computational and memory limitations.

The performance of the Gaussian Naive Bayes and Random Forest models substantially improved after conducting feature selection. The dimensionality reduction technique boosts the classification capabilities of traditional models so they can better extract useful information from redundant and noisy features. Feature engineering, together with dimensionality reduction, maintains central importance when implementing classical machine learning approaches.

The study confirms that recall stands as a primary measurement tool because incorrect negations in cyberbullying detection lead to high operational costs. After adjusting features in their models, BERT and CNN reached the best recall scores, which improved their capability to identify real cyberbullying examples.

Some models failed to perform optimally between training and testing phases, thus necessitating additional regularization methods, including dropout, early stopping, or k-fold cross-validation, to combat overfitting. The selection process for models should be determined by application-specific demands since Random Forest proves best at minimizing false positives, but BERT and CNN excel at detecting all true instances. Deployed models should balance performance quality with operational processing load between each other.

However, BERT and comparable models function best for cyberbullying system detection when sufficient computing resources are available for their implementation. CNN presents itself as a highly effective solution in resource-efficient situations. According to the evaluation, the accuracy of the CNN proved optimal at 91.6

6 Conclusion

Recent worldwide surveys indicate that cyberbullying rates have surged because popular social media promotes this behavior, thus 60% of teenagers and 42% of adults now face online harassment. Identifying this behavior remains demanding because of implicit verbalization, along with the usage of sarcasm and the contextual nature of the responses. The research combines BERT with linguistic, lexical, and structural features and manual annotation to improve data reliability while addressing conventional methods' limitations and insufficient labeled data availability.

The examined model showed notable advancement through an in-depth feature selection process. Before adding refinement features, the baseline BERT system reached a 91.4% accuracy level. The BERT model achieved a final prediction accuracy rate of 98.6% when combined with TF-IDF vectors, n-grams, profanity and threat scores, sentiment polarity, NER tags, and punctuation usage. An outstanding impact occurred in difficult class detection, resulting in F1 scores that jumped from 82-85% to 93-95% among Hate, Threat, and Sexual categories. The combination of feature selection strategies alongside preprocessing methods achieves double improvements in classification metrics while maintaining system dependability.

Research on multilingual detection systems remains crucial because most current research focuses on English datasets. The detection

solutions for cyberbullying that exist today only monitor less than 10% of online communication contents that operate in Arabic or Urdu, or Roman Urdu languages.

Future research will target the development of this model to extend linguistic capabilities, which will create culturally appropriate cyberbullying detection across diverse social networks.

The dataset used in this study is available from the corresponding author upon request.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, and grants 20241816, 20241819, and 20240951 of the Secretaria de Investigacion y Posgrado of the Instituto Politecnico Nacional, Mexico. The authors thank CONACYT for the computing resources provided through the Plataforma de Aprendizaje Profundo para Tecnologias del Lenguaje of the Laboratorio de Supercomputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

1. **Abdelsalam, A. M., Sorour, A. S., Abdelaziz, M. M. (2020).** Cyberbullying victimization and its association with self-esteem and emotional intelligence among adolescents: An intervention study. Journal of Nursing Education and Practice, Vol. 10, No. 7, pp. 62–73. DOI: 10.5430/jnep.v10n7p62.
2. **Abdullah, Ather, M. A., Kolesnikova, O., Sidorov, G. (2025).** Detection of biased phrases in the Wiki Neutrality Corpus for Fairer Digital Content Management using Artificial Intelligence. Big Data and Cognitive Computing, Vol. 9, No. 7, pp. 190. DOI: 10.3390/bdcc9070190.
3. **Abdullah, Hafeez, N., Nasir, M. U., Shabbir, M., Mehmood, S., et al. (2024).** Personalized email marketing: A machine learning approach for higher engagement and conversion rates. 2024 Horizons of Information Technology and Engineering (HITE), IEEE, pp. 1–6.
4. **Abdullah, Hafeez, N., Sidorov, G., Gelbukh, A., Oropeza Rodríguez, J. L. (2025).** Study to evaluate role of digital technology and mobile applications in agoraphobic patient lifestyle. Journal of Population Therapeutics and Clinical Pharmacology, Vol. 32, No. 1, pp. 1407–1450. DOI: 10.53555/r6bw9e39.
5. **Abdullah, Hafeez, N., Ullah, F., Ather, M. A., Hasan, A., Gelbukh, A., Oropeza-Rodríguez, J. L., Sidorov, G., Kolesnikova, O. (2025).** Performance tradeoffs in adaptive hybrid encryption and decryption techniques security analysis for optimized protection in IoT-environmental data systems. Contemporary Mathematics, Vol. 6, No. 5, pp. 5407–5442.
6. **Agbaje, M., Afolabi, O. (2024).** Neural network-based cyber-bullying and cyber-aggression detection using Twitter (X) text.
7. **Ali, A., Syed, A. M. (2020).** Cyberbullying detection using Machine Learning. Vol. 3, No. 2, pp. 45–50.
8. **Ali, B. S., Ullah, I., Al Shloul, T., Khan, I. A., Khan, I., Ghadi, Y. Y., Abdusalomov, A., Nasimov, R., Ouahada, K., Hamam, H. (2024).** ICS-IDS: application of big data analysis in AI-based intrusion detection systems to identify cyberattacks in ICS networks. The Journal of Supercomputing, Vol. 80, No. 6, pp. 7876–7905.
9. **Amjad, M., Ashraf, N., Zhila, A., Sidorov, G., Zubiaga, A., Gelbukh, A. (2022).** Threatening language detecting and threatening target identification in Urdu tweets. Preprint.
10. **Amjad, M., Zhila, A., Sidorov, G., Labunets, A., Butta, S., Amjad, H. I., Gelbukh, A. (2022).** Overview of abusive and threatening language detection in Urdu at FIRE 2021. arXiv preprint arXiv:2207.06710.

11. **Aroyehun, S. T., Gelbukh, A. (2018).** Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pp. 90–97.

12. **Ashraf, N., Mustafa, R., Sidorov, G., Gelbukh, A. (2020).** Individual vs. group violent threats classification in online discussions. Companion Proceedings of the Web Conference 2020, Association for Computing Machinery, New York, NY, USA, pp. 629–633. DOI: 10.1145/3366424.3385778.

13. **Berrar, D. (2018).** Bayes' theorem and naive Bayes classifier. Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics, Vol. 403, No. 412.

14. **Biagioni, S., Baroni, M., Melis, F., Baldini, F., Menicucci, D., Potente, R., Molinaro, S. (2023).** Cyberbullying roles and the use of psychoactive substances: A systematic review. Adolescent research review, Vol. 8, No. 4, pp. 423–455.

15. **Bokolo, B. G., Liu, Q. (2023).** Combating cyberbullying in various digital media using machine learning. In Combatting Cyberbullying in Digital Media with Artificial Intelligence. Chapman and Hall/CRC, pp. 71–97.

16. **Bovill, H. (2023).** Too much information: Exploring technology-mediated abuse in Higher Education Online Learning and Teaching Spaces Resulting from COVID-19 and Emergency Remote Education. Higher education, Vol. 86, No. 2, pp. 467–483.

17. **Butt, S., Amjad, M., Balouchzahi, F., Ashraf, N., Sharma, R., Sidorov, G., Gelbukh, A. F. (2022).** Overview of EmoThreat: Emotions and threat detection in Urdu at FIRE 2022. Proceedings of FIRE 2022 (Working Notes), pp. 220–230.

18. **Camacho-Vázquez, L. A., Camacho-Vázquez, V. A., Orantes-Jiménez, S. D., Sidorov, G. (2025).** Detection of negative emotions in short texts using deep neural networks. Cyberpsychology, Behavior, and Social Networking.

19. **Camacho-Vázquez, V. A., Sidorov, G., Galicia-Haro, S. N. (2018).** Automatic detection of negative emotions within a balanced corpus of informal short texts. Cyberpsychology, Behavior, and Social Networking, Vol. 21, No. 12, pp. 781–787. DOI: 10.1089/cyber.2018.0185.

20. **Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A. (2017).** Mean birds: Detecting aggression and bullying on twitter. Proceedings of the 2017 ACM on web science conference, pp. 13–22.

21. **Chen, Y., Zhou, Y., Zhu, S., Xu, H. (2012).** Detecting offensive language in social media to protect adolescent online safety. , pp. 71–80.

22. **Daoud, W. B., Othmen, S., Hamdi, M., Khdhir, R., Hamam, H. (2023).** Fog computing network security based on resources management. EURASIP Journal on Wireless Communications and Networking, Vol. 2023, No. 1, pp. 50.

23. **Dewani, A., Memon, M. A., Bhatti, S. (2021).** Cyberbullying detection: advanced preprocessing techniques & Deep Learning architecture for Roman Urdu data. Journal of Big Data, Vol. 8, No. 1, pp. 160. DOI: 10.1186/s40537-021-00526-x.

24. **Gelbukh, A., Zamir, M. T., Ullah, F., Ali, M., Taiba, T., Usman, M., Hafeez, N., Dudaeva, L., Fasoldt, C. (2024).** State-of-the-art review in explainable machine learning for smart-cities applications. Springer, pp. 67–76.

25. **Ghadi, Y. Y., Shah, S. F. A., Mazhar, T., Shahzad, T., Ouahada, K., Hamam, H. (2024).** Enhancing patient healthcare with mobile edge computing and 5G: Challenges and solutions for secure online health tools. Journal of Cloud Computing, Vol. 13, No. 1, pp. 93.

26. **Haq, I. U., Zarkoon, S. M. (2023).** Cyber stalking: A critical analysis of prevention of

electronic crimes Act-2016 and its effectiveness in combating cyber crimes, A perspective from Pakistan. Pakistan's Multidisciplinary Journal for Arts & Science, pp. 43–62.

27. **Haque, R., Islam, N., Tasneem, M., Das, A. K. (2023).** Multi-class sentiment classification on Bengali social media comments using machine learning. International journal of cognitive computing in engineering, Vol. 4, pp. 21–35.

28. **Hosseinmardi, H., Ghasemianlangroodi, A., Han, R., Lv, Q., Mishra, S. (2014).** Towards understanding cyberbullying behavior in a semi-anonymous social network. 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014), IEEE, pp. 244–252.

29. **Hosseinmardi, H., Li, S., Yang, Z., Lv, Q., Rafiq, R. I., Han, R., Mishra, S. (2014).** A comparison of common users across Instagram and Ask.fm to better understand cyberbullying. 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, IEEE, pp. 355–362.

30. **Iparraguirre-Villanueva, O., Alvarez-Risco, A., Herrera Salazar, J. L., Beltozar-Clemente, S., Zapata-Paulini, J., Yáñez, J. A., Cabanillas-Carbonell, M. (2023).** The public health contribution of sentiment analysis of Monkeypox Tweets to detect polarities using the CNN-LSTM model. Vaccines, Vol. 11, No. 2, pp. 312.

31. **Jahromi, A. H., Taheri, M. (2017).** A non-parametric mixture of Gaussian Naive Bayes classifiers based on local independent features. 2017 Artificial intelligence and signal processing conference (AISP), IEEE, pp. 209–212.

32. **Jain, P. K., Quamer, W., Saravanan, V., Pamula, R. (2023).** Employing BERT-DCNN with sentic knowledge base for social media sentiment analysis. Journal of Ambient Intelligence and Humanized Computing, Vol. 14, No. 8, pp. 10417–10429.

33. **Kaur, M., Saini, M. (2023).** Indian Government initiatives on cyberbullying: A case study on cyberbullying in Indian Higher Education Institutions. Education and Information Technologies, Vol. 28, No. 1, pp. 581–615.

34. **Ktari, J., frikha, T., Hamdi, M., Affes, N., Hamam, H. (2025).** Enhancing blockchain security and efficiency through FPGA-based consensus mechanisms and post-quantum cryptography. Recent Advances in Electrical & Electronic Engineering, Vol. 18, No. 7, pp. 946–958.

35. **Maitra, P., Sarkhel, R. (2018).** A k-competitive autoencoder for aggression detection in social media text. Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pp. 80–89.

36. **Maity, K., Jha, P., Jain, R., Saha, S., Bhattacharyya, P. (2023).** "explain thyself bully": Sentiment aided cyberbullying detection with explanation. International Conference on Document Analysis and Recognition, Springer, pp. 132–148.

37. **Medina Nieto, M. A., de la Calleja Mora, J., López Domínguez, E., Hernández Velázquez, Y., Arrieta Díaz, D. (2025).** Semantic MOCIBA 2021: A vocabulary for cyberbullying based on open data analysis. Computación y Sistemas, Vol. 29, No. 1, pp. 409–422.

38. **Muhammad, K. B., Burney, S. A. (2023).** Innovations in Urdu sentiment analysis using machine and deep learning techniques for two-class classification of symmetric datasets. Symmetry, Vol. 15, No. 5, pp. 1027.

39. **Murnion, S., Buchanan, W. J., Smales, A., Russell, G. (2018).** Machine learning and semantic analysis of in-game chat for cyberbullying. Computers & Security, Vol. 76, pp. 197–213.

40. **Oladepo, T., Abiola, O., Abiola, T., Abdullah, Muhammad, U., Abiola, B. (2025).** Predicting Emotion Intensity in Text Using Transformer-Based Models. Proceedings of the 19th International Workshop on Semantic

Evaluation (SemEval-2025), Association for Computational Linguistics, Vienna, Austria, pp. 1677–1682.

41. **Pal, S., Singha, P. (2023).** Linking trophic state with the eco-hydrological state of dam-induced floodplain wetland in Barind Tract. Arabian Journal of Geosciences, Vol. 16, No. 4, pp. 246.
42. **Powers, D. M. (2020).** Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.
43. **Rahman, H., Tariq, J., Masood, M. A., Subahi, A. F., Khalaf, O. I., Alotaibi, Y. (2023).** Multi-tier sentiment analysis of social media text using supervised machine learning. Comput. Mater. Contin, Vol. 74, No. 3, pp. 5527–5543.
44. **Rahman-Laskar, S., Gupta, G., Badhani, R., Pinto-Avendaño, D. E. (2024).** Cyberbullying detection in a multi-classification codemixed dataset. Computación y Sistemas, Vol. 28, No. 3, pp. 1091–1113.
45. **Sainz, V., Martín-Moya, B. (2023).** The importance of prevention programs to reduce bullying: A comparative study. Frontiers in psychology, Vol. 13, pp. 1066358.
46. **Salminen, J., Almerekhi, H., Milenković, M., Jung, S.-g., An, J., Kwak, H., Jansen, B. (2018).** Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. Proceedings of the International AAAI Conference on Web and Social Media, Vol. 12, No. 1.
47. **Saranya, S., Usha, G. (2023).** A machine learning-based technique with intelligentwordnet lemmatize for twitter sentiment analysis. Intelligent Automation & Soft Computing, Vol. 36, No. 1.
48. **Schaffer, C. (1993).** Selecting a classification method by cross-validation. Machine Learning, Vol. 13, No. 1, pp. 135–143. DOI: 10.1007/BF00993106.
49. **Shaukat, K., Parveen, Q., Dahar, M. A., Ehsan, T. (2023).** A study of social and psychological factors related to bullying victimization at elementary level. Russian Law Journal, Vol. 11, No. 3, pp. 3206–3221.
50. **Shushkevich, E., Cardiff, J., Rosso, P., Akhtyamova, L. (2020).** Offensive language recognition in social media. Computación y Sistemas, Vol. 24, No. 2, pp. 523–532.
51. **Siddiqui, S., Schultze-Krumbholz, A. (2023).** Bullying prevalence in Pakistan's educational institutes: Preclusion to the framework for a teacher-led antibullying intervention. PLoS one, Vol. 18, No. 4, pp. e0284864.
52. **Siddiqui, S., Schultze-Krumbholz, A. (2023).** Successful and emerging cyberbullying prevention programs: A narrative review of seventeen interventions applied worldwide. Societies, Vol. 13, No. 9, pp. 212.
53. **Sidorov, G., Gelbukh, A. (2001).** Automatic detection of semantically primitive words using their reachability in an explanatory dictionary. Proceedings of the 2001 IEEE International Conference on Systems, Man and Cybernetics (ICSMC), IEEE, Tucson, AZ, USA, pp. 1683–1687 vol.3. DOI: 10.1109/ICSMC.2001.973527.
54. **Tian, Y., Zhang, Y. (2022).** A comprehensive survey on regularization strategies in Machine Learning. Information Fusion, Vol. 80, pp. 146–166.
55. **Ullah, F., Gelbukh, A., Zamir, M. T., Felipe-Riverón, E. M., Sidorov, G. (2024).** Enhancement of named entity recognition in low-resource languages with data augmentation and BERT models: A case study on Urdu. Computers, Vol. 13, No. 10, pp. 258.
56. **Ullah, S., Ilyas, M., Ullah, I.**, **.** Role of social media in perpetuating bullying in Pakistan. .
57. **Wang, H., Bragg, F., Guan, Y., Zhong, J., Li, N., Yu, M. (2023).** Association of bullying victimization with suicidal ideation and suicide attempt among school students:

A school-based study in Zhejiang province, China. Journal of affective disorders, Vol. 323, pp. 361–367.

58. **Xu, L., Yan, Y., Huang, X. (2022).** Deep Learning in Solar Astronomy. Springer.
59. **Yi, P., Zubiaga, A. (2023).** Session-based cyberbullying detection in social media: A survey. Online Social Networks and Media, Vol. 36, pp. 100250. DOI: 10.1016/j.osnem.2023.100250.
60. **Ying, X. (2019).** An overview of overfitting and its solutions. Vol. 1168, pp. 022022.
61. **Zamir, M. T., Ullah, F., Tariq, R., Bangyal, W. H., Arif, M., Gelbukh, A. (2024).** Machine and deep learning algorithms for sentiment analysis during COVID-19: A vision to create fake news resistant society. PloS one, Vol. 19, No. 12, pp. e0315407.

Article received on 20/02/2025; accepted on 20/05/2025.
**Corresponding author is Irfan Latif.*