A logo of a university

Description automatically generated

**Natural Language Processing**

**DSCI 6004**

**Final Project ON**

**Comparative Analysis of NLP Architectures for Cyberbullying Detection and Classification Across Social Media and Online Gaming Platforms**

**Done BY**

**REVANTH PAINENI(**[**rpain1@unh.newhaven.edu**](mailto:rpain1@unh.newhaven.edu)**)**

**SRIDHAR YEMUNDLA(**[**syemu1@unh.newhaven.edu**](mailto:syemu1@unh.newhaven.edu)**)**

**Contents**

**[Abstract:](#_Toc7043)** [3](#_Toc7043)

[**Introduction:** 3](#_Toc7044)

[**Related Work:** 4](#_Toc7045)

**Methods**[**:** 5](#_Toc7046)

[Data Preprocessing: 5](#_Toc7047)

Feature Engineering……………………………………………………………………………………………………………………………6

[Model Architectures: 7](#_Toc7048)

Model Training and Evaluation……………………………………………………………………………………………………………8

**Results:**………………..………………………………………………………………………………………………………………………………9

Model Performance.…..………………….……………………………………………………10

Comparative Analysis………………………………………………………………………....10

Training Results…………………………...…………………………………………………..11

Text Data Analysis…………………………………………………………………………….12

Audio Data Analysis……………………...…………………………………………………..13

[**Discussion:** 13](#_Toc7052)

[**Conclusion:** 14](#_Toc7053)

**References**:………………………………………………………………………………………………..15

**Abstract:**

This report compares various machine learning architectures for the detection of cyberbullying in textual datasets. Our focus is on the effectiveness of Generalized Recurrent Units (GRU), Convolutional Neural Networks (CNN) and BERT (Bidirectional Encoder Representations from Transformers). The model uses different preprocessing methodologies and embedding techniques and incorporates them in each phase to improve the predictive performance. This report compares the uniqueness of each model and finds out the best strategies to curb the problem of cyberbullying in digital communication channels.

**Introduction:**

The most pressing issue in digital interaction is cyberbullying. It affects users in almost all digital spaces, and the more active we become in the digital world, the more important it is to find protective methods to detect and reduce cyberbullying. The goal of this report is to provide a solution to this growing problem by employing the latest machine learning techniques to recognize cyberbullying and stop it. Using Natural Language Processing techniques, this research aims to provide useful tools to protect users from cyberbullying, for example, for moderators of platforms and social media. We evaluated three different machine learning models – GRU, CNN and BERT – that are well-suited for processing and interpreting the intricacies of human language (which is important for effective harmful content identification).

The Generalised Recurrent Unit (GRU) model is a type of recurrent neural network suited to learn temporal dependencies of data, while Convolutional Neural Networks (CNN) are well-suited to learn hierarchical features of spatial data, which we apply to textual contexts in order to robustly learn increasingly complex features. Lastly, BERT is a deep learning model that becomes context-aware by looking in both directions in text.

The fine-tuning of these models using thoroughly preprocessed datasets and the use of advanced embedding vectors such as GloVe can result in a sensitivity that can capture the linguistic nuances of cyberbullying, not only for immediate detection but to analyse patterns that are forever changing as the online society adapts to a brand-new form of harassment. The findings of this study will serve as basis to the development of new, more effective methods of preventing cyberbullying, paving the way towards a more civilised form of discourse online.

**Related work:**

Because of these realities, current advances in cyberbullying identification highlight that it’s an urgent and multilingual, cross-cultural problem, and that the use of deep learning techniques to classify are increasingly a focus of research in this area thanks to their capability of understanding the subtleties of cyberbullying language.

Dewani et al (2021), in their work on cyberbullying in a less-resourced language Roman Urdu, have used deep neural network architectures such as RNN-LSTM and RNN-BiLSTM with promising efficacy and with high validation accuracies and F1 scores. This further emphasises the need for a language-specific approach to preprocessing, in addition to the need especially for low-resource languages for the development of linguistic resources, including slang/insult dictionaries. Neelakandan et all’s work on a novel framework that integrates deep learning with feature selection algorithms to improve cyberbullying detection on social media also reported a high rate of accuracy using their SSA-DBN model. This reveals the potential of utilising sophisticated algorithmic strategies to refine detection processes.

Meanwhile, Akinyemi et al (2023) used a combination of traditional machine learning techniques and a deep learning sequence model to identify cyberbullying from nine different categories such as age, religion, and ethnicity. Their study reported high accuracy on Random Forests, demonstrating the feasibility of ensemble and hybrid models. Shankar et al (2022) used supervised machine learning methods to classify abusive and not-so-abusive content from Twitter and found that SVMs and Naive Bayes models gave them a high degree of precision in their results. Such results convey the importance of applying bespoke approaches for various types of social media and its content.

Moreover, a second study also highlighted the difficulty of detecting cyberbullying that combines several expressions of hate, using the multifaceted approach. The teams took advantage of the power of deep learning algorithms to handle many types and sources of data. Furthermore, this approach stressed the need to continue developing new algorithmic approaches to keep up with the ever-evolving cyberbullying tactics and a growing sphere of digital presence.

Collectively, these studies suggest that models are being adapted to specific linguistic and cultural settings and are being paired with sophisticated preprocessing methods to better identify bullying and better characterize cyberbullying on different platforms and across different languages. In summary, these examples represent a series of important steps toward alleviating cyberbullying in an increasingly interconnected online society.

**Methods**

**Data Collection and Preprocessing**

The dataset that I used for this study is "Cyberbullying Tweets" which was obtained from Kaggle, and it contains: Bullying based on gender, ethnicity or age. This is a good dataset because you can find different kinds of cyberbullying that are based on specified groups, gender, ethnicity or age.

Some basic preprocessing included scrubbing the data for URLs, user mentions and hashtags, to get to the meat of the actual text. Tokenisation was then used to split the text into words, or tokens, and stopwords were pruned to remove the most common, irrelevant words. More sophisticated preprocessing involved using a Porter Stemmer to reduce words to their base or root. Reducing words to their base form makes the input for modelling more uniform.A computer screen shot of a program code

Description automatically generated

**Feature Engineering**

Textual data was first tokenised into numbers using the TensorFlow Keras Tokenizer class to convert text into sequences of integers. Each sequence was then padded to the same length (200 words) to ensure the same input size for all models and to construct the same input as microblog comments. Secondly, pre-trained GloVe word vectors were used to vectorise words into a 200D space that contains rich pre-learned relationships between words which are model-friendly.A screenshot of a computer program

Description automatically generated

**Model Architectures**

**Three main deep learning architectures were implemented:**

* **GRU-Based Model:** This model comprised of Gated Recurrent Unit (GRU) layer to deal with sequence data by capturing the relations between data and time, where the architecture is tuned with the help of dropout layers to reduce overfitting, and lastly trained with dense output layer including SoftMax activation for classifying the input into different types of cyberbullying communication.

A computer screen shot of a program code

Description automatically generated

* **CNN-Based Model:** The architecture of our CNN model consisted of 1D convolutional layers, which are especially suited for the extraction of features from sequences of data. After each convolutional layer, the data was passed through a layer of max pooling (in order to reduce its dimension), and batch normalization (in order to speed up training). Spatial dropout was also used, to make the model more robust to overfitting.A screen shot of a computer

  Description automatically generated
* **BERT-Based Model:** This model was fine-tuned on the cyberbullying data using a pre-trained BERT model. By using bidirectional training of text, which is an important feature of BERT, this model is more effective at modelling context compared with one-directional models.A screen shot of a computer program

  Description automatically generated

**Model Training and Evaluation**

Models were trained using the Adam optimiser with a categorical crossentropy loss function due to the multiclass nature of the output. Two callbacks were used for the training process, namely Early Stopping and ReduceLROnPlateau to stop training when the validation loss has saturated and reduce the learning rate when the validation score has plateaued respectively.

The performance of each model was then evaluated by measuring the accuracy, precision, recall and F1-score metrics to indicate the success of each approach in detecting and classifying the cyberbullying instances.

**Results**

***Model Performance***

***GRU-Based Model***

The GRU-based model was able to effectively process sequential text data, since all words in a sentence have a certain temporal order, mimicing human language. After 30 epochs of training the model achieved 82.52% of accuracy at training set, and 82.45% of accuracy at validation set. The F1-score was balanced, at about 88.58 and 75.79 for precision and recall, respectively, for a total score of 0.8169, which shows that the GRU model performed well but was somewhat conservative, valuing accuracy over recall when it came to predicting instances of cyberbullying.A group of graphs showing different results

Description automatically generated with medium confidence

***CNN-Based Model***

The CNN model in particular did a great job identifying interesting features in the text data (it used convolutional layers to learn salient features from the padded sequences along the length of the input). In comparison with the GRU, it was much less computationally intensive and even managed a slightly better validation accuracy of 83 per cent. However, the F1-score measured across the training and validation phases also indicated that, while it was accurate.A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

***BERT-Based Model***

The BERT model with deep bidirectional architecture had the highest accuracy in the validation score of 87%; it also had higher precision and recall than other models tested; recall is a measure of accuracy relative to the true occurrence of a phenomenon, and it demonstrates the ability of the BERT model to understand the contextual and affective variations in the input text. The F1-score for the BERT model remained over 0.85 in all epochs, which demonstrates its ability to classify cyberbullying reliably.A graph of a line and a line

Description automatically generated with medium confidence

**Comparative Analysis**

Comparing the models indicates that although GRU and CNN were faster and efficient and both computationally less intensive than BERT, BERT’s deep learning achieved the best results and came out on top of the traditional models overall performance measures, especially in extreme examples of cyberbullying, where contextual information is important.

**Training Results**

Amongst all the graphs that illustrated the training process for each model clearly, shown were the learning curves of accuracy and loss for each model in the respective graphs. In all, improvement of the models over epochs were consistently observed with the early stopping mechanisms working efficiently by preventing overfitting from happening. Unlike the accuracy and loss graphs in each of the models discussed earlier, graphs that illustrated precision, recall and F1-score showed the strong and weak points of the models.

To conclude, BERT was the most stable and performed the best and the highest in all the stated metrics.

**Text Data Analysis**

The research team tested if the BERT model can be applied in a real-world task by classifying a piece of text it hadn’t seen before. The text was tokenised with the tokeniser we trained the BERT model with, and then classified: This text contains evidence of other\_cyberbullying according to our BERT model, demonstrating how the model can successfully help us interpret unseen data.A screen shot of a computer program

Description automatically generated

**Audio Data Analysis**

The secondary usage of the model was to analyze an audio file to ensure that it is free from any cyberbullying content. The input audio file was converted from MP3 to WAV format for further processing. The WAV file was then converted to plain text using speech-to-text conversion. The text, as an input, was then subjected to classification as cyberbullying using the BERT model. The output of the classification demonstrated that the model classified the transcribed text in the audio file into 'other\_cyberbullying' class, with an accuracy of 0.69. This test showcases the model's continuous usage by analysing multimedia data while performing its predefined tasks and maintaining its accuracy.A screen shot of a computer program

Description automatically generated

**Discussion**

As we see from the results, although models such as GRU and CNN can work well in the simpler cases, when the cyberbullying dilemma becomes more difficult , the BERT performance improves substantially and becomes a necessity to achieve the best detection rates. Moreover, the use of pre-trained embeddings such as GloVe also contributed significantly to the model performance, hinting that better linguistic representational features could be a key to make the models work well.

**Conclusion**

The investigation into the application of GRU, CNN, and BERT models for cyberbullying detection highlights the substantial benefits of integrating deep learning techniques in the fight against online harassment. Among the models tested, the BERT model exhibited superior performance due to its ability to deeply understand the context and nuances embedded within language, demonstrating an impressive balance of accuracy, precision, and recall. This robust performance extended to successfully predicting cyberbullying instances in unknown text data and accurately classifying content from audio files, showcasing the model's versatility and efficacy across different media types. Cyberbullying that while traditional models remain useful for certain applications, the complex nature of human language and the subtleties of cyberbullying are best captured using advanced, context-aware systems.

Looking forward, enhancing cyberbullying detection frameworks with deep learning not only improves detection rates but also helps in crafting more nuanced responses to prevent harassment. Future work should explore the integration of multimodal data sources, such as images and videos, to create more comprehensive detection systems. Additionally, extending these models to accommodate multiple languages and cultural contexts will enhance their applicability and effectiveness in global cyber safety efforts, ensuring safer online environments for users worldwide. The demonstrated success in handling unknown data and audio inputs further supports the potential of these models to adapt to a wide array of situations and content formats, reinforcing their utility in dynamic real-world applications.

**References:**

Dewani, A., Memon, M. A., & Bhatti, S. (2021). Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data. Journal of Big Data, 8(160).

Neelakandan, S., Sridevi, M., Chandrasekaran, S., Murugeswari, K., Pundir, A. K. S., Sridevi, R., & Lingaiah, T. B. (2022). Deep learning approaches for cyberbullying detection and classification on social media.

Akinyemi, J. D., Ibitoye, A. O. J., Oyewale, C. T., & Onifade, O. F. W. (2023). Cyberbullying detection and classification in social media texts using machine learning techniques. In Advances in Computer Science for Engineering and Education VI (pp. 440-449).

Shankar, K., Abirami, A. M., Indira, K., Angeline, C. V. N., & Shubhavya, K. (2022). Cyberbullying detection in social media using supervised ML and NLP techniques. In Communication and Intelligent Systems (pp. 817-828).

Dewani, A., Memon, M. A., & Bhatti, S. (2021). Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data. Journal of Big Data, 8(160).

**GitHub Link:**

<https://github.com/rpain1/Comparative-Analysis-of-NLP-Architectures-for-Cyberbullying-Detection->