www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(4), Sep-Dec 2025, 497-504. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2025.v16i4.589

Analysis of Cyberbullying using Bernoulli Restricted Boltzmann Machine and Multilayer Perceptron Neural Networks

Ana Laura Lezama-Sánchez[1], *Mireya Tovar Vidal*[2]
[1] Escuela de Artes Plasticas y Audiovisuales, Benemerita Universidad Autonoma de Puebla, Puebla, Mexico
[2] Faculty of Computer Science, Benemerita Universidad Autonoma de Puebla, Puebla, Mexico
analaura.lezama@correo.buap.mx, mireya.tovar@correo.buap.mx

Abstract. This article proposes a model based on natural language processing and deep learning for detecting cyberbullying in Spanish texts from social media. Since this phenomenon has become an alarming problem in recent years, particularly in Spanish-speaking communities, automatic detection has become crucial to address this challenge. In this context, the results obtained indicate that the model not only correctly identifies bullying cases but also minimizes false positives.
Keywords: deep learning, detection cyberbullying, Boltzmann

Article Info
Received November 29, 2024
Accepted December 4, 2024

1 Introduction

In the last decade, the exponential growth of social media has transformed the way we communicate and interact. Children, adolescents and adults use these platforms to connect, share and express their everyday thoughts and activities. However, the use, in many cases poorly supervised, has led to an alarming phenomenon: cyberbullying. This type of harassment can affect both mental health and physical well-being, becoming a problem that affects people of all ages. However, cyberbullying is not the only type of abuse that occurs on social media; there are also other harmful behaviors in different environments, such as school and workplace harassment. In addition, a new form of harassment has emerged: hate speech, along with verbal abuse, insults and toxicity, which affect people in various aspects of their lives.

Cyberbullying poses significant risks, as it can reach people anytime and anywhere, without geographical boundaries. In addition, the ease with which information is disseminated online amplifies the speed and scope of harassment, making it even more damaging to victims. Anyone can be susceptible to cyberbullying and become a target of online intimidation and harassment. The consequences of cyberbullying are diverse and severe. Psychologically, it can lead to anxiety, depression, low self-esteem and even suicidal thoughts in the victims. Socially, it often results in isolation and difficulties in establishing or maintaining interpersonal relationships.

The implementation of natural language processing (NLP), machine learning (ML) and deep learning (DL) techniques has significantly advanced in the detection and analysis of cyber-bullying. These approaches have allowed the development of computational models capable of identifying this behavior in different data sets, including those in Spanish and English Afrifa et al., (2022).

The effectiveness of classification models is assessed using metrics such as accuracy, precision, recall and F_1, which provide objective measures of model performance. Accuracy indicates the proportion of messages correctly classified as bullying among all messages identified as such. Accuracy represents the overall proportion of messages that are correctly classified, whether or not they are intimidating. Recovery measures the proportion of successfully detected bullying messages within the total number of actual harassment messages in the data set. Finally, the F_1 combines precision and recall, offering a balanced assessment of the model's performance by counting both false positives and false negatives.

Therefore, this work proposes a model that classifies a dataset gathered from various social media platforms, labeled as "bullying" or "non-bullying," develop by Ejaz et al., (2024). The model is based on advanced techniques, such as the Bernoulli Restricted

Boltzmann Machine and Multilayer Neural Networks, which enable precise and efficient classification of these harmful behaviors. Through this research, we aim to play a role in developing tools for fostering a more secure and healthier online setting.

The rest of the article is organized as follows: Section 2 reviews some of the existing works on cyberbullying detection. Section 3 describes the methodology employed in this work, while Section 4 presents the results obtained by applying the proposed methodology. The conclusions, as well as future works, are discussed in Section 5. Finally, the references consulted during the development of this work are listed.

2 Related Works

In this section, works in the same field are presented. Some authors have incorporated cyberbullying detection into their methodologies, addressing specific aspects such as sarcasm, statistical techniques for analyzing data from social media, linear regression, sentiment analysis, among others. In some cases, not only text analysis techniques are used, but also methods for analyzing images and/or PDF documents.

In contrast, the study carried out by Recalde (2021) examines cases of cyberbullying in Ecuador over the past five years are analyzed, using statistical techniques to examine data from social media. The main objective was to identify deficiencies in the detection capabilities of regulatory authorities. The authors found a significant increase in reports related to privacy violations and juvenile pornography. Additionally, they projected a 70% rise in cyberbullying cases by 2025. Based on these findings, the authors suggested the need for more effective measures on social media platforms, such as algorithms, architectures, software, or processes to address these risks. The data used for the analysis came from the Public Prosecutor's Office, based on public access to information laws, and linear regression was applied to forecast future reports. The trend showed an average annual increase of 13.45%.

On the other hand, Llampa et al., (2020) addresses the gap in NLP between English and Spanish. The authors aimed to develop an effective automated model in Spanish for sentiment analysis of YouTube video comments. To this end, they propose an approach that allows for the effective detection of verbal abuse, identifying both the victims and the perpetrators, thus facilitating informed decision-making. This analysis combines sentiment analysis with a supervised learning algorithm, specifically SVM. The process includes several stages of text processing adapted to Spanish, as well as the selection and training of the SVM model. As a programming language, they used Python along with libraries such as spaCy, scikit-learn, and matplotlib for development, data analysis, and visualization. The YouTube comments were obtained through Google's YouTube Data API. The authors emphasize the importance of transforming textual data into numerical representations, a process known as feature extraction. In this case, they used word vectorization (Word Embeddings) to convert each comment into a vector compatible with machine learning models.

In Sultan et al., (2023), a study is proposed that provides a useful framework for detecting cyberbullying in screenshots. The study employs OCR, NLP, and machine learning to detect cyberbullying in images. The results show that by applying the linear Support Vector Classifier (SVC) after OCR and combining it with logistic regression, the highest accuracy is achieved, reaching 96%. The dataset used was extracted from Kaggle and is related to cyberbullying and toxicity. This dataset collects various relevant data for the automated detection of cyberbullying, sourced from a wide range of platforms, including social media, Kaggle, Twitter, Wikipedia talk pages, and YouTube.

The work presented by Mamani (2020) aims to use Artificial Intelligence to identify conflictive conversations that affect children and young people around the world. However, one challenge the authors encountered was the evolution of linguistic complexities, such as slang and jargon, used to perpetuate virtual violence. Additionally, they point out that the mood of the message sender influences the semantic meaning of the texts. Therefore, the authors considered analyzing the context of the texts, even when they are written in a confusing manner. As a result, they employed procedures for natural language processing and appropriately processed the information for analysis, training a Recurrent Neural Network (RNN) to better understand the forms of written expression used by individuals in their communication.

The research conducted by Al-Garadi et al., (2019) provides a thorough review of the process of detecting cyberbullying on social media. The authors performed a literature survey to identify aggressive behaviors on social networking sites using machine learning techniques. Their analysis focused on four main areas: data collection, feature engineering, the development of the cyberbullying detection model, and the assessment of the models created. Furthermore, they provided a summary of the various types of discriminative features used in detecting online cyberbullying and assessed the most effective supervised machine learning classifiers for this task. A notable contribution of the study was the establishment of specific evaluation metrics to evaluate model performance, which allowed the authors to identify key parameters and compare different machine learning algorithms.

The work by Altowairgi et al., (2023), focuses on the development of models to detect cyberbullying, utilizing algorithms such as SVM, Naïve Bayes, and Random Forests. The authors worked with a cyberbullying dataset from the Mendeley website, which covers various forms of cyberbullying, including aggression, racism, insults, and toxicity. To represent the ideational metafunction, Latent Dirichlet Allocation (LDA) was applied as the primary feature. The models were evaluated using metrics such as precision, recall, and F_1. The findings demonstrated that the algorithms used were effective at detecting cyberbullying, achieving 92% precision for Naïve Bayes and 93% accuracy for both SVM and Random Forests.

On the other hand, the research proposed by Mahmud et al., (2024), presents a study on the detection of cyberbullying in the Bangla and Chittagonian languages, using machine learning and deep learning techniques. The authors collected more than 5,000 text samples from social media and evaluated the reliability of annotations using Krippendorff's alpha coefficients and Cohen's Kappa. They applied various machine learning methods, the best one being the SVM model with an accuracy of 0.711. In contrast, convolutional neural networks (CNN) outperformed these results, reaching precisions between 0.69 and 0.811. The authors also conducted experiments with models such as BiLSTM-GRU and CNN-LSTM, with the model CNN-LSTM-BiLSTM achieving an accuracy of 0.82. In addition, transformer-based models such as XLM-Roberta and Bangla BERT showed significant improvements with accuracies of 0.841 and 0.822 respectively.

In the study by Chen et al., (2024), the authors develop machine learning and deep learning models for the Chinese language. Additionally, they propose a model based on BERT, XLNet, and deep Bi-LSTM for detecting cyberbullying in Chinese. They also collected real cyberbullying comments to enrich the Chinese offensive language dataset, known as COLDATASET. The performance of the model surpassed all reference models in this dataset, achieving a 4.29% improvement over the highest-performing SVM, a 1.49% improvement over the best-performing deep learning mode, and a 1.13% improvement compared to BERT.

In the study by Alqahtani et al., (2024), they present an approach to improve a system for detecting different types of cyberbullying. The authors used multiclass classification algorithms and, for feature extraction, employed TF-IDF. The results of their experiment showed superior performance compared to previous studies that used the same dataset. Additionally, they implemented two machine learning algorithms by combining N-grams with TF-IDF features, which resulted in good performance in classification tasks. The experimental results demonstrated that the proposed approach can effectively detect different types of cyberbullying, with a precision of 0.9071.

In the study by Dewani et al., (2023), the authors developed a method to detect cyberbullying in Roman Urdu. The approach was based on statistical feature extraction techniques, N-gram extraction and TF-IDF weighting using the SVM classification model. The results obtained were evaluated using GridSearchCV and cross-validation. By applying N-grams, an accuracy of 83% was achieved, whereas by using statistical characteristics, an accuracy of 79% was achieved. The authors also mention that they categorized the severity of cyberbullying, classifying it into implicit and explicit cases.

In Ali et al., (2020), a machine learning-based method is introduced to detect various forms of cyberbullying, including text, images, and PDF documents. The approach utilizes machine learning techniques such as Support Vector Machine (SVM), k-nearest neighbors, Decision Trees, and Random Forest, along with image recognition algorithms, Optical Character Recognition (OCR), and NLP methods. The system extracts feature like the use of offensive language, insults, and a hostile tone in the text, as well as other factors such as sentiment, grammar, and semantics. The proposed system supports decision-making by flagging content for further review and action by a human when cyberbullying is detected.

This paper presents a model to detect whether a text from social media is intended to harass its recipient. To achieve this, a dataset extracted from Kaggle was used, which comes from platforms such as Wikipedia, Twitter and YouTube. The data were preprocessed by removing HTML tags, URLs, punctuation marks, and performing tokenization and lemmatization. As a result of this process, a single clean CSV file was generated containing the preprocessed texts. Next, text vectorization techniques using TF-IDF (Term Frequency-Inverse Document Frequency) were applied to obtain numerical representations of the texts. Subsequently, to create a model that predicts whether a text corresponds to harassment or not, a Bernoulli Restricted Boltzmann Machine (RBM) was used to reduce the dimensionality of the data, and then a classifier based on a Multilayer Perceptron (MLP) neural network was trained. The model was evaluated using performance metrics such as precision, recall, accuracy, and F_1. The results obtained were accuracy and recall of 0.9094, precision of 0.9012, and F_1 of 0.8961.

3 Proposed Approach

This paper presents a model to detect whether a text from social media is intended to harass its recipient or not.

The architecture of the proposed model is presented in Figure 1 and includes the following stages: the dataset collection, detailed in Table 1, and the preprocessing of the texts to standardize and prepare them for analysis. In this phase, cleaning techniques such as the removal of HTML tags, URLs, and punctuation are applied, along with tokenization and lemmatization of the texts. Subsequently, the TF-IDF vectorization technique is used to represent the texts in numerical format, which facilitates computational analysis. For text classification, a Bernoulli RBM and a MLP neural network are employed. Finally, the results obtained are evaluated using standard metrics such as accuracy, recall, precision, and F_1.

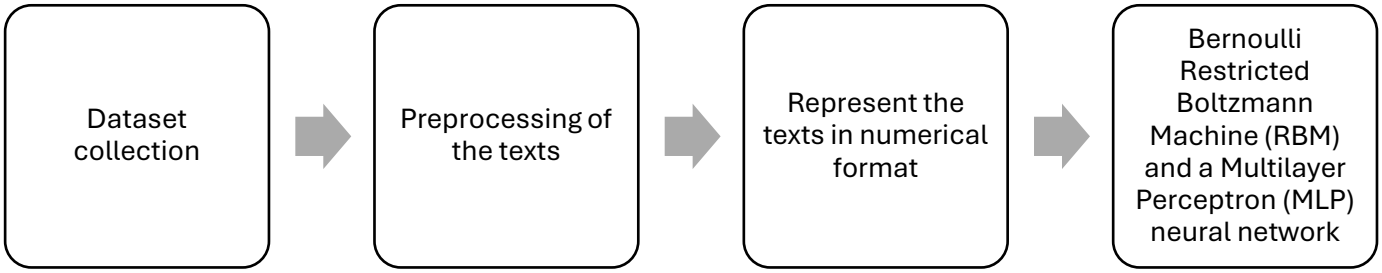

Fig 1. Proposed Architecture

The proposed approach consists of 4 phases, which are made up of the following tasks:

Data Collection: In this phase, the dataset collected by Ejaz et al., (2024) was used. This dataset consists of texts related to cyberbullying, extracted from Twitter and YouTube, and is available in CSV files. The dataset was obtained from the Kaggle database (see Table 1). This dataset aggregates information from various sources associated with the automated identification of cyberbullying. The data comes from several social media platforms, such as Kaggle, Twitter, Wikipedia Talk pages, and YouTube. The texts in the data are classified as bullying or non-bullying and include six different forms of cyberbullying.
This data set provides a rich and diverse representation of cyberbullying texts, which enables effective training and evaluation of detection models. The variety of sources and types of bullying make the dataset suitable for developing robust classification models in this field.

Table 1. Dataset

Type of cyberbullying	Number of messages
Aggression	115,863
Attack	115,863
Insults	8,798
Toxicity	159,685
Sexism, racism	16,852
Racism	13,472
Sexism	14,882
Hate	3,465

Text Preprocessing: In this phase, the text preprocessing is carried out. The specific tasks include:

1. HTML Tag Removal: HTML tags present in the text are removed using regular expressions.
2. URL Removal: Web links (URLs) appearing in the text are removed.
3. Text Normalization: The entire text is converted to lowercase to ensure uniformity
4. Removal of Unwanted Characters: Punctuation marks and other special characters are removed, leaving only letters and spaces.

5. Tokenization and Lemmatization: The text is segmented into tokens (words) and lemmatized, meaning it is transformed into its base form (for example, "run" instead of "running").
6. Stopword Removal: Irrelevant words for analysis (such as "and", "the", "of", etc.) are removed.

At the end of this phase, a clean CSV file containing the preprocessed texts is generated.

Represent the Texts in Numerical Format: In this stage, TF-IDF is used to convert the texts into numerical representations. The existing Python function TfidfVectorizer is employed to create a feature matrix, where each row represents a document (text) and each column represents a feature (word) with its weight calculated through TF-IDF. This process transforms the texts into numerical vectors that can be processed by machine learning models (Azeez et al, 2024).

The process applied to include the TF-IDF technique is described below:

TF-IDF (Term Frequency - Inverse Document Frequency) is a technique used to convert texts into numerical representations that reflect the importance of each word in the context of a set of documents. The calculation of TF-IDF is done in two steps:

- TF (Term Frequency): Measures the frequency of a term in a document relative to the total number of terms in that document. It is calculated using the formula:

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d}$$

- IDF (Inverse Document Frequency): Measures the importance of a term across the entire set of documents. The rarer a term is across documents, the higher its weight. It is calculated with the formula:

$$IDF(t)\frac{N}{Numbers\ of\ documents\ containing\ term\ t}$$

  where N is the total number of documents in the dataset.

- TF-IDF: The final representation of a term in a document is the product of the two previous measures:

$$TF - IDF(t,d) = TF(t,d)\ x\ IDF(t)$$

  This value reflects the relevance of a word in a document relative to the entire corpus, helping to identify the most significant and distinctive words. This vectorized representation is used to transform texts into numerical vectors, which can then be processed by machine learning models.

Bernoulli Restricted Boltzmann Machine and Multilayer Perceptron Neural Network

In this phase, the model is trained for text classification**:**

1. Dimensionality Reduction with RBM: A Bernoulli RBM is used to reduce the dimensionality of the data, which improves the model's efficiency and accuracy by eliminating irrelevant or redundant features.
    a. Number of hidden layers: A single hidden layer was used in the RBM to reduce the data's dimensionality.
    b. Number of visible units: This corresponds to the number of features or words extracted from the texts using TF-IDF.
    c. Number of hidden units: It determines how many latent representations are used to compress the information. 256 configurations were experimented with to optimize accuracy and computational efficiency.
    d. Learning rate: An adaptive learning rate of 0.01 was used to adjust the weights during the training of the RBM.
    e. Mini-batch size: A mini-batch size of 64 was used to balance training time and convergence.
2. Classification with MLP: Next, a classifier based on a MLP neural network is trained, which is used to predict whether a text corresponds to cyberbullying or not.
    a. Number of hidden layers: The MLP was configured with two hidden layers.
    b. Activation function: The activation function used was ReLU (Rectified Linear Unit), which has proven effective for text classification tasks.
    c. Number of neurons in the output layer: For binary classification (cyberbullying or not), a single neuron was used in the output layer with a sigmoid activation function.
    d. Learning rate: A learning rate of 0.001 was used to train the model, which allowed for adequate convergence without overfitting.
    e. Number of epochs: The model was trained for 50 epochs, with cross-validation at the end of each epoch to prevent overfitting.
    f. Regularization: Dropout (with a rate of 0.5) was used to prevent overfitting during the training of the neural network.

g. During this phase, the RBM is first applied to reduce dimensionality and improve the efficiency of the MLP, which then classifies the texts as cyberbullying or not. The model was trained using 70% of the data for training, and final predictions were made on the remaining 30% of the test dataset.

Evaluation and Results

The performance of the model is assessed using standard classification metrics, including precision, recall, accuracy, and F1. These metrics are used to evaluate how effectively the model classifies texts as either cyberbullying or non-cyberbullying.

Accuracy measures (see equation 1) the percentage of correct predictions relative to the total number of predictions made. It is a simple metric, but it is not always representative of imbalanced problems, as it can give a high value even if the model is not correctly classifying the minority classes (López et. al, 2016).

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} = \frac{TP + TN}{TP + TN + FP + FN} \quad (\mathbf{1})$$

TP (True Positives): Correct predictions of the positive class.
TN (True Negatives): Correct predictions of the negative class.
FP (False Positives): Incorrect predictions where the model predicted positive when the class was negative.
FN (False Negatives): Incorrect predictions where the model predicted negative when the class was positive.

Recall (see equation 2) measures the model's ability to correctly identify all instances of the positive class. That is, of all the samples that truly belong to the positive class, what percentage was correctly identified as such? A high recall indicates that the model is efficient at identifying the positive class, but it may also have more false positives (incorrect predictions where the model labeled samples as positive when they were not) (Sultan et al., 2023).

$$Recall = \frac{TP}{TP + FN} \quad (\mathbf{2})$$

Precision is the proportion of true positives among all positive predictions (see equation 3)

$$Precision = \frac{TP}{TP + FP} \quad (\mathbf{3})$$

The F1 (see equation 4) is the harmonic mean of precision and recall. It is especially useful when there is an imbalance between classes, as it takes both false positives and false negatives into account. The F1 is a useful metric when a balance between recall and precision is needed. If either of them is low, the F1 will also be low, indicating that the model is not performing well in terms of both metrics (Mercado et al., 2018).

$$\mathbf{F_1} = 2x\frac{Precision\ x\ recall}{Precision + recall} \quad (\mathbf{4})$$

4 Results and Discussion

In this section, the results obtained in the classification of messages as harassment, non-harassment, or neutral topics are presented.

In Table 2 presents the results obtained from the corpus classification using the proposed model with dimensionality reduction (RBM). The results were evaluated using metrics such as accuracy, precision, recall, and F1. These metrics were calculated using a weighted average, considering the distribution of the classes in the dataset. Specifically, the results obtained were an accuracy of 0.9094, a precision of 0.9012, a recall of 0.9094, and an F1 of 0.8961. These values reflect a solid and consistent model performance, suggesting that the classification of bullying behavior on social media.

Table 2 Results obtained with the proposed model with RBM

Metrics	Results
Precision	0.9012
Recall	0.9094
Accuracy	0.9094
F_1	0.8961

In contrast, in Table 3, the results obtained without dimensionality reduction (RBM) are shown. The model performed better without RBM, achieved an accuracy of 0.9491, a precision of 0.9474, a recall of 0.9491, and an F_1 of 0.9479. These results suggest that the model without RBM performs better than the version with RBM, highlighting a greater ability to classify bullying behavior in texts.

Table 3 Results obtained without RBM

Metrics	Results
Precision	0.9474
Recall	0.9491
Accuracy	0.9491
F_1	0.9479

When comparing these results with those obtained using RBM, we can observe that the model without dimensionality reduction achieves better precision and recall. In fact, the improvement is significant across all metrics (accuracy, precision, recall, F_1). This superior performance suggests that the model without RBM is more effective at classifying bullying behavior on social media, as the dimensionality reduction has not affected its ability to learn relevant patterns.

Although dimensionality reduction with RBM could offer benefits in terms of reduced complexity and overfitting, the results indicate that, in this case, the model without RBM has shown better overall performance. The higher metrics without RBM suggest that the crucial information for effective classification has remained intact in the original model, while the dimensionality reduction with RBM might have impacted the model's ability to capture specific patterns in the data.

5 Conclusions

The results obtained allow us to conclude that the application of the proposed model demonstrates solid performance in classifying bullying behaviors. The accuracy of 0.9094 indicates that the model with dimensionality reduction (RBM) has a high correct classification rate, suggesting that it is efficient in distinguishing between bullying and non-bullying texts. On the other hand, there is a good balance between precision (0.9012) and recall (0.9094), as these values are very close. This indicates that the model maintains an adequate ability to correctly identify bullying cases (high recall), while also making precise predictions (high precision). This balance is crucial in applications where it is important to minimize false positives while accurately identifying bullying cases.

However, the results obtained without dimensionality reduction (RBM) show superior performance, with an accuracy of 0.9491, a precision of 0.9474, a recall of 0.9491, and an F_1 of 0.9479. These results suggest that the model without RBM has a greater capacity to correctly classify bullying behavior in texts, highlighting an improvement in both precision and recall compared to the version with RBM.

Furthermore, the F_1 of 0.8961 from the model with RBM, which combines both precision and recall, reinforces the conclusion that the model performs well. However, the model without RBM has an even higher F_1, further supporting its superior performance. This relatively high value, close to both precision and recall, suggests that the model efficiently handles both positive and negative cases without sacrificing one for the other. Taken together, the results reflect consistent performance in classifying cyberbullying texts.

The high precision, recall, and accuracy of the model without RBM demonstrate that it is more effective in identifying bullying behavior on social media compared to the model with RBM. Nevertheless, this does not detract from the merits of the model with RBM, which also shows robust performance, although the model without dimensionality reduction achieves a more accurate and reliable classification.

As future work, it is considered feasible to adjust the model's parameters, test it with additional data, explore alternative approaches, or carry out the extraction of syntactic, semantic, and pragmatic features.

References

Afrifa, S., & Varadarajan, V. (2022). Cyberbullying detection on twitter using natural language processing and machine learning techniques. *International Journal of Innovative Technology and Interdisciplinary Sciences, 5*(4), 1069-1080. https://doi.org/10.15157/IJITIS.2022.5.4.1069-1080.

Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., & Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. *IEEE Access, 7(1),* 70701-70718. https://doi.org/10.1109/ACCESS.2019.2918354.

Ali, A., & Syed, A. M. (2020). Cyberbullying detection using machine learning. Pakistan Journal of Engineering and Technology. *Pakistan Journal of Engineering and Technology, 3*(2), 45-50. https://doi.org/10.51846/vol3iss2pp45-50.

Alqahtani, A. F., & Ilyas, M. (2024). An Ensemble-Based Multi-Classification Machine Learning Classifiers Approach to Detect Multiple Classes of Cyberbullying. *Machine Learning and Knowledge Extraction*, *6*(1), 156-170. https://doi.org/10.3390/make6010009.

Altowairgi, R., Eshamwi, A., & Hsairi, L. (2023). Language Matters: A Systemic Functional Linguistics-Enhanced Machine Learning Framework for Cyberbullying Detection. *International Journal of Computer Science and Network Security, 23*(9), 192-198. https://doi.org/10.22937/IJCSNS.2023.23.9.25.

Azeez, N. A., Misra, S., Ogaraku, D. O., & Abidoye, A. P. (2024). A Predictive Model for Benchmarking the Performance of Algorithms for Fake and Counterfeit News Classification in Global Networks. *Sensors (Basel, Switzerland), 24*(17), 1-34. https://doi.org/10.3390/s24175817.

Chen, S., Wang, J., & He, K. (2024). Chinese Cyberbullying Detection Using XLNet and Deep Bi-LSTM Hybrid Model. *Information*, *15*(2), 2-18. https://doi.org/10.3390/info15020093.

Dewani, A., Memon, M. A., Bhatti, S., Sulaiman, A., Hamdi, M., Alshahrani, H., Alghamdi, A., & Shaikh, A. (2023). Detection of Cyberbullying Patterns in Low Resource Colloquial Roman Urdu Microtext using Natural Language Processing, Machine Learning, and Ensemble Techniques. *Applied Sciences*, *13*(4), 2-22. https://doi.org/10.3390/app13042062.

Ejaz, N., Razi, F., & Choudhury, S. (2024). Towards comprehensive cyberbullying detection: A dataset incorporating aggressive texts, repetition, peerness, and intent to harm. *Computers in Human Behavior*, 153(1), 108123. https://doi.org/10.1016/j.chb.2023.108123.

Llampa, Á. F., Farfán, J., & Rodríguez, M. E. (2020). *Análisis de sentimiento aplicado a la detección de ciberacoso en YouTube como red social*. In *VI Simposio Argentino de Ciencia de Datos y Grandes Datos (AGRANDA 2020) - JAIIO 49* (pp. 77-88). Sociedad Argentina de Informática.

López, S. T., Cuza, M. L. A., Pérez, P. Y. P., & Diéguez, L. A. P. (2016). Red neuronal multicapa para la evaluación de competencias laborales. *Revista Cubana de Ciencias Informáticas, 10*(1), 210-223.

Mahmud, T., Ptaszynski, M., & Masui, F. (2024). Exhaustive Study into Machine Learning and Deep Learning Methods for Multilingual Cyberbullying Detection in Bangla and Chittagonian Texts. *Electronics*, *13*(9), 1-36. https://doi.org/10.3390/electronics13091677.

Mamani A., M. (2020). Modelo basado en inteligencia artificial para la detección de ciberacoso (Bachelor's thesis). Universidad Mayor de San Andrés Ciudad de la Paz Estado Plurinacional de Bolivia.

Mercado, R. N. M., Chuctaya, H. F. C., & Gutiérrez, E. G. C. (2018). Automatic cyberbullying detection in spanish-language social networks using sentiment analysis techniques. *International Journal of Advanced Computer Science and Applications, 9*(7), *228-235*. https://doi.org/10.14569/IJACSA.2018.090733

Recalde M., J. A. (2021). El ciberacoso por redes sociales en el Ecuador (Bachelor's thesis). Universidad Politécnica Salesiana Sede Guayaquil.

Sultan, T., Jahan, N., Basak, R., Jony, M. S. A., & Nabil, R. H. (2023). Machine learning in cyberbullying detection from social-media image or screenshot with optical character recognition. *International Journal of Intelligent Systems and Applications, 15*(2), 2-13. https://doi.org/10.5815/ijisa.2023.02.01