ISSN 2007-9737

Comparative Analysis of Machine Learning and Deep Learning Models for Harassment and Discrimination Detection in Text

Ana Laura Lezama Sánchez[1,*], Mireya Tovar Vidal[2]

[1] Universidad Autónoma de Puebla (BUAP),
Facultad de Artes Plásticas y Audiovisuales,
Mexico

[2] Universidad Autónoma de Puebla (BUAP),
Faculty of Computer Science,
Mexico

{analaura.lezama,mireya.tovar}@correo.buap.mx

Abstract. Harassment and discrimination affect both workplace environments and online platforms. To To address this issue, we focus on automatically detecting such behaviors in textual data to help create safer digital spaces. In this article, we compare traditional machine learning and deep learning models for detecting harassment and discrimination. We evaluate four approaches: TF-IDF with logistic regression, BERT-based classification, a CNN with GloVe embeddings, and a GRU model enhanced with attention mechanisms and capsule networks. For all experiments, we rely on the Everyday Sexism Project dataset, which groups the texts into five categories: Workplace Harassment, Harassment, Discrimination, Sexism, and Other. We evaluate their performance applying accuracy, precision, recall, and F_1. The obtained results show that deep learning models outperform traditional methods in identifying complex linguistic patterns in abusive content.

Keywords. Machine learning, deep learning, harassment, discrimination

1 Introduction

Harassment and discrimination have become more visible all the time, especially in workplaces and educational environments. These behaviors arise in both direct, face-to-face interactions and digital communication. Because such content is produced at a fast pace and in large volumes, maintaining consistent human review has become increasingly complex. This situation has created a need for tools that can more efficiently and scalably identify abusive language [6].

In response to these behaviors, researchers in this Area have combined Natural Language Processing (NLP) with machine learning and deep learning techniques. These methods enable the examination of large amounts of text, hence the identification of linguistic patterns that often appear in harassment or discrimination. Systems built on these techniques can support moderation efforts and promote safer online environments [17].

In this paper, we propose four modeling approaches. First, we use TF-IDF combined with logistic regression. Second, we implement a BERT-based classifier. Third, we test a Convolutional Neural Network (CNN) with GloVe embeddings. Finally, we evaluate a GRU model that integrates attention mechanisms and capsule networks. We compare the results and assess their performance using precision, recall, accuracy, and F_1 to identify the best-performing model for this classification task.

This paper makes three main contributions.

First, we compare traditional machine learning models with deep learning approaches for detecting harassment in text. Second, we investigate how attention and capsule layers can enhance the interpretability of these models. Third, we carry out

an empirical evaluation using the Everyday Sexism Project dataset.

The paper is organized as follows. Section 2 summarizes related research. Section 3 introduces the background concepts. Section 4 describes the text classification models. Section 5 reports and analyzes the experimental results. Finally, Section 6 presents the conclusions of the study and future work.

2 Related Work

In this section, we review prior research related to the detection of harmful online behavior, with particular Attention to studies that incorporate deep learning, machine learning, and other artificial intelligence techniques. The works discussed here explore how we can use machine learning and natural language processing (NLP) to identify cyberbullying across various social media platforms.

In [11], a neural network-based method is presented for classifying texts related to harassment and discrimination across multiple domains, including workplaces, educational settings, and gender-based issues. Using an LSTM model, the researchers captured linguistic patterns commonly associated with harassment, reporting a precision of 0.8212, recall of 0.6883, F_1 of 0.7489, and overall accuracy of 0.7782.

In another work [10], their focus was on digital harassment on social media and the effectiveness of deep learning architectures. The study evaluated dense neural networks, CNNs, LSTMs, and Graph Convolutional Networks (GCNs) independently. Among these, a hybrid CNN-LSTM model performed best, achieving a precision of 0.840 and F_1 of 0.836, while an autoencoder with k-means clustering performed significantly worse.

These findings suggest promising directions for early detection and automated intervention.

Kennedy et al. (2025) [9] analyzed cyberbullying across multiple social media platforms. They used semi-structured interviews and focus groups with U.S. college students. From these discussions, they identified six key factors behind cyberbullying patterns. Among them are anonymity, post visibility, community norms, and moderation systems. Their findings suggest that platform design can reduce abusive behavior or make it worse, depending on how it is implemented. This perspective supports larger research efforts on online abuse.

Daouadi et al. (2024) [5] compared machine learning, deep learning, and transformer-based approaches for detecting hate speech in Arabic.

Their results show that fine-tuning large pre-trained models substantially improves performance.

Aklouche et al. (2024) [1] studied offensive content in English, evaluating multiple transformer architectures and ensemble methods. The study found that ensemble strategies enhance robustness across different datasets.

Arslan (2024) [3] presents a study about cyberbullying directed at the LGBTQ+ community. The authors evaluate transformer-based models, BERT, RoBERTa, and GPT-2 2 and find that RoBERTa performs the best. However, the authors noted that all models struggle to identify subtle or context-dependent forms of abuse. To address class imbalance, the authors apply oversampling techniques, including SMOTE and ADASYN. They conclude by emphasizing the need for more diverse datasets, multimodal information, and fairness-oriented training strategies to improve online safety.

Raju (2024) [13] investigates hate speech and personal attacks on platforms such as Wikipedia and Twitter. Classifiers and feature-extraction methods were evaluated. Showing that even simple techniques like Bag-of-Words and TF-IDF can achieve over 90% accuracy. Their findings indicate that the approach captures several forms of harassment effectively. However, the study emphasizes the need for future models that account for contextual information and platform-specific language patterns.

Gencoglu (2020) [7] further examines the role of machine learning in cyberbullying detection. The author proposes a fairness-based training strategy that effectively reduces bias while maintaining model performance. The object is aiming to encourage more transparent and ethical automated detection practices.

Srinath et al. (2021) [16] introduce BullyNet, a method based on signed networks. By combining a custom centrality metric, a bullying score system,

and contextual analysis of over 5.6 million tweets, the author successfully identified users exhibiting abusive behavior, achieving 80% precision and 81% accuracy.

Al-Garadi et al. (2019) [2] examined the entire cyberbullying detection process—from data collection and feature engineering to model construction and evaluation—with a focus on predictive modeling. They argued that a thorough assessment requires going beyond simple accuracy metrics, incorporating precision, recall, F_1, and AUC. Their study also stressed the importance of simulating aggressive online behavior through supervised learning and discriminative feature selection, pointing to new directions for creating generalizable and interpretable models.

In Salawu et al. (2017) [14], a review of cyberbullying detection methods was conducted.

The authors classify them into mixed-initiative, supervised learning, lexicon-based, and rule-based systems. Their analysis revealed persistent challenges, including the lack of high-quality labeled datasets. The simplification of bullying to mere swearing, and the limited Attention given to subtler forms of abuse, such as social isolation or repeated victimization.

3 Principal Concepts

In this section, we outline the main concepts that guide our study, the problem context, and the algorithms we rely on.

Harassment and discrimination still occur in workplaces, schools, and online, often harming people's physical and emotional health [8].

Since much of this behavior happens through written messages, automatic text detection is now a key way to reduce harm and make digital spaces safer. Natural Language Processing (NLP), a part of artificial intelligence, offers tools to analyze large amounts of text and find patterns linked to abusive or discriminatory language [4].

Text classification is a common way to detect harmful language. This task works by labeling pieces of text like posts or comments, based on what they say. Earlier methods mainly used features such as Term Frequency–Inverse Document Frequency (TF-IDF) with simple models, such as logistic regression [12]. As deep learning has advanced, models have become much better.

Newer architectures such as BERT, Convolutional Neural Networks (CNNs), and Gated Recurrent Units (GRUs) can understand more complex meanings and context, which helps them spot subtle forms of harassment [15].

The word embeddings (such as GloVe) convert words into dense vectors. Furthermore, this reflects their semantic similarity. Attention mechanisms help models focus on the most critical parts of the text. Capsule networks add another way to capture how language is structured, which is useful when harmful intent is suggested rather than directly stated [18].

In this paper, we test four text classification models (TF-IDF with Logistic Regression, BERT, a TextCNN model using GloVe embeddings, and a GRU model with attention and capsule layers). Each model offers a unique approach: TF-IDF analyzes word patterns, BERT provides context-aware word meanings, CNNs extract local meaning from text, and the GRU model captures both word order and deeper structure.

For all models, we measure performance using accuracy, precision, recall, and F_1. These metrics help us see how well each method balances false positives and false negatives in different categories.

4 Proposed Approach

In this section, we expose the proposed approach. First, the dataset, preprocessing, and implemented models are detailed. Figure 1 illustrates the overall workflow. Next, the four models are trained (TF-IDF + Logistic Regression, BERT, TextCNN with GloVe, and GRU with Multi-Head Attention and Capsules).

And finally, we evaluated the obtained results using performance metrics. The dataset used in the experiments was the Everyday Sexism Project, which contains 95,746 stories (see Table 1). The dataset is labeled across five categories: Workplace Harassment, Harassment, Discrimination, Sexism, and Other.

Table 1. Corpus description

Description	Value
Corpus source	Extracted from `everydaysexism.com`
Number of stories	95,746
Purpose of corpus	Experiences related to harassment and discrimination
Classes	Workplace harassment, Harassment, Discrimination, Sexism, Other

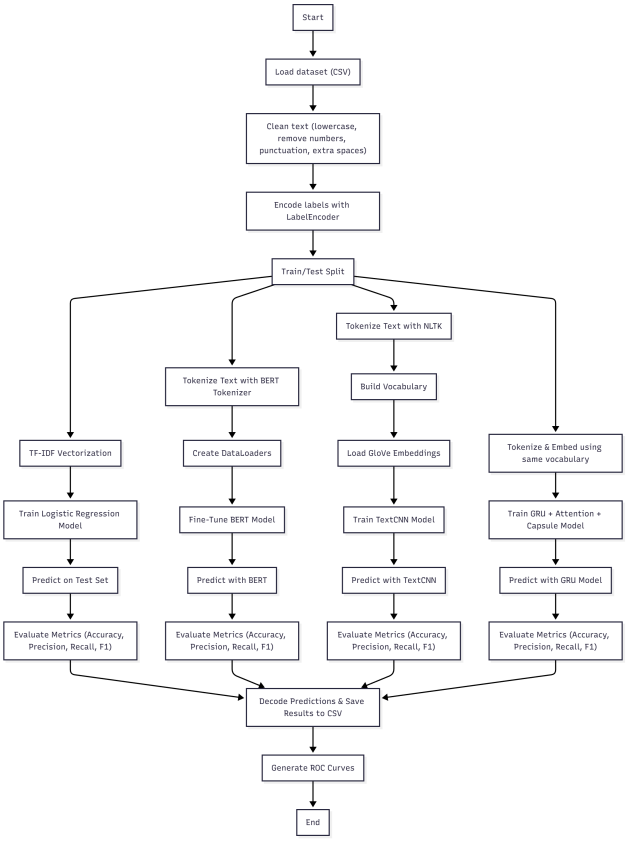

Fig. 1. General flowchart of the proposed process for the automatic detection of bullying and cyberbullying

4.1 Data Preprocessing

For data preprocessing, the text is converted to lowercase, and we removed numbers, punctuation, and extra spaces. We encode the labels with LabelEncoder to transform the categorical classes into numeric values. We divide the dataset into 80% for training and 20% for testing.

4.2 Implemented Models

We implemented four text classification models, each capturing semantic and syntactic patterns in different ways. The models used are described below:

— TF-IDF + Logistic Regression: Lexical features were extracted using the TF-IDF (Term Frequency-Inverse Document Frequency) technique. Selecting the 5,000 most frequent features, and trained a logistic regression model for classification. We trained the logistic regression model with a learning rate of 0.01 over 10 epochs.

— BERT (Bidirectional Transformer Encoder Representations) The pre-trained `bert-base-uncased` model was adjusted for multiclass classification. Training was performed in one epoch using the AdamW optimizer. We used DataLoaders for batch processing, with a learning rate of 5e-5 and a batch size of 16.

— TextCNN with GloVe Embeddings: We built a convolutional neural network (CNN). Pre-trained GloVe embeddings were used, and convolutional filters of sizes 3, 4, and 5 were applied, along with max-pooling and dropout layers to reduce overfitting. We trained for 10 epochs with the Adam optimizer, a learning rate of 0.001, and a dropout rate of 0.5.

— GRU with Multi-Head Attention and Capsules: A bidirectional GRU network was designed with a multi-head attention mechanism and a capsule layer to enrich the hierarchical representation of text features. The model was trained for 10 epochs using the Adam optimizer with a learning rate of 0.001 and a batch size of 32.

Table 2. Hardware and Training Configuration for the Models

Component / Parameter	Specification
Machine	Dell Laptop (Model unspecified)
CPU	Intel Core i7
RAM	64 GB
GPU	Training executed on GPU
Operating System	Windows 11
Python Version	3.10
Libraries	PyTorch, Transformers, Scikit-learn, NLTK

4.3 Evaluation Metrics

Models were evaluated using accuracy, precision, recall, and F_1, applying weighted averaging to address class imbalance. The predictions of each model, along with the correct labels, were exported to a CSV file for further analysis. In addition to overall accuracy, precision, recall, and F_1.

4.4 Training Configuration and Computational Environment

For the BERT model, we trained for one epoch due to computational constraints. Therefore, we trained TextCNN and the GRU model with attention and capsule networks for 10 epochs. We are using the Adam optimizer with a learning rate of 5e-5 and a batch size of 16. To address class imbalance across the dataset categories, we applied weighted loss functions during training. This type of strategy promotes balanced predictions and prevents the majority of classes from dominating the learning process. We evaluated all models on the validation set using accuracy, precision, recall, and F_1.

Therefore, Table 2 summarizes the hardware, computational environment, and software libraries used for training each model. This information allows reproducibility and comparison across different hardware setups.

5 Results and Discussion

In this section, we present the analysis based on the results obtained.

5.1 Discussion

As shown in Table 3, the TextCNN model achieves the strongest performance across all metrics. BERT also performs well, especially in recall, which is important for reducing false negatives in harassment detection. Deep learning models generally outperform the TF-IDF baseline, likely because they capture contextual and semantic cues more effectively. Our findings indicate that deep learning architectures—such as TextCNN and BERT—capture complex patterns of abusive language. These models are more effective than traditional approaches like logistic regression, particularly when paired with dense semantic embeddings. Model 4 (the GRU with attention and capsule networks) also achieves strong results and benefits from the attention mechanism, although its precision is slightly lower. Overall, the results show that modern NLP techniques, such as pre-trained embeddings and transformer-based architectures, offer notable improvements in the automatic detection of harassment and discrimination compared with approaches built on simpler linguistic representations.

While models achieve strong performance, further evaluation of fairness and potential bias is needed. Future work will include more comprehensive fairness-aware evaluation to mitigate potential bias across categories. After evaluating the models using standard metrics, we further analyze their discriminative ability using ROC curves for each class, providing insights beyond accuracy and F_1.

5.2 ROC Curve Analysis

To summarize each model's behavior, we report the Area Under the Curve (AUC) in Figure 2 shows the Receiver Operating Characteristic (ROC) curves for the multiclass classification models. The x-axis corresponds to the False Positive Rate (FPR), and the y-axis represents the True Positive Rate (TPR).

Table 3. Performance metrics for different detection models

Model	Accuracy	Precision	Recall	F_1
Model 1 (TF-IDF + Logistic Regression)	0.7640	0.7453	0.7640	0.7379
Model 2 (BERT)	0.7972	0.7383	0.7972	0.7647
Model 3 (TextCNN with GloVe embeddings)	0.8025	0.7784	0.8025	0.7795
Model 4 (GRU with attention and capsule networks)	0.7776	0.7113	0.7776	0.7418

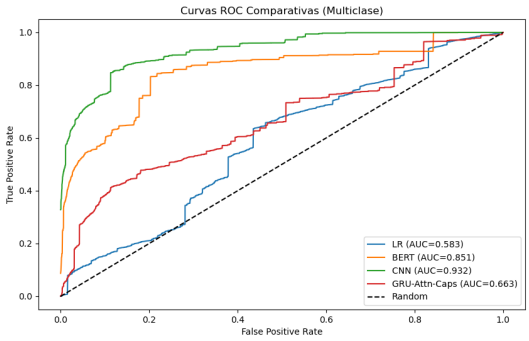

Fig. 2. Comparative multiclass ROC curves for different models. The CNN model achieves the highest AUC (0.932), followed by BERT (0.851), GRU-Attn-Caps (0.663), and Logistic Regression (0.583).

Curves that rise toward the upper-left corner (0, 1) indicate stronger classification performance.

The results indicate that:

— CNN (AUC = 0.932): This model achieves the strongest performance, showing excellent discrimination across the classes.

— BERT (AUC = 0.851): BERT also performs well, although its curve lies slightly below that of the CNN.

— GRU-Attn-Caps (AUC = 0.663): This model reaches a moderate level of performance.

— Logistic Regression (AUC = 0.583): This approach performs the worst, only slightly above the random baseline (dashed line).

Taken together, the curves show that deep learning models such as CNNs and BERTs substantially outperform the traditional approach. The CNN obtains the highest AUC, underscoring its ability to capture discriminative patterns in a multiclass setting.

While ROC curves indicate overall discriminative performance, model interpretability techniques, such as attention visualization and SHAP/LIME, help explain the specific features that influence predictions.

5.3 Explainability and Interpretability

To better understand model decisions, we analyzed attention weights from the GRU + Attention + Capsule model. Figure 3 highlights the most influential words contributing to correct classifications. Models with pre-trained embeddings, such as TextCNN + GloVe and BERT, effectively capture semantic patterns that help explain predictions.

For BERT and TextCNN, SHAP and LIME analyses were applied to interpret model predictions. Table 4 shows example words with high importance scores contributing to correct classifications. For instance, terms such as ”ignored”, ”mocked”, and ”harassment” were consistently highlighted as influential features for the Harassment class.

Finally, we examine misclassified instances to identify limitations and patterns in which the models struggle, highlighting areas for future improvement.

5.4 Error Analysis

Table 5 provides examples of misclassified instances. Most errors occurred in subtle harassment cases, such as sarcasm or context-dependent language, underscoring the challenges of automated detection.

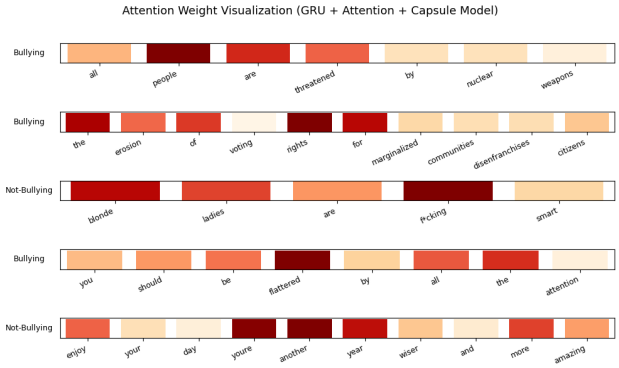

Fig. 3. Visualisation of attention weights for the GRU + Attention + Capsule model using examples from the Bullying and Not-Bullying classes. Darker colors indicate higher attention scores, highlighting the words that contributed most to each classification

Table 4. Example words with high SHAP/LIME importance for selected classes

Class	Top words
Harassment	ignored, mocked, jokes
Discrimination	gender, unfairly, criticized
Sexism	sexist, comments, appearance

Table 5. Examples of misclassified instances

Text	True Label	Predicted Label	Model
He constantly makes inappropriate jokes at work	Workplace harassment	Harassment	GRU + Attention + Capsules
She was ignored in meetings because of her gender	Discrimination	Other	TextCNN + GloVe
Colleagues mocked my appearance repeatedly	Harassment	Other	BERT
Boss criticized me unfairly for no reason	Workplace harassment	Discrimination	TF-IDF + LR

5.5 Discussion

Deep learning models with pre-trained embeddings (TextCNN + GloVe and BERT) outperform traditional TF-IDF + Logistic Regression models in terms of comprehensiveness. This is a critical factor for minimizing false negatives in bullying detection. The GRU + Attention + Capsule model demonstrates strong discrimination capacity, enabling interpretability through attention weights, though its accuracy can be improved. These results suggest that combining semantic embeddings with contextual attention mechanisms enhances automatic detection and discrimination of bullying in textual data.

Therefore, future work should explore other evaluation methods, such as hyperparameter optimization, fairness-based techniques, and more comprehensive interpretability analyses, to obtain higher-performing models and improve their social applicability.

6 Conclusion and Future Work

In this paper, we propose how several machine learning and deep learning models perform in detecting harassment and discrimination in textual narratives. The experiments show that Model 3 (TextCNN with GloVe embeddings) achieves the strongest results across all metrics, including precision, accuracy, recall, and F_1. Model 2 (BERT) also performs well, particularly in terms of recall, a key factor in sensitive tasks such as harassment detection, where missing relevant cases can have serious consequences.

We compared with traditional approaches reported in the literature, such as logistic regression and LSTM. The models evaluated here demonstrate a stronger ability to identify abusive content. These results show that benefits of combining deep learning architectures with dense semantic representations that can capture context-dependent language.

This work makes the next contributions. First, it evaluates both classical and modern NLP techniques on a real-world dataset. Second, it proposes hybrid architectures that provide a deeper understanding of implicit abuse patterns.

Our paper has some limitations. The models were trained and tested solely on the Everyday Sexism Project dataset, which may limit their generalizability to other domains.

In future work, we will focus on expanding the dataset, incorporating multimodal features, and applying training strategies to reduce potential biases. We also plan to explore real-time deployment with human oversight to evaluate model performance in dynamic environments better.

Acknowledgments

We thank the Ontological Engineering Laboratory, directed by Dra. Mireya Tovar Vidal, located at the Faculty of Computer Science of the Benemérita Universidad Autónoma de Puebla,

References

1. **Aklouche, B., Bazine, Y., Ghalia-Bououchma, Z. (2024).** Offensive language and hate speech detection using transformers and ensemble learning approaches. Computación y Sistemas, Vol. 28, No. 3, pp. 1031–1039.
2. **Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., Mujtaba, G., Chiroma, H., Khattak, H. A., 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, Vol. 7, pp. 70701–70718.
3. **Arslan, M., Madrigal, M. S., Abuhamad, M., Hall, D. L., Silva, Y. N. (2024).** Detecting lgbtq+ instances of cyberbullying. arXiv preprint arXiv:2409.12263.
4. **Baclic, O., Tunis, M., Young, K., Doan, C., Swerdfeger, H., Schonfeld, J., Data, P., Hub, I. (2020).** Natural language processing (nlp) a subfield of artificial intelligence. CCDR, Vol. 46, No. 6, pp. 1–10.
5. **Daouadi, K. E., Boualleg, Y., Guehairia, O. (2024).** Comparing pre-trained language model for arabic hate speech detection. Computación y Sistemas, Vol. 28, No. 2, pp. 681–693.
6. **Fnais, N., Soobiah, C., Chen, M. H., Lillie, E., Perrier, L., Tashkhandi, M., Straus, S. E., Mamdani, M., Al-Omran, M., Tricco, A. C. (2014).** Harassment and discrimination in medical training: a systematic review and meta-analysis. Academic Medicine, Vol. 89, No. 5, pp. 817–827.
7. **Gencoglu, O. (2020).** Cyberbullying detection with fairness constraints. IEEE Internet Computing, Vol. 25, No. 1, pp. 20–29.
8. **Ikbal, T. (2023).** Empowering change: A comprehensive framework to combat women's harassment and foster equality. International Journal of Multidisciplinary Research and Technology, Vol. 4, No. 8, pp. 23–47. DOI: 10.5281/zenodo.8281005.
9. **Kennedy, R., Bryson, S., Ellsworth, K., Kapur, I. (2025).** Fake profiles, mean comments, and toxic communities: College students' perspectives on cyberbullying across social media platforms. International Journal of Bullying Prevention, pp. 1–14.
10. **Lezama-Sánchez, A. L., Tovar Vidal, M. (2025).** Cyberbullying text classification using neural networks, generative models, and graph analysis. International Conference on Advances in Computing Research, Springer, pp. 197–206.
11. **Lezama-Sánchez, A. L., Tovar Vidal, M. (2025).** Multi-label classification of texts on harassment and discrimination. Pattern Recognition: 17th Mexican Conference, MCPR 2025, Guanajuato, Mexico, June 25–28, 2025, Proceedings, Springer Nature, Vol. 15715, pp. 165.
12. **Naeem, M. Z., Rustam, F., Mehmood, A., Ashraf, I., Choi, G. S., et al. (2022).** Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms. PeerJ Computer Science, Vol. 8, pp. e914.
13. **Raju, K., PROFF, M. B. J. B. A. (2024).** A systematic investigation of cyber harrassment intention behaviours and its impact on social media platforms. MATERIAL SCIENCE, Vol. 23, No. 04.
14. **Salawu, S., He, Y., Lumsden, J. (2017).** Approaches to automated detection of cyberbullying: A survey. IEEE Transactions on Affective Computing, Vol. 11, No. 1, pp. 3–24.
15. **Samant, R. M., Bachute, M. R., Gite, S., Kotecha, K. (2022).** Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review

and future directions. IEEE Access, Vol. 10, pp. 17078–17097.

16. **Srinath, A. S., Johnson, H., Dagher, G. G., Long, M. (2021).** Bullynet: Unmasking cyberbullies on social networks. IEEE Transactions on Computational Social Systems, Vol. 8, No. 2, pp. 332–344.
17. **Todorovic, M., Kozakijevic, S., Jovanovic, L., Babic, L., Antonijevic, M., Zeljkovic, V., Zivkovic, M., Bacanin, N. (2025).** Detecting harassment in user comments: Two-tier machine learning and metaheuristics approach with natural language processing. SN Computer Science, Vol. 6, No. 6, pp. 573.
18. **Worth, P. J. (2023).** Word embeddings and semantic spaces in natural language processing. International journal of intelligence science, Vol. 13, No. 1, pp. 1–21.

Article received on 27/08/2025; accepted on 02/12/2025.
**Corresponding author is Ana Laura Lezama Sánchez.*