www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 17(2),2026, 55-61. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2026.v17i2.1268

Evaluation of GRU with Attention and DistilBERT in Text Classification Tasks

Ana Laura Lezama Sánchez[1,2], *Mireya Tovar Vidal*[2]
[1] Secretaría de Ciencia, Humanidades, Tecnología e Innovacion, Av. Insurgentes Sur 1582, Col Credito Constructor, C.P.03940, Benito Juarez, Ciudad de Mexico, Mexico
[2] Faculty of Computer Science, Benemérita Universidad Autónoma de Puebla (BUAP), Avenida 14 Sur y Avenida San Claudio, Ciudad Universitaria, 72570 Puebla, Mexico
analaura.lezama@correo.buap.mx, mireya.tovar@correo.buap.mx

Abstract. In this paper compares two deep learning models for multiclass text classification in the cyberbullying domain. The corpus is written in English and includes six classes representing different types of harassment. The trained models are a GRU network with an attention mechanism and the DistilBERT transformer. The aim applies to a 5-fold cross-validation scheme. And uses accuracy, precision, recall, and F_1 as evaluation metrics. The GRU model achieves significantly higher performance (accuracy = 0.83, F_1 = 0.45) than DistilBERT (accuracy = 0.018, F_1 = 0.004). The results obtained indicate that the GRU architecture with word embedding performs more effectively on this task. This is due to the dataset's linguistic characteristics, as well as the limited fine-tuning capacity of the lightweight transformer. The findings emphasize the need to select model architectures that align with the corpus properties and the classification task's complexity. Future work will involve fine-tuning strategies and evaluating additional transformer-based models.

Keywords: Cyberbullying, Machine Learning, Deep Learning

Article Info
Received December 2, 2025
Accepted Jan 11, 2026

1 Introduction

The use of social media has driven the exponential growth of information. This activity has profoundly transformed the way people communicate. However, it has given rise to negative and difficult-to-control phenomena, such as cyberbullying. This problem, although a form of non-physical harassment, is considered digital violence. This type of aggression affects millions of users worldwide, primarily young people and adolescents. However, it also impacts a portion of the adult population, causing psychological, social, and, in extreme cases, legal consequences. To address this growing problem, implementing automated systems that can accurately detect abusive behavior in real-time has become increasingly necessary.

In the development of these systems, deep neural networks have demonstrated exemplary performance in recent years in natural language processing (NLP) tasks, such as text classification in different domains, including the detection of offensive content. Attention mechanisms have become key components of the most advanced models, as they allow the networks to selectively focus on the most relevant parts of the text when making decisions and thus identify texts with some cyberbullying (Anand et al., 2023).

Therefore, to contribute knowledge in this field, in this article, we propose a comparative evaluation of two deep learning models: the GRU network with an attention mechanism and the DistilBERT transformer for text classification in the field of cyberbullying. The corpus used is in English and consists of six different classes, i.e., five different types of harassment, as well as a class without harassment. For validation, we applied a *k*-fold cross-validation scheme with five partitions, using accuracy, precision, recall, and F_1 as evaluation metrics.

Transformer-based models such as BERT and DistilBERT have demonstrated excellent performance across many NLP tasks, but they do not always perform effectively in every context. The dataset's size and domain, the text's linguistic characteristics, and the chosen fine-tuning strategy strongly influence model performance. In contrast, recurrent models such as GRU, particularly when combined with attention mechanisms and domain-adapted embeddings, can better capture patterns in smaller or more topic-specific corpora. Cyberbullying datasets include implicit biases related to demographic groups and linguistic stereotypes. Therefore, in this paper, we compare the two architectures not only to measure performance but also to understand how data properties and task complexity should guide model selection.

Our objective is to determine how model architecture and text representation affect performance on a real cyberbullying dataset. For the evaluation, we used accuracy, precision, recall, and F_1 under 5-fold cross-validation to ensure the reliability of results.

The article is structured as follows: Section 2 provides a review of prior literature on the topic. Section 3 presents the key concepts. Section 4 describes the proposed text classification models. Section 5 analyzes the results obtained. Section 6 presents the discussion. Section 7 presents the conclusions and future work. Finally, the bibliography consulted during the development of this work.

2 Related Work

In this section, we review previous research relevant to the analysis of cyberbullying in text, with special emphasis on studies that use deep learning, machine learning and other artificial intelligence techniques. Early approaches to cyberbullying detection relied on lexical features and classical machine learning models. These approaches used manually engineered features, which limited their generalization to unseen or informal text.

The introduction of distributed representations, such as Word2Vec and GloVe, enabled models to capture semantic relationships between words, improving performance when paired with neural architectures such as LSTM and GRU. Transformer-based models (BERT, RoBERTa, DistilBERT) have demonstrated state-of-the-art performance in text classification tasks due to their capacity to learn bidirectional representations. However, their effectiveness depends on domain adaptation and the availability of computational resources. In scenarios where datasets are imbalanced, informal, or small, simpler recurrent architectures may still outperform transformers.

Sultan et al. (2023) exposes a study that examines how digital communication has profoundly changed the way people interact. This allows them to overcome traditional barriers of time and space. The authors examined the effectiveness of deep learning and machine learning techniques, such as bidirectional long-term memory (BiLSTM), for detecting cyberbullying. Additionally, the authors suggested a universal metadata-driven architecture that, based on their findings, performs better than alternative approaches when used for social network data flows.

Barlett (2023) describes Cyberbullying from a theoretical standpoint. He exposes a study, and results lend credence to the notion that psychological theory-based therapies are typically more successful for opposing this problem. The suggested method pinpoints emotional and cognitive elements that greatly impact violent online conduct, including hostile attribution biases, digital disinhibition, and anonymity.

In Nghiem et al. (2024) there is a classification architecture based on prototypical networks. The authors propose a meta-learning approach for identifying harmful language. Notably, this method achieves over 75% of the maximum F_1 with less than 10% of the training set, demonstrating its effectiveness even with sparse training data. The usage of joint representations in this approach, which integrate data and label semantics, may expand its applicability to further natural language processing applications like pose identification or sentiment analysis.

To contextualize the complex nature of online harassment both technically and morally the writers' viewpoint is crucial. However, Hasan et al. (2023) investigate deep learning techniques for cyberbullying detection and their benefits over conventional models, especially about automated feature extraction and scalability. The authors also highlight important gaps in literature, including the usage of multimedia formats like deepfakes, the psychological profiles of bullies, and the scant attention given to individual behavioral patterns.

In Aldhyani et al. (2022), they propose a hybrid CNN-BiLSTM detection model, which they evaluate in both binary and multiclass bullying scenarios. The authors' reported results indicate that BiLSTM outperforms previous models in multiclass classification, achieving accuracy rates close to 99%. However, the authors also acknowledge limitations, such as the risk of overfitting and the exclusive use of English-language data. The authors mention that these findings point to the need for more robust multilingual transformer architecture as a next step in advancing cyberbullying detection.

Therefore, this article addresses automatic text classification in the field of cyberbullying, a problem explored in the literature using different deep learning-based approaches.

To contribute to this field, we propose a comparative evaluation of two architectures widely used in NLP: a GRU network with an attention mechanism and the DistilBERT transformer. The corpus used is in English and consists of six different classes with different type of cyberbulling. For validation, we used a k-fold cross-validation scheme with five partitions, using accuracy, precision, recall, and F_1 as evaluation metrics.

3 Principal Concepts

In this section, we present the key concepts that support our research. The emergence of social platforms for communicating from anywhere in the world and with people we don't know has given way to the increasing digitalization of social life, and this, in turn, has given rise to new forms of violence, particularly cyberbullying. This phenomenon manifests itself through repeated attacks, harassment, or intimidation through digital media such as social networks, instant messaging, or online platforms (Kazan, 2022). Cyberbullying spreads quickly which increases its psychological and social impact. The growing volume of user-generated content has made automatic detection and challenge for data science, computational linguistics, and artificial intelligence (Kazan, 2022).

Cyberbullying arises because it occurs in contexts where we in any context maybe don´t know the face and/or name of the aggressor. That is, digital technologies are used to send offensive messages to ridicule or humiliate the victim. These actions can affect individuals of any age and lead to problems such as anxiety, depression, and even suicidal thoughts. As online information continues to expand, manual efforts to identify these behaviors have become impractical, as they require considerable time and sustained attention from human reviewers.

In response to this issue, artificial intelligence plays a key role through Natural Language Processing (NLP), a field that enables computers to interpret and analyze human language. NLP analyzes text to identify patterns associated with aggressive or inappropriate expressions, enabling the detection of cyberbullying or other type of conduct. This process relies on converting text into numerical representations that algorithms can process, using methods such as tokenization, vectorization, and word embeddings (Ghosh et al., 2025). Embedding map words into dense vectors that capture semantic relationships, allowing deep learning models to interpret context more effectively.

Among the architectures used for this task are Recurrent Neural Networks (RNNs), which are designed to process sequential data such as language. However, RNNs struggle to capture long-term dependencies because of the vanishing gradient problem (Almufareh et al., 2025). To address this limitation, researchers developed variants like Gated Recurrent Units (GRUs), which regulate the flow of information through internal gating mechanisms that determine what to retain and what to discard during sequence processing.

Attention mechanisms allow the model to dynamically focus its analysis on the most relevant parts of the text for the task at hand. Other approaches that treat all words equally, attention gives greater weight to the most significant terms for classification. GRU with attention achieves a richer and more contextualized representation of the text, which improves the model's ability to recognize cyberbullying patterns.

DistilBERT, have gained popularity recently because of their higher accuracy. Based on the Transformer architecture and trained on massive amounts of text, DistilBERT is a lighter and more efficient variant of BERT (Bidirectional Encoder Representations from Transformers). BERT transformed natural language processing. DistilBERT is appropriate for text classification problems with constrained computational resources since it maintains around 97% of BERT's performance while having a more effective structure and shorter training times (Devi et al., 2025).

Therefore, to understand the results obtained with each of the models described, it is necessary to evaluate the performance of the implemented classification models.
The metrics used are:

- Accuracy: proportion of correct predictions.
- Precision: proportion of true positives among all optimistic predictions.
- Recall: the ability of the model to find all positive cases.
- F_1: harmonic mean between precision and recall.

4 Proposed Approach

The proposed methodology is structured into five main phases, covering the entire process from data preparation to model evaluation:

- Phase 1: Data Preparation
  - A pre-labeled dataset of short text messages, categorized into six classes, was used.
    We applied the following preprocessing steps:
    - Normalization of text (lowercasing, punctuation removal, and whitespace cleaning).
    - Stopword removal and basic token normalization.
    - Label encoding using LabelEncoder and OneHotEncoder to convert class labels into categorical form suitable for multiclass classification.

- Phase 2: Text Tokenization and Representation
  - GRU model
    - We tokenized the text using the Keras Tokenizer, which converted each text into an integer sequence.
    - We padded all sequences to a fixed length of 300 using pad_sequences to ensure a uniform input size.
    - We obtained word representations from pre-trained 300-dimensional GloVe embeddings.
    - We constructed an embedding matrix that mapped tokenizer indices to their corresponding GloVe vectors and loaded it into the embedding layer.
  - Distilbert model
    - The Hugging Face AutoTokenizer was used to convert raw text into token IDs compatible with transformer-based input.
    - No external embedding layer was required, as DistilBERT learns contextual representations internally.
- Phase 3: Model Construction
  - GRU with Attention
    The architecture consists of the following components:
    - Embedding Layer initialized with GloVe weights (frozen during training to preserve semantic structure).
    - GRU Layer with a hidden state size of 128, which processes sequential information.
    - An Attention Mechanism is implemented to assign varying importance to different tokens in the sequence.
    - Fully Connected Dense Layers, including:
      - One intermediate dense layer with ReLU activation.
      - A final softmax output layer for six-class classification.

  - DistilBERT
    - We used the TFAutoModelForSequenceClassification architecture.
    - This model includes the transformer encoder and a classification head.
    - We fine-tuned the entire model end-to-end but used a low learning rate due to the sensitivity of transformer weights.
- Phase 4: Training
  - Both models were trained using *K*-Fold cross-validation (*k*=5), enabling performance assessment across different subsets of the data.
  - GRU model
    - Loss function: categorical_crossentropy
    - Optimizer: Adam (Learning rate = 0.001 (1e-3))
    - Batch size = 32
    - Epochs = 10
  - DistilBERT
    - Optimizer: Adam with learning rate 3e-5
    - Fine-tuning performed end-to-end, using the model's built-in classification loss.
- Phase 5: Evaluation

- Each fold in the cross-validation was evaluated using standard performance metrics: accuracy, precision, recall, and F_1.
- Classification reports and saved predictions for further analysis.

5 Results and Discussion

In this section, we present the results obtained from the developed model. Additionally, we describe the dataset used for evaluating the model.

5.1 Dataset Description

In this section, the dataset used in this work is described (see Table 1 (Wang et al., 2020)). The corpus, which is in English, consists of 6 different classes, representing 6 distinct categories. The classes include gender, religion, age, and ethnicity, which correspond to 4 types of cyberbullying. Additionally, there is a class representing instances with no cyberbullying and another class labeled others.

Table 1. Dataset

Class	Size
Age	8,000
Gender	8,000
Religion	8,000
Ethnicity	8,000
No cyberbullying	8,000
Others	8,000

5.2 Results

In this work, two text classification models, DistilBERT and GRU, were evaluated. Five-fold cross-validation was applied. In Table 2, the DistilBERT model performed relatively poorly across all metrics. The model's average precision was 0.0322 with a standard deviation of 0.0522, while the average recall was 0.0556 ($\sigma = 0.0052$) and the average F_1 was 0.0041 ($\sigma = 0.0008$). This suggested that, although DistilBERT can identify some positive cases, its overall performance on this dataset is limited, reflecting high variability between folds.

On the other hand, the GRU model showed significantly superior performance. The average precision was 0.4803 ($\sigma = 0.0528$), the average recall was 0.4578 ($\sigma = 0.0588$), and the average F_1 reached 0.4513 ($\sigma = 0.0537$). This indicates that the GRU model is not only more accurate but also more consistent across folds, with relatively low standard deviations.

In terms of overall accuracy, DistilBERT achieved a value of 0.0188, while GRU achieved 0.8313, demonstrating that the GRU model is clearly more effective for the specific task evaluated.

Therefore, the results suggest that, for this dataset and experimental setup, GRU is the most suitable model, showing a much more favorable balance between precision, recall, and F_1, as well as lower variability between folds, compared to DistilBERT.

Table 2 presents the GRU model achieved significantly higher accuracy and F_1, indicating more effective handling of the classification task. DistilBERT, despite its theoretical capacity for contextual representation, underperformed in all evaluation metrics.

Table 2 presents the average performance across 5 folds for both models.

Model	Accuracy	Precision	Recall	F_1
GRU + Attention	0.83	0.47	0.44	0.45
DistilBERT	0.018	0.006	0.004	0.004

Table 3 provides qualitative examples of model predictions. Here, we compare the original labels with the projections from both GRU and DistilBERT. The examples demonstrate that GRU is effective in correctly classifying bullying content, including highly explicit or offensive examples. At the same time, DistilBERT frequently fails to identify such content, often misclassifying bullying as non-bullying.

Table 3 Example Predictions by GRU and DistilBERT Models for Bullying Detection in Text

Text	Label_original	GRU_pred	DistilBERT_pred
Ten outside soon doctor shake everyone treatment seat radio land.	Not-Bullying	Not-Bullying	Not-Bulllying
I wanna fuck you	Bullying	Bullying	Not -Bullying
Male action movies...ehhhh @That_Damn_Ram Not sexist but female action movies...ehhhh	Bullying	Bullying	Not Bullying
I hope this round humbled the girls, they can't cook	Not- Bullying	Not- Bullying	Not Bullying
you not only a nigger, you're a dumb nigger. Fuck you and your riots. You want war?	Bullying	Bullying	Not -Bullying
Really miss my classmates n schoolmates. See you all soon people	Not-Bullying	Not-Bullying	Not Bullying
Idiot I'm talking about Halal Products not Muslims. Are you saying Halal products means Muslims ?	Bullying	Bullying	Not Bullying
Me your big friend from bangladesh	Bullying	Bullying	Not Bullying

6 Discussion

The performance difference suggests that, for this corpus, the GRU model with GloVe embeddings was better at representing linguistic structure. Three key factors explain this result:

1. Embedding Adaptation:
    1. GloVe embeddings were semantically aligned with the dataset. While DistilBERT may not have sufficiently adapted during the limited fine-tuning.
2. Data Characteristics:
    1. The dataset contains informal language, abbreviations, and non-standard writing. GRU can rely on sequential learning. On the other hand, DistilBERT requires deeper contextual alignment that may not occur without more extensive training.
3. Cross-Validation Stability:
    1. The GRU model showed consistent performance across folds. DistilBERT exhibited high variance and failed to converge effectively.

Therefore, this supports the idea that model selection must consider dataset scale, writing style, and computational constraints, not only state-of-the-art trends.

When comparing our results with those reported by Wang et al. (2020), substantial differences are evident. Their study introduced a semi-supervised Dynamic Query Expansion (DQE) framework to automatically generate a balanced version of the original dataset, thereby addressing severe class imbalance. They also evaluated multiple models, including DistilBERT, achieving an F_1 of approximately 0.91. In the present work, DistilBERT achieved an F_1 of 0.004 under a five-fold cross-validation scheme using the original, unbalanced dataset. These differences highlight that the superior performance reported in (Wang et al., 2020) is primarily due to dataset rebalancing, data augmentation, and extensive fine-tuning. In contrast, our approach emphasizes baseline evaluation with minimal preprocessing. Consequently, our findings serve as a reference point for understanding model behavior in raw, real-world conditions.

7 Conclusions and Future Work

In this paper, we demonstrate that a GRU model with attention and GloVe embeddings outperforms a DistilBERT transformer in multiclass cyberbullying detection. The GRU achieved higher accuracy and F_1, highlighting the effectiveness of recurrent architectures when the dataset contains informal and domain-specific language.
We propose that Future work will address:

- Experimentation with RoBERTa or BERT-base to improve transformer adaptability.
- Data augmentation to address class imbalance.
- Integration of linguistic and semantic metadata to enrich classification context.

The findings emphasize that the most advanced model is not always the most effective, and we must evaluate model performance relative to dataset characteristics.

References

Aldhyani, T. H., Al-Adhaileh, M. H., & Alsubari, S. N. (2022). Cyberbullying identification system based on deep learning algorithms. *Electronics, 11*(20), 3273. https://doi.org/10.3390/electronics11203273

Almufareh, M. F., Jhanjhi, N., Humayun, M., Alwakid, G. N., Javed, D., & Almuayqil, S. N. (2025). Integrating sentiment analysis with machine learning for cyberbullying detection on social media. *IEEE Access*.

Anand, M., Sahay, K. B., Ahmed, M. A., Sultan, D., Chandan, R. R., & Singh, B. (2023). Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. *Theoretical Computer Science, 943*, 203–218. https://doi.org/10.1016/j.tcs.2022.11.047

Barlett, C. P. (2023). Cyberbullying as a learned behavior: Theoretical and applied implications. *Children, 10*(2), 325. https://doi.org/10.3390/children10020325

Devi, K. N., Jayanthi, P., Rajasekar, V., Rathinasamy, R., Dhandapani, P., & Risam, R. (2025, February). Cyberbullying detection using deep learning algorithms. In *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)* (pp. 1298–1306). IEEE.

Ghosh, R., Malhotra, M., & Kumar, N. (2025). Cyber bullying in the digital age: Challenges, impact, and strategies for prevention. In *Combating cyberbullying with generative AI* (pp. 151–180). IGI Global.

Hasan, M. T., Hossain, M. A. E., Mukta, M. S. H., Akter, A., Ahmed, M., & Islam, S. (2023). A review on deep-learning-based cyberbullying detection. *Future Internet, 15*(5), 179. https://doi.org/10.3390/fi15050179

Kazan, H. (2022). Cyber bullying and violence literacy in the context of digitalization. In *Research anthology on combating cyber-aggression and online negativity* (pp. 496–519). IGI Global.

Nghiem, H., Gupta, U., & Morstatter, F. (2024). *"Define your terms": Enhancing efficient offensive speech classification with definition* (arXiv:2402.03221). https://arxiv.org/abs/2402.03221

Sultan, D., Toktarova, A., Zhumadillayeva, A., Aldeshov, S., Mussiraliyeva, S., Beissenova, G., & Imanbayeva, A. (2023). Cyberbullying-related hate speech detection using shallow-to-deep learning. *Computers, Materials & Continua, 75*(1), 2115–2131. https://doi.org/10.32604/cmc.2023.032993

Wang, J., Fu, K., & Lu, C.-T. (2020). SosNet: A graph convolutional network approach to fine-grained cyberbullying detection. En *2020 IEEE International Conference on Big Data* (pp. 1699–1708). IEEE.