CYBERBULLYING DETECTION IN SOCIAL MEDIA USING PRE-TRAINED LANGUAGE MODELS

Jasmeen Kah Ying Bong[1], Kasturi Dewi Varathan[2*], and Teoh Hwai Teng[3]

[1,2,3]Department of Information Systems,
Faculty of Computer Science & Information Technology,
Universiti Malaya, 50603 Kuala Lumpur, Malaysia

[2]Vel Tech Rangarajan Dr Sagunthala R and D Institute of Science and Technology,
Avadi, Chennai - 600062 India

Emails: jasmeenbky96@gmail.com[1], kasturi@um.edu.my[2*] (Corresponding Author), teoh0821@gmail.com[3]

ABSTRACT

The rapid integration of Information and Communication Technologies (ICT) has revolutionized online communication, yet it has also led to the emergence of cyberbullying, a harmful digital behaviour. This study addresses the urgency of combating cyberbullying and its negative impacts by using advanced pre-trained language models (PLMs) through transfer learning in detecting cyberbullying in social media. The goal is to enhance cyberbullying detection's effectiveness to create safer online spaces. Cyberbullying detection model using transfer learning, DistilBERT, DistilELECTRA, and MiniLM PLMs were explored. The PLMs' evaluation using the AMiCA dataset, MiniLM achieves the highest performance in detecting cyberbullying, with an accuracy of 97.84% in cross-validation and 98.57% in hold-out testing, while DistilBERT and DistilELECTRA also perform well, achieving accuracies of 97.34% and 98.03%, and 97.58% and 92.97%, respectively. MiniLM consistently maintains competitive F-measures, addressing class imbalance. Overall, MiniLM stands out with high accuracy and micro F1-scores, outperforming other models. Comparative analysis reaffirms MiniLM's excellence in binary classes and overall evaluation showcasing the effectiveness of transfer learning compared to previous studies. In conclusion, this study demonstrates the capabilities of PLMs for cyberbullying detection and suggests future research directions

Keywords: Cyberbullying Detection, Transfer Learning, Pre-trained Language Models, AMiCA Dataset, Text Classification

1.0 INTRODUCTION

The rise of Information and Communication Technologies (ICT) has significantly impacted social communication. It has transformed the way we connect with others, making communication more accessible. This widespread use of social media platforms allows people to connect with others from all over the world, creating a sense of shared global community. As a result, individuals from different parts of the world can interact, exchange ideas, and build connections that transcend geographic boundaries. The World Bank shows that by the end of 2022, 4.66 billion people would have been online, accounting for 60% of the world's population[1]. Most of these users are from Asia. The number of internet users in Malaysia increased from 88.7% in 2020 to 92.7% in 2022, as reported in the Internet Users Survey 2022 [2]. According to the same survey, 24.6 million Malaysians used social networking sites, with Facebook, Instagram, and YouTube being the most popular platforms. With the increasing use of technology, people are experiencing more problems related to the internet, including excessive social media use and online harassment.

Within the realm of ICT's benefits, a darker facet emerges as individuals exploit technological progress for abusive behavior, exemplified in cases of cyberbullying [3]. The term “Cyber” emphasizes the digital nature of the behavior, highlighting its occurrence through electronic communication channels. "Cyberbullying" underscores that the behavior takes place within the realm of the internet, encompassing diverse digital formats. The inclusion of "bullying" underscores the continuum between traditional bullying and its digital counterpart. While the medium has changed, the fundamental nature of the behavior remains akin.

Amid the advantages of ICT's integration, cyberbullying has emerged as a concerning phenomenon [3]. Social media platforms, which have become virtual hubs for social interaction, have unfortunately also transformed into arenas for bullying. The digital environment allows perpetrators to conceal their identities, complicating the detection of cyberbullying—an intricate challenge in safeguarding the integrity of online communities. As noted by [4], the rise in Internet usage has directly contributed to a surge in cyberbullying incidents, with significant public health consequences, including mental, psychological, and social issues [5]. In the context of social media platforms, 79 % of young adults faced cyberbullying on YouTube, 69 % on Snapchat, 64 % on TikTok, and 50 % on Facebook [6]. Cyberbullying poses a severe risk to public health and can have long-lasting impacts that often last into early adulthood, according to [7]. The consequences of cyberbullying are far-reaching and deeply concerning. Those subjected to cyberbullying often experience severe mental health challenges, such as depression, anxiety, feelings of isolation, and a diminished caKeepacity to experience pleasure (anhedonia) [3]. These distressing outcomes underscore the urgent need to address cyberbullying as a pivotal concern within the realm of online interactions.

Recent research suggests that pre-trained language models show significant promise for detecting cyberbullying due to their ability to capture nuanced linguistic patterns for large datasets [3][8]. These models including BERT, RoBERTa, and DistilBERT, have demonstrated impressive performance in other domains of natural language processing. Despite their potential, the application of pre-trained language models in cyberbullying detection remains underexplored, with only a few studies, such as those by [3] and [8], delving into this area. These works have shown the fine-tuned models outperform traditional approaches, achieving high accuracy and F1 scores. However, challenges such as data diversity, contextual nuances, and ethical implications of automated detection remains barriers to widespread implementation.

This study seeks to bridge the gap by investing the efficacy of pre-trained language models in cyberbullying detection. To address these gaps, this study investigates how different pre-trained language models can be applied and fine-tuned for specific datasets and contexts. By systematically experimenting with transfer learning methods, this research aims to optimize detection mechanisms and address growing concern of cyberbullying in online spaces. Pre-trained models such as DistilBERT, DistilELECTRA and MiniLM are examined to their suitability and efficiency or cyberbullying detection.

The observed disparity in performance between transfer learning models and other approaches has ignited a curiosity to dissect the intricacies underlying this discrepancy. The prospect of leveraging different pre-trained models offers a new avenue for investigation, presenting an opportunity to refine and enhance transfer learning's potential in addressing the complexities of cyberbullying detection. Section 2 is catered for related works on cyberbullying detection. Followed by section 3 which is on methodology. Section 4 is catered for results obtained from the models and section 5 emphasize on the discussions. Finally, section 6 concludes the research with summary of findings and future works

2.0 RELATED WORKS

In recent years, the detection of cyberbullying has gained significant attention, prompting researchers to explore diverse techniques for effective identification. Several studies have delved into the realm of conventional machine learning, deep learning and transfer learning approaches to tackle the intricate task of cyberbullying detection.

2.1 Analysis of Cyberbullying Detection Research using Conventional Machine Learning and Deep Learning Approach

Cyberbullying detection research has seen significant advancements through both conventional machine learning and deep learning approaches. Several studies have contributed to this field, each employing diverse techniques and evaluation metrics. Table 1 shows the analysis of past research on cyberbullying detection using conventional machine learning and deep learning approach.

The field of cyberbullying detection in social media has been explored using various machine learning and deep learning techniques. Alduailaj and [9] utilized SVM and Naïve Bayes for detecting cyberbullying, with SVM achieving an accuracy of 95.74% on a dataset of 30,000 tweets sourced from Twitter. Similarly, [10] applied SVM, KNN, and Naïve Bayes, with SVM using the RBF kernel achieving an F1 score of 0.92 on a smaller dataset of 652 tweets. [11] employed a wide range of classifiers, including Logistic Regression and Linear SVC, achieving high performance metrics (accuracy and F1 score of 0.96) using Bag of Words and TF-IDF features across a large dataset of 160,000 samples.

[12] explored Gaussian Naïve Bayes, Logistic Regression, Random Forest (RF), and other classifiers, with RF achieving an accuracy and F1 score of 0.92 on a dataset of 20,001 tweets. [13] explored Logistic Regression and Multinomial Naïve Bayes using character-level TF-IDF n-grams, with Logistic Regression achieving slightly better performance (F1 score of 0.798) on a dataset of 7,625 samples.

Table 1: Analysis of Cyberbullying Detection Research Using Conventional Machine Learning and Deep Learning approach

Reference	Technique	Best Technique	Evaluation	Dataset	Research Gap
[9]	SVM, NB	SVM	Accuracy: 0.96	Twitter (Twitter API, uploaded to Kaggle) **Size :** 30,000 **Availability:** https://www.kaggle.com/datasets/alanoudaldealij/arabic-cyberbullying-tweets	1. Only focus on Arabic dataset 2. Only focus on Text Analysis 3. Only focus on single-platform dataset 4. Only focus on Machine Learning, no deep learning
[10]	SVM, KNN, NB	SVM with RBF Kernel	F1 Score: 0.92	**Source:** Twitter (Twitter API) **Size:** 652 **Availability**: NA	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis 4. Small dataset size 5. Only focus on Machine Learning, no deep learning
[11]	Logistic regression, Decision tree, Gradient Boosting, Random-forest, Bagging, SGD, Linear SVC, AdaBoost	Linear SVC and Logistic Regression	**Logistic Regression** Bag of Words Accuracy: 0.96 F1 Score: 0.96 **TF-IDF** Accuracy: 0.96 F1 Score: 0.95 **Linear SVC** Bag of Words Accuracy: 0.95 F1 Score: 0.95 **TF-IDF** Accuracy: 0.96 F1 Score: 0.96	**Source:** Kaggle, Twitter, Wikipedia, YouTube **Size:** 160,000 **Availability**: NA	1. Only focus on English dataset 2. Only focus on Text Analysis
[12]	Gaussian Naïve Bayes, Logistic Regression, Decision Tree, RF, AdaBoost	RF	Accuracy: 0.92 Precision: 0.92 Recall: 0.92 F-1 Score: 0.92	**Source:** Twitter (Kaggle) - Cyberbullying **Size :** 20,001 **Availability:** https://github.com/aymeam/Datasets-for-Hate-Speech-Detection/blob/master/README.md	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis

Table 1: Continued

Reference	Technique	Best Technique	Evaluation	Dataset	Research Gap
[13]	XGBoost, Extr Tree Classifier, K-Nearest Neighbours, Logistic Regression, RF, Linear SVC, Decision Tree, Multinomial NB	LR and Multinomial NB	**Char TF-IDF Bi-gram** **1) LR:** Accuracy: 0.75 F1 Score: 0.80 Training Time: 0.96 **2) Multinomial** Accuracy: 0.75 F1 Score: 0.79 Traning Time: 0.00 **Char TF-IDF Trii-gram** **1) LR:** Accuracy: 0.75 F1 Score: 0.80 Training Time: 1.00 **2) Multinomial** Accuracy: 0.75 F1 Score: 0.80 Traning Time: 0.00	**Source:** Twitter (snscrape) **Size :** 7,625 **Availability:** NA	1. Only focus on single-platform dataset 2. Only focus on Urdu dataset 3. Only focus on Text Analysis
[14]	SVM, Random Forest, NB, Decistion Tree	NB	Precision: 0.85 Recall: 0.79 F Score: 0.82 Accuracy: 0.83	**Source:** Twitter (Kaggle) - Cyberbullying **Size :** 20,001 **Availability:** https://github.com/aymeam/Datasets-for-Hate-Speech-Detection/blob/master/README.md	1. Only focus on English dataset 2. Only focus on single platform dataset 3. Only focus on Text Analysis 4. Only focus on Machine Learning, no deep learning
[15]	Naives Bayes, Logistics Regression, SVM, RF, XGBoost, LSTM, GRU	GRU	Accuracy: 0.92 Precision: 0.92 Recall: 0.92 F-1 Score: 0.92	**Source:** Twitter (Kaggle) **Size :** 47,694 **Availability:** https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification	1. Only focus on English dataset 2. Only focus on single platform dataset 3. Only focus on Text Analysis
[16]	SVM, DT	DT	Accuracy: 94.42 Precision: 42.33 Recall: 30.24 F1 Score: 35.28 AUC: 73.09	**Source:** Ask.fm (AMiCA) **Size:** 113,698 – English, 78387 - Dutch **Availability**: NA	1. Only focus on English dataset 2. Only focus on single platform dataset 3. Only focus on Text Analysis

Table 1: Continued

Reference	Technique	Best Technique	Evaluation	Dataset	Research Gap
[17]	Logistic Regression, SVM, KNN, RF	Logistic Regression	Accuracy: 0.92 Precision: 0.87 Recall: 0.96 F-1 Score: 0.91	**Source**: Automated Hate Speech Detection and the Problem of Offensive Language – Twitter (tdavidson/hate-speech-and-offensive-language), Toxic Comment Classification - Wikipedia (Kaggle), Twitter Sentiment Dataset (Kaggle) **Size:** Automated Hate Speech Detection and the Problem of Offensive Language – 24,783 Toxic Comment Classification – 159,571 Twitter Sentiment Dataset – 162,973 **Availability:** Automated Hate Speech Detection and the Problem of Offensive Language https://github.com/t-davidson/hate-speech-and-offensive-language Toxic Comment Classification https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge Twitter Sentiment Dataset https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset	1. Only focus on English dataset 2. Only focus on Text Analysis Only focus on Machine Learning, no deep learning
[18]	Decision Tree, NB, RF, XgBoost, SVM, SVM (rbf), Logistic Regression, GloVe	GloVe840 embedding with BLSTM	Accuracy: 0.9260 F1 Score: 0.9420	**Source:** Twiiter **Size :** 35,787 **Availability:** https://github.com/aymeam/Datasets-for-Hate-Speech-Detection/blob/master/README.md	1. Only focus on English dataset 2. Only focus on single platform dataset 3. Only focus on Text Analysis

Table 1: Continued

Reference	Technique	Best Technique	Evaluation	Dataset	Research Gap
[19]	SVM, Logistic Regression, KNN, Naïve Bayes, AdaBoost, RF	AdaBoost	1st Variant Accuracy: 0.89 Precision: 0.84 Recall: 0.86 F1-Score: 0.88 2nd Variant Accuracy: 0.90 Precision: 0.90 Recall: 0.88 F1-Score: 0.89	**Source:** Twitter (API) **Size :** 5,000 **Availability: NA**	1. Only focus on English dataset 2. Only focus on single platform dataset 3. Only focus on Text Analysis
[20]	Logistic Regression, RF, XGBoost, LSTM-CNN	LSTM-CNN	Precision: 0.76 Recall: 0.31 ROC AUC: 0.97 F1 Score: 0.44 Accuracy: 0.952	**Source:** Twitter (Kaggle) + YouTube **Size :** Twitter - 1.6 Million YouTube - 1,000 **Availability:** https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge	1. Only focus on English dataset 2. Only focus on Text Analysis
[21]	Linear SVM, Logistic Regression, RF, Multi Layered Perceptron	Linear SVM	TF-IDF: Twitter Dataset F1 Score: 0.94 Wikipedia Dataset: F1 Score: 0.84	**Source:** Wikepedia (figshare), Twitter **Size :** Wikepedia (figshare) – 40,000 Twitter - 35,787 **Availability:** Wikepedia (figshare) https://figshare.com/articles/dataset/Wikipedia_Talk_Labels_Personal_Attacks/4054689 Twitter https://github.com/aymeam/Datasets-for-Hate-Speech-Detection/blob/master/README.md	1. Only focus on English dataset Only focus on Text Analysis
[22]	RF, AdaBoost, Extra Tree, Linear SVC, Logistic Regression	Dual Model	Performance: 0.42	**Source:** Vine **Size :** 436,000 **Availability:** NA	1. Only focus on single platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis

Table 1: Continued

Reference	Technique	Best Technique	Evaluation	Dataset	Research Gap
[23]	NB, KNN	KNN	With Chi-Square: Non-Bullying Precision: 0.73 Recall: 0.99 F1-Score: 0.84 Bullying Precision:0.69 Recall: 0.07 F1-Score: 0.13	**Source:** Facebook **Size :** 8,818 **Availability:** https://github.com/joshimiloni/Cyber-Bullying-Detection	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis
[24]	RF, NB & J48 - Baseline - Baseline + Personalities - Baseline + Personalities + Sentiment - Baseline + Personalities + Sentiment + Emotion - Baseline + Sentiment: - Baseline + Sentiment + Emotion	Baseline + Personalities + Sentiment	Accuracy: 0.92 F1-Score: 0.92	**Source:** Twitter (Twitter API) + Twitter hate speech (Kaggle) **Size :** 9,484 **Availability:** NA	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis 4. Only focus on Machine Learning, no deep learning
[25]	Logistic Regression	Logistic Regression	Kasture dataset Cyberbullying – 376 Non Cyberbullying – 9 37 Search API Cyberbullying – 21042 Non-Cyberbullying – 33852	**Source:** Twitter + Twitter (Search API) **Size :** Kasture - 1,313 Search API - 54,894 **Availability:** Twitter https://research.cs.wisc.edu/bullying/data.html Twiiter (Search API) https://chatcoder.com/	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis 4. Only focus on Machine Learning, no deep learning

Table 1: Continued

Reference	Technique	Best Technique	Evaluation	Dataset	Research Gap
[26]	SVM, KNN, NB	SVM with RBF Kernel	Mean Accuracy: 0.86	**Source:** Twitter (Twitter REST API) **Size :** 652 **Availability:** NA	1. Only focus on single-platform dataset 2. Only focus on Sinhala dataset 3. Only focus on Text Analysis 4. Only focus on Machine Learning, no deep learning
[27]	SVC, Logistic Regression, Multinomial NB, RF, SGD	Logistic Regression	Accuracy: 0.93 Precision: 0.91 Recall: 0.96 F1-Score: 0.93	**Source:** Twitter (Twitter API) + Twitter hate speech (Kaggle) **Size :** 2000 **Availability:** Twitter (Twitter API) NA Twitter hate speech (Kaggle) https://www.kaggle.com/vkrahul/twitter-hate-speech	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis 4. Small dataset size 5. Only focus on Machine Learning, no deep learning
[28]	RF - Baseline - Baseline + Big 5 + Dark Triad - Baseline + Trait - Baseline + Key Traits	Baseline + Key Traits	Precision: 0.96 Recall: 0.95 F-Measure: 0.93	**Source:** Twitter **Size :** 9,484 **Availability:** NA	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis 4. Only focus on Machine Learning, no deep learning
[29]	SVM, Neutral Network model	Neural Network Model	Accuracy: 0.93 F1-Score: 0.92	**Source:** Formspring.me **Size :** 12773 **Availability:** NA	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis
[30]	SVM	SVM	F1 Score: 0.75	**Source:** Facebook (Facebook API) **Size :** 1,182 **Availability:** NA	1. Only focus on single-platform dataset 2. Only focus on Myanmar dataset 3. Only focus on Text Analysis 4. Small dataset size

Table 1: Continued

Reference	Technique	Best Technique	Evaluation	Dataset	Research Gap
[31]	Liner SVM, Logistic Regression, Decision Tree, RF, Gradient Boosting, Regression Tree Multilayer perceptron	Logistic Regression and Gradient Boosting Regression Tree for all features	**All Features** **Logistic Regression** Accuracy: 0.93 Precision: 0.93 Recall: 0.94 F-Measure: 0.94 **Gradient Boosting Re gression Tree** Accuracy: 0.93 Precision: 0.92 Recall: 0.95 F-Measure: 0.94	**Source:** Twitter **Size :** 2,349,052 **Availability:** NA	1. Only focus on single-platform dataset 2. Only focus on Japanese dataset 3. Only focus on Text Analysis
[32]	SVM, CNN-CB	CNN-CB	Accuracy: 0.95	**Source:** Twitter (Twitter streaming API) **Size :** 39,000 **Availability:** NA	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis
[33]	Linear SVM	Linear SVM	Engligh F1 Score: 0.64 Dutch F1 Score: 0.62	**Source:** Ask.fm (AMiCA) **Size:** 113,698 – English, 78387 - Dutch **Availability**: NA	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis
[34]	Naives Bayes, SVM	SVM	Precision: 0.93 Recall: 0.94 F-Measure: 0.93	**Source:** Twitter (Twitter Scrapper) & Facebook (Facebook Scrapper) **Size:** Arabic - 35,273 English - 91,431 **Availability**: NA	1. Only focus on single-platform dataset 2. Only focus on Arabic dataset 3. Only focus on Text Analysis
[35]	BWM, BoW, Semantic-enhanced BoW, LSA, LDA, mSDA, smSDA smSDAu	smSDA	Twitter: Accuracies: 0.85 F1 scores: 0.72 MySpace Accuracies: 0.90 F1 Scores: 0.78	**Source:** Twitter (Twitter API) + MySpace **Size:** Twitter (Twitter API) - 7,321 My Space - 1,753 **Availability**: Twitter - https://research.cs.wisc.edu/bullying/data.html My Space - https://chatcoder.com/	1. Only focus on English dataset 2. Only focus on Text Analysis

Table 1: Continued

Reference	Technique	Best Technique	Evaluation	Dataset	Research Gap
[36]	RF, SVM, Multiplayer Perceptron, J48 Decision Tree, CNN Pre-trained, CNN Random, PCNN	PCNN	Twitter Dataset with TM CFA: Accuracy: 0.99 Precision: 0.97 Recall: 0.98 F1-Score: 0.98 Formspring.me Dataset with TM CFA: Accuracy: 0.88 Precision: 0.25 Recall: 0.79 F1-Score: 0.38	**Source:** Twitter + Formspring.me **Size:** Twitter – 1,313 Formspring.me – 13,000 **Availability**: NA	1. Only focus on English dataset Only focus on Text Analysis

[17] demonstrated that Logistic Regression outperformed SVM, KNN, and RF in detecting cyberbullying, achieving an F1 score of 0.91 on datasets aggregated from Twitter and Wikipedia. [14] utilized SVM, RF, Decision Tree and Naïve Bayes, with NB achieving a balanced accuracy of 0.83 on 20,001 tweets. [15a] experimented with multiple models, highlighting GRU's consistent performance with an F1 score of 0.92 on a dataset of 47,694 tweets. [16] used Decision Trees and SVM for cyberbullying detection, with Decision Trees achieving an accuracy of 94.42% on a dataset sourced from Ask.fm.

[18] applied GloVe embedding with BLSTM, achieving an F1 score of 0.942 on datasets aggregated from Twitter, highlighting the effectiveness of embeddings for detecting offensive language. [19] tested AdaBoost and RF, with AdaBoost's second variant achieving an F1 score of 0.894 on a small dataset of 5,000 tweets. [20] proposed an LSTM-CNN hybrid model, achieving an accuracy of 95.2% on a large Twitter and YouTube dataset. [21] highlighted Linear SVM's superiority in detecting hate speech using TF-IDF features, achieving an F1 score of 0.939 on datasets from Twitter and Wikipedia.

[22] proposed a dual model for cyberbullying detection, though achieving moderate performance (0.42). [23] utilized KNN with Chi-Square, achieving higher precision for non-bullying instances but struggled with identifying bullying instances effectively. [24] emphasized the importance of incorporating user personality traits in improving detection accuracy, achieving 91.88% accuracy. [25] showcased Logistic Regression's high performance in datasets sourced from Twitter APIs, with balanced metrics across datasets. [26] compared SVM, KNN, and Naïve Bayes for cyberbullying detection, achieving a mean accuracy of 85.28% with SVM using the RBF kernel on a small dataset of 652 tweets. [27] explored several classifiers, with Logistic Regression achieving the best results (F1 score of 0.93) on a dataset of 2,000 tweets aggregated from Twitter and Kaggle.

[28] enhanced cyberbullying detection by incorporating psychological traits into Random Forest models, achieving high performance metrics (precision of 0.960 and recall of 0.952) on a dataset of 9,484 tweets. [29] proposed a neural network model, which outperformed SVM, achieving an accuracy of 92.8% and an F1 score of 0.919 on a dataset sourced from Formspring.me. [30] applied SVM for cyberbullying detection on Facebook posts, achieving an F1 score of 0.754 on a small dataset of 1,182 entries.

[31] compared Logistic Regression, Gradient Boosting Regression Tree, and other models, with both Logistic Regression and Gradient Boosting performing exceptionally well (F1 score of 0.935) on a massive dataset of over 2.3 million entries aggregated from Twitter. [32] utilized CNN-CB (Convolutional Neural Network for Cyberbullying) and achieved an accuracy of 95% on a dataset of 39,000 tweets sourced from the Twitter streaming API. [33] employed Linear SVM for detecting cyberbullying in English and Dutch datasets from Ask.fm, achieving moderate F1 scores of 0.64 (English) and 0.62 (Dutch).

[34] demonstrated the strong performance of SVM for cyberbullying detection, achieving high precision (0.934) and recall (0.941) on datasets aggregated from Twitter and Facebook. [35] explored various machine learning techniques for cyberbullying detection, including Bag of Words (BoW), Semantic-enhanced BoW, Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and multiple Stacked Denoising Autoencoders (mSDA, smSDA, smSDAu). Among these methods, smSDA (semantic-enhanced Stacked Denoising Autoencoder) achieved the best results, with accuracies of 84.9% (Twitter dataset) and 89.7% (MySpace dataset), alongside F1 scores of 0.719 and 0.776, respectively. [36] explored CNN and PCNN, achieving exceptional accuracy of 0.990 on a Twitter dataset, highlighting the potential of neural network models for cyberbullying detection.

Through these diverse approaches, researchers have explored the intricacies of cyberbullying detection, harnessing the capabilities of both conventional machine learning and deep learning techniques to develop effective and robust detection models. However, differences in dataset sizes, sources, and evaluation metrics highlight the need for standardized evaluation methods to better compare the efficacy of these models.

2.2 Analysis of Cyberbullying Detection with Transfer Learning Approach

The advancement of cyberbullying detection techniques through transfer learning has garnered significant attention in recent research. Different studies have explored the effectiveness of various transfer learning models and techniques in enhancing cyberbullying detection accuracy and performance. Table 2 shows the analysis of cyberbullying detection research using transfer learning approach. [37] tackled cyberbullying, spamming and multilingual detection using models ranging from Gaussian Naïve Bayes (GNB) and Random Forest (RF) to BERT and MBERT. They reported RF as optimal for cyberbullying with an accuracy of 91.77%, BERT for spamming with an accuracy of 97%, and MBERT for cyberbullying detection in Tanglish with an accuracy of 75%.

[8] explored transformer-based models such as BERT, XLNet, RoBERTa, and XLM-RoBERTa, highlighting RoBERTa as the most effective with an F1-score of 0.87. [3] employed a range of techniques, including LR, Linear SVC, DistilBert, DistilRoBerta, and Electra-small, with their fine-tuned DistilBert yielding impressive results after preprocessing steps such as noise removal, normalization, cleaning, sentiment analysis, and word embedding. This approach achieved an accuracy of 97% and an F1 score of 0.68 through cross-validation, and even higher performance with a hold-out method.

[38] focused on detecting slang-based cyberbullying using custom models, GloVe embeddings, and BERT. Their BERT model achieved an F1-score of 0.72, with high precision and accuracy. [39] employed a diverse range

of techniques, including SVM, LR, CNN, LSTM, TF-IDF, BERT, VecMap, and more. Their combination of BERT and VecMap in a CNN architecture yielded promising results, with accuracies of 81% for the Twitter dataset and 62% for the Facebook dataset. [40] introduced the HateBERT model through GDPR-compliant preprocessing, including anonymization, normalization, and entity extraction. Their approach achieved notable results, with F1 scores of 0.76, 0.63, and 0.68 for the UC, QA, and Twitter datasets, respectively.

[41] explored cross-platform cyberbullying detection by employing BiL–TM, BERT, RoBERTa, XP-CB-BERT, and XP-CB-RoBERTa models. Their XP-CB-RoBERTa technique integrated various preprocessing steps, including noise removal, normalization, final cleaning, sentiment analysis, and word embedding. This cross-platform approach resulted in a macro-average F1 score of 0.69, showcasing the effectiveness of transfer learning across different platforms. [42] compared various BERT model sizes and reported BERT-Base as the most effective with an F1-score of 0.91 and AUC of 0.97 on a Twitter dataset from [43].

[44] focused on leveraging transfer learning with BERT to address cyberbullying detection. They conducted thorough preprocessing steps involving tokenization, lowercasing, stemming, stop words removal, punctuation, and more. Their BERT-based approach showcased commendable performance, achieving F1 scores of 0.75(racism tweets) and 0.76 (sexism tweets) for detecting racism and sexism on Twitter, respectively. Similarly, Kaggle dataset experiments demonstrated an F1 score of 0.76 for detecting insults. [45] harnessed the power of text and image embeddings, leveraging RoBERTa and Xception models, which resulted in impressive recall (0.92) and F1-score (0.86) values. [46] focused on BERT for cyberbullying detection, showcasing remarkable accuracy with 0.98 for the Formspring.me dataset and 0.96 for Wikipedia.

These studies collectively underscore the significance of transfer learning in enhancing cyberbullying detection performance. By leveraging pre-trained language models and well-designed preprocessing steps, researchers have successfully addressed the challenges of cyberbullying detection across various platforms and datasets, offering insights into effective techniques for combating online harassment.

3.0 METHODOLOGY

Figure 1 displays the structure outlining the process of employing transfer learning methodologies for the purpose of detecting cyberbullying (binary text classification). The research framework employed follows the General Research Approach process by [47] which is the process employed by most machine learning-based studies.

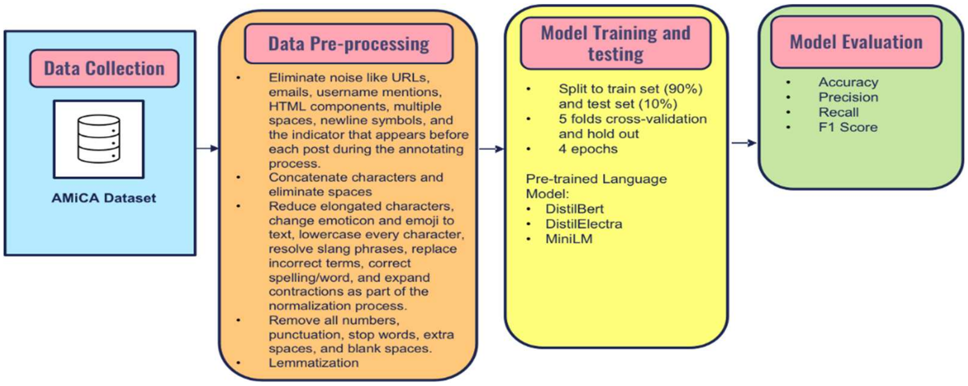
Data Collection
AMiCA Dataset
Data Pre-processing
• Eliminate noise like URLs, emails, username mentions, HTML components, multiple spaces, newline symbols, and the indicator that appears before each post during the annotating process.
• Concatenate characters and eliminate spaces
• Reduce elongated characters, change emoticon and emoji to text, lowercase every character, resolve slang phrases, replace incorrect terms, correct spelling/word, and expand contractions as part of the normalization process.
• Remove all numbers, punctuation, stop words, extra spaces, and blank spaces.
• Lemmatization
Model Training and testing
• Split to train set (90%) and test set (10%)
• 5 folds cross-validation and hold out
• 4 epochs
Pre-trained Language Model:
• DistilBert
• DistilElectra
• MiniLM
Model Evaluation
• Accuracy
• Precision
• Recall
• F1 Score

Fig. 1: Model for Identifying Cyberbullying Occurrences

Table 2: Analysis of Cyberbullying Detection Research using Transfer Learning approach

Reference	Technique	Best Technique	Evaluation	Dataset	Research Gap
[37]	**For Cyberbullying:** Gaussian Naïve Bayes, Logistic Regression, Decision Tree, RF, AdaBoost **For Spamming:** BERT, Gaussian NB, Logistic Regression **For Cyberbullying in Tanglish:** MBERT	**For Cyberbullying:** Random Forest **For Spamming:** BERT **For Cyberbullying in Tanglish:** MBERT	**For Cyberbullying** RF Accuracy: 0.91 **For Spamming**: BERT Accuracy: 0.97 **For Cyberbullying in Tanglish:** MBERT Accuracy: 0.75	**Source:** Kaggle (Tweets Dataset for Detection of Cyber-Trolls) & Dravidian code-mix HAVOC 2021 **Size:** 19,999 **Availability:** Kaggle (Tweets Dataset for Detection of Cyber-Trolls https://www.kaggle.com/datasets/dataturks/dataset-for-detection-of-cybertrolls Dravidian code-mix HAVOC 2021 https://github.com/dravidian-codemix/HASOC-2021	1. Only focus on single-platform dataset 2. Only focus on Text Analysis
[8]	BERT, XLNet, RoBERTa, XLM-RoBERTa	RoBERTa	F1 Score: 0.87	**Source:** Formspring.me, Twitter **Size:** Formspring.me – 12,773 Twitter – 47,692 **Availability**: Formspring.me https://github.com/sweta20/Detecting-Cyberbullying-Across-SMPs Twitter https://drive.google.com/drive/folders/1oB2fan6GVGG83Eog66Ad4wK2ZoOjwu3F	1. Only focus on English dataset 2. Only focus on Text Analysis
[38]	Custom Model, GloVe Twitter, BERT	BERT	**Three dataset with SLANG** Precision: 0.67 F1 Score: 0.72 Accuracy: 0.84 AUC:0.88	**Source:** Facebook **Size:** 15,000 **Availability**: NA	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis

Table 2: Continued

Reference	Technique	Best Technique	Evaluation	Dataset	Research Gap
[3]	LR, Linear SVC, DistilBert, DistilRoBerta, Electra-small	Fine-tuned DistilBert	**Cross-validation** Accuracy – 0.97 Precision – 0.76 Recall – 0.62 F1 Score – 0.68 **Hold-out** Accuracy – 0.97 Precision – 0.74 Recall – 0.71 F1 Score – 0.72	**Source:** Ask.fm (AMiCA) **Size:** 113,698 – English, 78387 - Dutch **Availability**: NA	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis
[39]	SVM, LR, CNN, LSTM Combine with TF-IDF, BERT, VecMap, BERT+VecMap, BERT + VecMap-CNN	BERT + VecMap-CNN	**Twitter Dataset:** Accuracy: 0.81 +/- 0.0047, F1-Score: 0.81 +/- 0.0052 **Facebook Dataset:** Accuracy: 0.62 +/- 0.0056, F1-Score: 0.62 +/- 0.0053	**Source:** Facebook **Size:** 15,000 **Availability**: NA	1. Only focus on single-platform dataset 2. Only focus on Hindi-English dataset 3. Only focus on Text Analysis
[40]	BERT, HateBERT, Bi-LSTM, SVM	HateBERT	**Train set Dataset – UC** UC F1 Score: 0.76 QA F1 Score: 0.63 Twitter F1 Score: 0.68 **Train set Dataset – QA** UC F1 Score: 0.69 QA F1 Score: 0.73 Twitter F1 Score: 0.63 **Train set Dataset – Twitter** UC F1 Score: 0.68 QA F1 Score: 0.67 Twitter F1 Score: 0.81	**Source:** User Comments (UC) + QA + Twitter **Size:** UC – 249,123 QA – 129,501 Twitter – 12,310 **Availability**: NA	1. Only focus on English dataset 2. Only focus on Text Analysis

Table 2: Continued

Reference	Technique	Best Technique	Evaluation	Dataset	Research Gap
[41]	BiL–TM, BERT, RoBERTa, XP-CB-BERT, XP-CB-RoBERTa	XP-CB-RoBERTa	**Cross-platform Macro average** F1 Score: 0.693	**Source:** Twitter **Size:** 16,090 **Availability**: https://github.com/sweta20/Detecting-Cyberbullying-Across-SMPs	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis
[42]	BERT-Base, BERT-Medium, BERT-Small. BERT-Mini, BERT-Tiny	BERT-Base	Accuracy – 0.92 AUC – 0.97 Precision – 0.91 Recall – 0.92 F1-Score – 0.91	**Source:** Twitter (Founta et al., 2018)[43] **Size:** 85,948 **Availability:** NA	1. Only focus on single-platform dataset 2. Only focus on English dataset 3. Only focus on Text Analysis
[44]	UD, SSWE, Glv-Twtr, Glv-CC, RI, W2V, Glv-WK, BERT, LR, MLP, LSTM, Bi-LSTM	BERT	**Twitter (Racism)** F1 score: 0.75 **Twitter (Sexism)** F1 score: 0.76 **Kaggle (Insults)** F1 Score: 0.76	**Source:** Twitter (Racism), Twitter (Sexism), Kaggle (Insult) **Size:** Twitter (Racism) – 13,741 Twitter (Sexism) – 14,881 Kaggle (Insult) – 7,557 **Availability:** https://github.com/aymeam/Datasets-for-Hate-Speech-Detection	1. Only focus on Text Analysis
[45]	RoBERTa and Xception	RoBERTa and Xception	Recall:0.92 F1-Score:0.86	**Source:** Facebook, Instagram and Twitter **Size :** 2,100 **Availability:** NA	1. Only focus on English dataset 2. Only focus on Text Analysis
[46]	BERT	BERT	**Formspring.me** Accuracy – 0.98 **Wikipedia** Accuracy – 0.96	**Source:** Formspring.me, Wikipedia **Size:** Formspring.me – 12,773 Wikipedia – 115,864 **Availability:** https://github.com/sweta20/Detecting-Cyberbullying-Across-SMPs	1. Only focus on English dataset 2. Only focus on Text Analysis

3.1 Data Collection

Ask.fm (AMiCA), which was collected from the Ask.fm platform by [33], served as the basis for the dataset used in this study. According to [48], Ask.fm, a popular site among teenagers, has become a focus topic for research on cyberbullying. The Automatic Monitoring for Cyberspace Applications (AMiCA) project in Belgium oversaw the data collecting, which took place from April to October 2013. This project started in 2018 and gained control of the dataset. Only the English component of the original dataset which included both English and Dutch was used in this analysis. The dataset comprises 113,694 posts, of which 108,319 are labeled as non-cyberbullying, and 5,375 are labeled as cyberbullying. Table 3 presents the class distribution of the data used for cyberbullying detection.

The AMiCA dataset was chosen out of a variety of publicly accessible cyberbullying-related datasets based on a number of criteria. It is notable for being an extensive open dataset that includes recent data [3]. Notably, the annotators consist of four qualified linguists in both English and Dutch with a track record of expertise. The technical rules with particular information that annotators should follow while labeling the data samples are what distinguishes the AMiCA dataset from others. The corpus included a broader range of cyberbullying topics, encompassing issues like curses, defamation, defense, insult, sexual content and threats.

Table 3: Class Distribution of AMiCA Dataset

Classes	Posts	%
Cyberbullying	5,375	4.7%
Non-cyberbullying	108,319	95.3%

3.2 Tools and Resources

Python 3.10, a versatile and open-source programming language with a large collection of easily accessible libraries, was used to compute the tasks in this study. Jupyter Notebook, an interactive web-based programming environment, was used to document the scrips. The source code is now accessible on a GitHub repository to guarantee reproducibility. High Performance Graphics Processing Units (GPU) were necessary for effective model training. Because Paperspace Gradient Notebooks offer complete access to JupyterLab, top-tier GPUs, lots of RAM, and availability, they were mostly used in the research for transfer learning tasks. With a basic free account, this platform is especially effective for deep learning applications. However, the transfer learning activities in this study required a Pro-plan subcriptions. An RTX 5000 GPU, 30 GB of RAM, 16 GB of dedicated GPU memory, and CUDA version 11.0 were all part of the system setup. The source code for this research is available on GitHub with the link https://github.com/JasmeenBongKahYing/Cyberbullying-Detection/

3.3 Data Preprocessing

It's crucial to prepare raw textual data before using it in applications for natural language processing. The basic input for any text-based application is this cleaned textual data [15b]. The Brat Repositories' stand-off document format was used to share the AMiCA dataset, which required further data processing [3]. The Brat Repositories document's text and annotation labels can now be extracted and integrated using new BratReader code. The pipeline for text preprocessing shown in Figure 1 was designed to handle text data processing. A Python package including different text preparation modules contained this pipeline. Preprocess_text can be found on the GitHub repository at this address: https://github.com/HwaiTengTeoh. Three unique steps make up the preprocessing process.

Preliminary text cleaning is done in the first stage to get rid of any noise in the data, including URLs, emails, username mentions, HTML components, too many spaces, newline characters, and special symbols for annotations. Additionally, single characters that are separated by spaces, such as 'W H A T,' are combined to form a unified whole.

The second stage then focuses on text normalization, which seeks to transform text into its fundamental form. To accomplish this, regular expressions are used to address redundantly represented words and reduce long letters (such as 'youuuuuuu'). Additionally, the 'emot' package converts embedded emojis and emoticons into textual representations. It is calculated how many emoticons and emojis each post has. To guarantee consistent word embeddings, all accented letters are transformed to lowercase and standardized to follow the English alphabet. A predetermined list drawn from internet chats, text messages, and social media platforms is used to resolve slang phrases. The slang can be accessed through the following links

Online chats slang: https://slang.net/terms/online_chat.
Text messaging slang: https://slang.net/terms/text_messaging.
Social media slang: https://slang.net/terms/social_media.

The challenges in cleaning up internet posts involve dealing with various complexities. For instance, there are spelling mistakes, slang terms, and words that are intentionally modified to bypass filters commonly used on online platforms. These manipulated words are strategically chosen to evade detection, making it difficult to effectively monitor and remove harmful content. To address these variations and substitute hidden profanity, a thorough preparation pipeline has been developed. Utilizing the 'FuzzyWuzzy' library to find the Levenshtein distance, which assesses string pattern similarity, is the basic idea. If 90% similarity criteria is met, the rule is set up to replace terms with their matching matches from the mapping dictionary. This level was found through experimentation to be ideal for handling the majority of instances. Furthermore, the free and open-source grammar and spelling checker LanguageTool assists in correcting word spellings. To establish a connection with the LanguageTool server, the 'language-tool-python' package is used. Using the 'pycontraction' package, which adapts based on contextual cues, contractions inside the text are expanded simultaneously.

The final stage involves eliminating numerical values, punctuation, and extra spaces. Intriguingly, stopwords are kept in the text because some words, including negations and pronouns, provide context for understanding cyberbullying. Then, except for pronouns, where the original form is maintained, lemmatization is used to transform words into their fundamental forms, ultimately lemmatizing to "be." The proposed order of preprocessing processes was found to be the most efficient one after extensive testing. Maintaining the order of the processes is crucial since skipping any steps could cause the preprocessing to break down altogether. For instance, removing punctuation too early might inadvertently affect emoticon structure, and removing numerals might alter slang terms containing numbers, such as '2day' becoming 'today' and '2morrow' becoming 'tomorrow'. Following text preparation, there are a total of 112, 247 postings after removing any empty text entries as the final stage.

3.4 Model Training and Testing

Deep learning-based PLMs support the idea of transfer learning, which is the process of using the knowledge gained or generated from specialized activities to successfully complete other tasks. A list of the default hyperparameter configurations used during the PLMs' fine-tuning process is presented in Table 4.

Table 4: Default Hyperparameter Configurations Applied during the Fine-Tuning Process of the PLMs

Hyperparameter	**DistilBERT**	**DistilELECTRA**	**MiniLM**
Repository path	distilbert-base	lsanochkin/distilelectra-base	sentence-transformers/all-MiniLM-L6-v2
Batch size for training	8	8	8
Batch size for testing	8	8	8
Maximum sequence length	512	512	512
Activation function	gelu	gelu	gelu
Learning rate	0.00005	0.00005	0.00005
Dropout Probability	0.1	0.1	0.1

Preprocessing played a pivotal role in aligning plain text with language model requirements, a crucial step to leverage transfer learning for enhancing the proficiency of DistilBERT, DistilELECTRA, and MiniLM PLMs in identifying cyberbullying within subsequent tasks. This preprocessing phase encompassed several key steps, including tokenization that partitions the text into individual tokens for each word, introducing the [CLS] token at the sentence's outset, appending the [SEP] token at the sentence's closure, and assigning unique identifiers to each token. These preparatory measures collectively optimized the textual data for effective utilization in the fine-tuning process, resulting in heightened cyberbullying detection capabilities across the chosen pre-trained models.The resulting 768-dimensional vector representations were created by coupling these token identifiers with their respective embeddings [3]. The embeddings of the [CLS] token underwent processing through a linear layer in order to make a prediction about the class when detecting cyberbullying from provided posts. The HuggingFace Transformers library provides access to pre-trained models and tokenizer classes for DistilBert, DistilElectra, and MiniLM [49].

The strategic choice of DistilBERT, DistilELECTRA, and MiniLM as the focal pre-trained language models (PLMs) in this research is grounded in their distinctive attributes, each contributing to the enhancement of cyberbullying detection through transfer learning. DistilBERT's size reduction and accelerated training speed make it a pragmatic option for processing extensive textual data efficiently, facilitating quicker experimentation.

DistilELECTRA's architecture, encompassing six layers, 768 dimensions, and 12 heads, combined with the innovative Replace Token Detection (RTD) technique, empowers the model to capture intricate nuances of abusive language beyond conventional masked language modeling (MLM) methods. While MiniLM's specifics might not be explicitly stated in the provided literature, its computational efficacy and versatility undoubtedly amplify the research's capability for fine-tuning diverse PLMs. By skillfully leveraging the distinct advantages of these models, the study achieves a comprehensive, efficient, and accurate approach to identifying cyberbullying instances within digital content.

3.5 Model Evaluation

The performance of various text categorization techniques was evaluated in this study using a comprehensive set of metrics: accuracy, precision, recall, and F1 score. These metrics were selected for their ability to capture different facets of a model's classification performance. Accuracy measures the proportion of correctly classified instances across all categories, providing an overall indication of model effectiveness. Precision quantifies the proportion of correctly identified positive instances among all instances predicted as positive, highlighting the reliability of positive predictions. Recall reflects the model's ability to identify all relevant positive instances, emphasizing sensitivity to positive cases. F1 score represent the harmonic mean of precision and recall, balancing these two metrics to provide a single measure of model performance.

4.0 RESULTS

This section presents a detailed analysis of the outcomes from evaluating various pre-trained language models (PLMs) for the detection of cyberbullying in textual data. The results are segmented into evaluations for the positive class (cyberbullying), the negative class (non-cyberbullying), overall performance, and a comparative analysis with previous research to contextualize the findings.

With a focus on optimizing efficiency, the research delved into the ramifications of fine-tuning these optimized PLMs. This strategic approach streamlined the training process, effectively circumventing complexities tied to their base PLM counterparts. The study revolved around the optimized versions—DistillBERT, DistilELECTRA, and the smaller-scaled MiniLM, serving as a point of comparison.

Acknowledging the effectiveness of transfer learning in mitigating data imbalance challenges, the fine-tuning procedure incorporated the original dataset size. With the retention of default hyperparameters, paramount attention was directed towards pinpointing the optimal number of epochs for fine-tuning, with a specific focus on achieving the highest F-measure during rigorous five-fold cross-validation. This critical determination, pivotal to our analysis, was discerned through an exhaustive evaluation across four epochs.

4.1 Performance Evaluation for Cyberbullying Class (Positive Class)

Table 5 provides performance evaluations for positive class of cyberbullying detection (cyberbullying class). Upon examining the outcomes meticulously presented in Table 5, this research unearthed a treasure trove of insights into cyberbullying detection. Of particular significance is the exceptional performance of the fine-tuned DistillBERT. This model exhibited its prowess with an overall F-measure of 69.57% during the five-fold cross-validation and an impressive 78.14% during the hold-out testing phase, both achieved at the fourth epoch.

Similarly, DistilELECTRA unveiled its potential, demonstrating peak F-measures of 72.30% and 81.00% for the cyberbullying class during five-fold cross-validation and hold-out testing, respectively, both observed at the fourth epoch. Furthermore, the smaller-scaled MiniLM showcased its aptitude, attaining peak F-measures of 76.01% and 84.41% for the cyberbullying class during five-fold cross-validation and hold-out testing, respectively, particularly evident at the third epoch. When viewed holistically, it becomes evident that MiniLM, despite its smaller scale, emerges as a performance exemplar, demonstrating competitive F-measure values across both cross-validation and hold-out testing scenarios. Its unwavering performance highlights its pivotal role in cyberbullying detection, particularly when navigating the intricate nuances of class imbalance.

4.2 Performance Evaluation for Non-Cyberbullying Class (Negative Class)

Table 6 provides the performance evaluation pertaining to the non-cyberbullying class (negative class). This comprehensive analysis sheds light on the effectiveness of various pre-trained language models (PLMs) when fine-tuned for the precise identification of content not associated with cyberbullying.

A particularly noteworthy standout in this evaluation is DistillBERT, which consistently demonstrates exceptional results across different epochs and evaluation methodologies. During the five-fold cross-validation process, DistillBERT achieves an impressive F-measure of 98.61% at the fourth epoch, underscoring its capability in recognizing instances attributed to the non-cyberbullying class. Importantly, during the hold-out testing phase, DistillBERT maintains its high standard, yielding an F-measure of 98.97%.

Table 5: Performance Evaluations for Cyberbullying Detection using PLMs (Cyberbullying Class)

		Cyberbullying class (positive class)					
		Cross-validation			Hold out		
PLM	Epoch Number	P	R	F	P	R	F
DistillBERT	1	76.37	59.74	66.38	78.41	71.56	74.83
	2	76.33	63.94	69.09	85.71	65.80	74.45
	3	75.93	63.98	69.27	80.59	76.39	78.44
	4	76.64	63.75	69.57	83.51	73.42	78.14
DistilELECTRA	1	77.34	63.35	69.50	85.49	79.93	82.61
	2	81.14	62.19	69.59	89.65	70.82	79.13
	3	79.19	66.17	72.04	85.60	77.32	81.25
	4	79.03	66.88	72.30	87.15	75.65	81.00
MiniLM	1	82.19	67.84	74.23	89.67	75.84	82.18
	2	80.96	70.26	74.91	91.11	78.07	84.08
	3	80.83	71.75	76.01	88.08	81.04	84.41
	4	79.53	72.45	75.78	90.08	79.37	84.39

Similarly, DistilELECTRA also delivers compelling outcomes, affirming its competence in accurately identifying content from the non-cyberbullying class. At the fourth epoch, DistilELECTRA achieves a robust F-measure of 98.74% during the five-fold cross-validation, further bolstering its efficacy. This remarkable performance extends seamlessly to the hold-out testing scenario, with DistilELECTRA sustaining its excellence and recording an F-measure of 99.11%.

Table 6: Performance Evaluations for Cyberbullying Detection using PLMs (Non-Cyberbullying Class)

		Non-cyberbullying class (negative class)					
		Cross-validation			Hold out		
PLM	Epoch Number	P	R	F	P	R	F
DistillBERT	1	98.00	99.04	98.52	98.57	99.01	98.79
	2	98.20	98.99	98.60	98.30	99.45	98.87
	3	98.20	98.93	98.56	98.81	99.07	98.94
	4	98.19	99.03	98.61	98.67	99.27	98.97
DistilELECTRA	1	98.17	99.08	98.63	98.99	99.32	99.15
	2	98.12	99.27	98.69	98.55	99.59	99.06
	3	98.31	99.14	98.72	98.86	99.35	99.10
	4	98.35	99.13	98.74	98.78	99.44	99.11
MiniLM	1	98.40	99.28	98.84	98.79	99.56	99.18
	2	98.52	99.19	98.85	98.90	99.62	99.26
	3	98.59	99.15	98.87	99.05	99.45	99.25
	4	98.62	99.08	98.85	98.97	99.56	99.26

Furthermore, MiniLM distinctively showcases its aptitude in detecting non-cyberbullying content. Across multiple epochs, MiniLM consistently achieves competitive F-measure values. Particularly remarkable are its F-measures of 98.87% and 98.85% during the third and fourth epochs, respectively, in the context of five-fold cross-validation.

In summary, the comprehensive analysis of these results underscores MiniLM's proficiency in identifying instances from the non-cyberbullying class due to its consistent and robust performance. DistilELECTRA and DistillBERT also make significant contributions, showcasing commendable abilities in effectively discerning content unrelated to cyberbullying.

4.3 Performance Evaluation of Overall Results

Table 7 outlines the comprehensive assessment of these PLMs across different epochs and evaluation metrics. Notably, DistillBERT emerged as a consistent performer, displaying robust outcomes throughout the evaluation process. Across various epochs, the model consistently achieved high accuracy, micro precision, micro recall, and micro F1-score. For instance, during five-fold cross-validation, DistillBERT exhibited an impressive accuracy of 97.34% at the fourth epoch, demonstrating its ability to make accurate predictions. This translated into notable micro precision, recall, and F1-score values of 87.42%, 81.39%, and 84.09%, respectively. These attributes persisted during the hold-out testing phase as well, further solidifying DistillBERT's proficiency.

Likewise, DistilELECTRA showcased its capabilities in overall cyberbullying detection. Its performance consistently maintained high accuracy levels, with micro F1-scores hovering around the 85-86% range. Particularly, during the fourth epoch of five-fold cross-validation, DistilELECTRA demonstrated a micro F-score of 85.52%, showcasing its adeptness in making balanced predictions. This trend persisted during hold-out testing, indicating its reliability.

MiniLM, while maintaining a higher accuracy level and micro F1-score, further accentuated its prowess in overall cyberbullying detection. Noteworthy performance highlights include the third epoch of five-fold cross-validation, where MiniLM achieved a micro F1-score of 87.44%, reflective of its precision and recall balance. This trend persisted throughout the evaluation, validating MiniLM's effectiveness.

In summary, the comprehensive analysis depicted in Table 7 elucidates the capabilities of DistillBERT, DistilELECTRA, and MiniLM in overall cyberbullying detection. While all models demonstrated proficiency, MiniLM emerged as a standout performer, consistently achieving higher accuracy and micro F1-scores across various evaluation metrics. This underlines MiniLM's potential in enhancing the precision of overall cyberbullying detection strategies.

Table 7: Performance Evaluations for Cyberbullying Detection using PLMs (Overall)

		Overall							
		Cross-validation				Hold out			
PLM	**Epoch Number**	**A**	**MP**	**MR**	**MF**	**A**	**MP**	**MR**	**MF**
DistillB ERT	1	97.16	87.19	79.39	82.45	97.69	88.49	85.28	86.81
	2	97.31	87.26	81.47	83.84	97.84	92.01	82.62	86.66
	3	97.25	87.06	81.45	83.91	97.99	89.70	87.73	88.69
	4	97.34	87.42	81.39	84.09	98.03	91.09	86.35	88.55
DistilEL ECTRA	1	97.37	87.76	81.21	84.07	92.24	89.62	90.88	92.24
	2	97.50	89.63	80.73	84.14	94.10	85.20	89.10	94.10
	3	97.56	88.75	82.66	85.38	92.23	88.33	90.18	92.23
	4	97.58	88.69	83.00	85.52	92.97	87.54	90.05	92.97
MiniLM	1	97.77	90.29	83.56	86.53	98.42	94.23	87.70	90.68
	2	97.80	89.74	84.72	86.88	98.58	95.01	88.84	91.67
	3	97.84	89.71	85.45	87.44	98.57	93.57	90.24	91.83
	4	97.80	89.07	85.77	87.31	98.59	94.53	89.46	91.83

4.4 Benchmarking with Previous Research

This section is emphasizing the benchmarking results obtained in this study with previous research. Table 8 provides an insightful comparison of the performance related to the cyberbullying class from various previous studies that have utilized the AMiCA dataset for the purpose of cyberbullying detection. This comparison is conducted across both cross-validation and hold-out testing scenarios, shedding light on the effectiveness of different approaches and algorithms.

[33] employed a Support Vector Machine (SVM) model and reported a cross-validation accuracy of 96.67% and a corresponding precision, recall, and F-score of 73.32%, 57.19%, and 64.26%, respectively. Similarly, [50] used a Cascading Ensemble approach, yet specific values for precision, recall, and F-score are not available. [16] adopted Decision Trees (DT) and SVM methods, showcasing varying performance levels, including precision, recall, and F-score values. Amidst this context, [3] introduced advanced pre-trained models DistilBERT. DistilBERT achieved a notable F-score of 67.90% during cross-validation and an even more impressive 72.42% during hold-out testing.

Presenting a significant leap in performance, the current research employed the pre-trained models (DistilBERT, DistilELECTRA, and MiniLM) across the cross-validation and hold-out testing stages. MiniLM demonstrated an F-score of 76.01 in cross validation and an F-score of 84.41% in hold-out-test results, significantly surpassing prior approaches. For the cross-validation and hold-out test results, respectively, the fine-tuned MiniLM in this research surpassed the prior best system employing DistilBERT by [3] by an increment of 8.11% and 11.99% of the F-measure metric. These results highlight the considerable advancement offered by transfer learning techniques, underscoring their capability to enhance cyberbullying detection on the AMiCA dataset.

Table 8: Benchmarking with Previous Research on Cyberbullying Detection based on AMiCA dataset

		Cross-validation				Hold Out			
Previous Studies	M/PLM	A	P	R	F	A	P	R	F
[33]	SVM	96.67	73.32	57.19	64.26	97.21	74.13	55.82	63.69
[16]	DT	NA	NA	NA	NA	94.42	42.33	30.24	35.28
[16]	SVM	NA	NA	NA	NA	96.57	75.23	44.86	56.21
[50]	Cascading Ensemble	NA	NA	NA	NA	NA	52.49	70.06	60.02
[3]	DistilBERT	97.22	75.60	61.71	67.90	97.41	73.89	71.00	72.42
This Research	MiniLM	97.84	80.83	71.75	76.01	98.57	88.08	81.04	84.41

5. DISCUSSIONS

The presented study offers a thorough analysis of various pre-trained language models (PLMs) fine-tuned for cyberbullying detection using the AMiCA dataset. The evaluation encompasses different epochs, cross-validation, and hold-out testing scenarios, providing a detailed understanding of the models' performance across varied conditions. The obtained results not only contribute to the advancement of cyberbullying detection methodologies but also shed light on the capabilities of these PLMs for text classification tasks.

In the cyberbullying class, the DistillBERT model displayed promising capabilities, achieving notable F-scores of 69.57% and 78.14% during five-fold cross-validation and hold-out testing, respectively. DistilELECTRA and MiniLM also demonstrated strong potential, with F-scores of 72.30% and 81.00%, and 76.01% and 84.41% respectively, indicating their effectiveness in accurately identifying instances of cyberbullying. Particularly noteworthy is MiniLM's performance, as it consistently showcased competitive F-measure values across both evaluation scenarios, reinforcing its role in addressing class imbalance intricacies.

The evaluation of the non-cyberbullying class revealed similar patterns of proficiency among the PLMs. DistillBERT consistently achieved high accuracy, precision, recall, and F-scores, reaffirming its capacity to discern content unrelated to cyberbullying. DistilELECTRA and MiniLM similarly exhibited robust performance, consistently achieving commendable results. Notably, MiniLM's ability to maintain precision and recall balance throughout multiple epochs underscores its effectiveness in detecting content outside the cyberbullying realm.

Considering overall performance, MiniLM emerged as a standout performer, consistently outperforming its counterparts in accuracy and micro F1-scores across different PLMs and epochs. DistillELECTRA and DistillBERT also demonstrated strong overall performance, with competitive accuracy and F1-scores, contributing to the advancement of cyberbullying detection strategies.

Comparative analysis of hold-out testing's performance metrics using transfer learning approaches highlighted the versatility of the PLMs across binary classes and overall evaluation. DistilBERT, DistilELECTRA, and MiniLM displayed distinctive capabilities, each excelling in different aspects. MiniLM's consistent performance across both binary classes and overall evaluation reaffirms its potential as the best-performing model.

Furthermore, comparing the metrics from the cyberbullying class with results from earlier research demonstrates the substantial progress made in cyberbullying detection. MiniLM in this study outperformed previous benchmarks, showcasing the effectiveness of transfer learning techniques and their capacity to address the challenges of cyberbullying detection on the AMiCA dataset.

The findings underscore the remarkable proficiency of MiniLM in comparison to DistillBERT and DistilELECTRA. MiniLM's consistent and robust performance across various evaluation metrics, epochs, and methodologies positions it as a standout performer in the realm of cyberbullying detection. This can be attributed to its smaller scale, which allows it to capture subtle nuances within the data and potentially avoid overfitting, enabling a finely balanced precision and recall. The research unveils the potential of transfer learning techniques, emphasizing MiniLM's capability to enhance precision and bolster overall cyberbullying detection strategies in the digital landscape, further solidifying its position as an efficient tool for addressing contemporary online challenges.

6. CONCLUSION

The research sought to enhance the precision and reliability of cyberbullying detection by leveraging advanced transfer learning models. To achieve this, three pre-trained language models (PLMs) – DistillBERT, DistilELECTRA, and MiniLM – were carefully fine-tuned. Transfer learning, which allows pre-trained models to apply their linguistic understanding to new, domain-specific challenges, proved instrumental in addressing the nuanced nature of cyberbullying language. By building upon pre-existing features and linguistic representations, the models demonstrated improved capabilities in capturing subtle expressions of harmful behavior, positioning them as valuable tools for creating more accurate and reliable detection systems.

To rigorously assess the models' effectiveness, the research adopted an exhaustive evaluation framework, analyzing performance across various epochs, methodologies, and metrics. This comprehensive approach not only highlighted the strengths and limitations of each PLM but also ensured a robust comparison of their abilities to classify content into binary categories – cyberbullying and non-cyberbullying. The results revealed that transfer learning significantly enhanced the models' precision and recall rates, leading to a substantial reduction in false positives and negatives. This demonstrated the models' ability to address the inherent challenges in cyberbullying detection, which often involve identifying subtle and context-dependent language patterns.

Among the three models, MiniLM consistently emerged as the top performer, outperforming DistillBERT and DistilELECTRA across multiple evaluation metrics. MiniLM excelled in striking a balance between precision and recall, achieving competitive F-measure values despite its smaller size. This balance enabled it to effectively minimize classification errors, such as misidentifying non-cyberbullying content as harmful or overlooking subtle instances of cyberbullying. The model's success underscores the importance of lightweight yet powerful PLMs in achieving the dual objectives of high accuracy and operational efficiency, making it particularly well-suited for real-world applications where computational resources may be limited.

The research marked a significant advancement over previous studies in cyberbullying detection, highlighting the transformative potential of fine-tuned transfer learning models. By demonstrating improved accuracy, reliability, and efficiency, the study validated the effectiveness of PLMs in addressing a critical societal challenge. The reduction in false positives and negatives not only enhanced the system's trustworthiness but also underscored its role in creating safer online environments. This achievement sets a strong precedent for the continued use of advanced PLMs in tackling complex language-related issues, paving the way for further innovation in cyberbullying detection and other related domains.

The foundation established by optimizing models like DistilBERT, DistilELECTRA, and MiniLM for binary text classification in cyberbullying detection provides a solid basis for future research. While binary classification effectively identifies cyberbullying, there is significant potential to expand into role classification. By analysing the dynamics between various actors, such as aggressors, victims, and bystanders, researchers can better understand the complexities of online interactions. Role classification would enable a deeper examination of the social factors and interactions driving cyberbullying, leading to more comprehensive prevention and intervention strategies.

Beyond binary classification, future research can explore multiclass classification to identify and differentiate between specific forms of cyberbullying, such as taunts, threats, hate speech, and more. Each form of harassment exhibits unique linguistic and contextual patterns and understanding these differences can enable the creation of more precise and targeted interventions. A multiclass framework would not only provide nuanced insights into cyberbullying behaviours but also offer tailored solutions to address different types of harassment, thus improving the effectiveness of detection and mitigation strategies.

Given the global and multilingual nature of the internet, cross-lingual cyberbullying detection is an essential area for future exploration. Training models on multilingual datasets would allow for the detection of cyberbullying across diverse linguistic and cultural contexts. This approach involves overcoming challenges such as language-specific nuances, translation complexities, and cultural variations in communication styles. Successfully addressing these issues can lead to inclusive solutions that effectively combat cyberbullying in different regions and languages, creating a safer digital environment for users worldwide.

Future research should also prioritize incorporating additional datasets from various social media platforms. Different platforms have distinct communication norms and user interactions, and expanding the datasets used for model training will improve the robustness and generalizability of detection systems. Including demographic and social context information, such as age, cultural background, and relationship dynamics, can further enhance model accuracy. Additionally, integrating advanced techniques like sentiment analysis and contextual understanding will provide deeper insights into the underlying emotional and situational aspects of cyberbullying incidents, distinguishing between harmful and benign interactions.

Lastly, leveraging state-of-the-art models, such as GPT and other advanced architectures, offers significant opportunities to refine cyberbullying detection frameworks. These models excel in handling complex linguistic patterns and contextual subtleties, making them well-suited to tackle the evolving challenges of online harassment. Addressing ethical challenges, such as false positives and privacy concerns, will also be critical to ensure responsible deployment. By focusing on these areas, future research can contribute to building more accurate, ethical, and globally effective cyberbullying detection systems, paving the way for a safer and more inclusive online environment.

Appendix A.
The source code for this research is available on GitHub with the link
https://github.com/JasmeenBongKahYing/Cyberbullying-Detection/.

REFERENCES

[1] World Bank. (2022). *Individuals Using the Internet (% of population) | Data*. World Bank. https://data.worldbank.org/indicator/IT.NET.USER.ZS

[2] Malaysian Communications and Multimedia Commission. (2022). *Internet Users Survey 2022 SURUHANJAYA KOMUNIKASI DAN MULTIMEDIA MALAYSIA MALAYSIAN COMMUNICATIONS AND MULTIMEDIA COMMISSION*. https://www.mcmc.gov.my/skmmgovmy/media/General/IUS-2022.pdf

[3] Teng, T. H., & Varathan, K. D. (2023). Cyberbullying Detection in Social Networks: A Comparison between Machine Learning and Transfer Learning Approaches. *IEEE Access*, 1–1. https://doi.org/10.1109/access.2023.3275130

[4] Kee, D. M. H., Al-Anesi, M. A. L., & Al-Anesi, S. A. L. (2022). Cyberbullying on social media under the influence of COVID-19. *Global Business and Organizational Excellence*, *41*(6). https://doi.org/10.1002/joe.22175

[5] Kwan, I., Dickson, K., Richardson, M., MacDowall, W., Burchett, H., Stansfield, C., Brunton, G., Sutcliffe, K., & Thomas, J. (2020). Cyberbullying and Children and Young People's Mental Health: A Systematic Map of Systematic Reviews. *Cyberpsychology, Behavior, and Social Networking*, *23*(2), 72–82. https://doi.org/10.1089/cyber.2019.0370

[6] Cook, S. (2024, January 10). *Cyberbullying Statistics and Facts for 2016 - 2018 | Comparitech*. Comparitech; Comparitech. https://www.comparitech.com/internet-providers/cyberbullying-statistics/

[7] CDC. (2023, September 28). *Preventing bullying*. Centers for Disease Control and Prevention. https://www.cdc.gov/violenceprevention/youthviolence/bullyingresearch/fastfact.html

[8] Ogunleye, B., & Babitha Dharmaraj. (2023). *Use of Large Language Model for Cyberbullying Detection*. https://doi.org/10.20944/preprints202306.1075.v1

[9] Alduailaj, A. M., & Belghith, A. (2023). Detecting Arabic Cyberbullying Tweets Using Machine Learning. *Machine Learning and Knowledge Extraction*, *5*(1), 29–42. https://doi.org/10.3390/make5010003

[10] Jeevitha, R., Chaitanya, K., Mathesh, N., Nithyanarayanan, B., & Darshan, P. (2023). Using Machine Learning to Identify Instances of Cyberbullying on Social Media. *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, *8*, 207–212. https://doi.org/10.1109/icscds56580.2023.10104720

[11] Sultan, T., Jahan, N., Basak, R., Jony, M., & Nabil, R. (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). 1-13.* https://doi.org/10.5815/ijisa.2023.02.01

[12] Agrawal, T., & Chakravarthy, V. D. (2022). Cyberbullying Detection and Hate Speech Identification using Machine Learning Techniques. *2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS).* https://doi.org/10.1109/icps55917.2022.00041

[13] Khan, S., & Qureshi, A. (2022). *Cyberbullying Detection in Urdu Language Using Machine Learning*. https://doi.org/10.1109/etecte55893.2022.10007379

[14] Siddhartha, K., Kumar, K. R., Varma, K. J., Amogh, M., & Samson, M. (2022, August 1). *Cyber Bullying Detection Using Machine Learning*. IEEE Xplore. https://doi.org/10.1109/ASIANCON55314.2022.9909201

[15a] Singh, Ksh. N., Devi, S. D., Devi, H. M., & Mahanta, A. K. (2022). A novel approach for dimension reduction using word embedding: An enhanced text classification approach. *International Journal of Information Management Data Insights*, *2*(1), 100061. https://doi.org/10.1016/j.jjimei.2022.100061

[15b] Singh, N. K., Singh, P., & Chand, S. (2022, November 1). *Deep Learning based Methods for Cyberbullying Detection on Social Media*. IEEE Xplore. https://doi.org/10.1109/ICCCIS56430.2022.10037729

[16] Ali, W.N.H.W., Mohd, M., & Fauzi, F. (2021). Cyberbullying Predictive Model, Implementation of Machine Learning Approach. 2021 *Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*. https://doi.org/10.1109/camp51653.2021.9497932

[17] Shrimali. S. (2022). *A Natural Language Processing and Machine Learning-Based Framework to Automatically Identify Cyberbullying and Hate Speech in Real-Time*. https://doi.org/10.1109/urtc56832.2022.10002243

[18] Bharti, S., Yadav, A. K., Kumar, M., & Yadav, D. (2021). Cyberbullying detection from tweets using deep learning. *Kybernetes*. https://doi.org/10.1108/k-01-2021-0061

[19] Bozyiğit, A., Utku, S., & Nasibov, E. (2021). Cyberbullying Detection: Utilizing Social Media Features. *Expert Systems with Applications*, *179*, 115001. https://doi.org/10.1016/j.eswa.2021.115001

[20] Gada, M., Damania, K., & Sankhe, S. (2021). Cyberbullying Detection using LSTM-CNN architecture and its applications. *2021 International Conference on Computer Communication and Informatics (ICCCI)*. https://doi.org/10.1109/iccci50826.2021.9402412

[21] Jain, V., Kumar, V., Pal, V., & Vishwakarma, D. K. (2021). Detection of Cyberbullying on Social Media Using Machine learning. *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. https://doi.org/10.1109/iccmc51019.2021.9418254

[22] López-Vizcaíno, M. F., Nóvoa, F. J., Carneiro, V., & Cacheda, F. (2021). Early detection of cyberbullying on social media networks. *Future Generation Computer Systems*, *118*. https://doi.org/10.1016/j.future.2021.01.006

[23] Nureni Ayofe AZEEZ, Misra, S., Omotola Ifeoluwa LAWAL, & Oluranti, J. (2021). Identification and Detection of Cyberbullying on Facebook Using Machine Learning Algorithms. *Journal of Cases on Information Technology*, *23*(4), 1–21. https://doi.org/10.4018/jcit.296254

[24] Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security*, *90*, 101710. https://doi.org/10.1016/j.cose.2019.101710

[25] Ho, S. M., Kao, D., Chiu-Huang, M.-J., Li, W., & Lai, C.-J. (2020). Detecting Cyberbullying "Hotspots" on Twitter: A Predictive Analytics Approach. *Forensic Science International: Digital Investigation*, *32*, 300906. https://doi.org/10.1016/j.fsidi.2020.300906

[26] Ishara Amali, H. M. A., & Jayalal, S. (2020, July 1). *Classification of Cyberbullying Sinhala Language Comments on Social Media*. IEEE Xplore. https://doi.org/10.1109/MERCon50084.2020.9185209

[27] Shah, R., Aparajit, S., Chopdekar, R., & Patil, R. (2020). Machine Learning based Approach for Detection of Cyberbullying Tweets. *International Journal of Computer Applications*, *175*(37), 51–56. https://doi.org/10.5120/ijca2020920946

[28] Balakrishnan, V., Khan, S., Fernandez, T., & Arabnia, H. R. (2019). Cyberbullying detection on twitter using Big Five and Dark Triad features. *Personality and Individual Differences*, *141*, 252–257. https://doi.org/10.1016/j.paid.2019.01.024

[29] Hani, J., Nashaat, M., Ahmed, M., Emad, Z., Amer, E., & Mohammed, A. (2019). Social Media Cyberbullying Detection using Machine Learning. *International Journal of Advanced Computer Science and Applications*, *10*(5). https://doi.org/10.14569/ijacsa.2019.0100587

[30] Win, Y. (2019). Classification using Support Vector Machine to Detect Cyberbullying in Social Media for Myanmar Language. *2022 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, 122–125. https://doi.org/10.1109/icce-asia46551.2019.8942212

[31] Zhang, J., Otomo, T., Li, L., & Nakajima, S. (2019). *Cyberbullying Detection on Twitter using Multiple Textual Features*. https://doi.org/10.1109/icawst.2019.8923186

[32] Al-Ajlan, M. A., & Ykhlef, M. (2018). Deep Learning Algorithm for Cyberbullying Detection. *International Journal of Advanced Computer Science and Applications*, *9*(9). https://doi.org/10.14569/ijacsa.2018.090927

[33] Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., Guy De Pauw, Daelemans, W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PLOS ONE*, *13*(10), e0203794. https://doi.org/10.1371/journal.pone.0203794 '

[34] Haidar, B., Chamoun, M., & Serhrouchni, A. (2017). A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning. *Advances in Science, Technology and Engineering Systems Journal*, *2*(6), 275–284. https://doi.org/10.25046/aj020634

[35] Zhao, R., & Mao, K. (2017). Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder. *IEEE Transactions on Affective Computing*, *8*(3), 328–339. https://doi.org/10.1109/taffc.2016.2531682

[36] Zhang, X., Tong, J., Vishwamitra, N., Whittaker, E., Mazer, J. P., Kowalski, R., Hu, H., Luo, F., Macbeth, J., & Dillon, E. (2016, December 1). *Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network*. IEEE Xplore. https://doi.org/10.1109/ICMLA.2016.0132

[37] Meenakshi, M., Babu, P. S., & Hemamalini, V. (2023, April). Deep learning techniques for spamming and cyberbullying detection. In *2023 International Conference on Networking and Communications (ICNWC)* (pp. 1-10). IEEE.

[38] Bhatia, B., Verma, A., Anjum, & Rahul Katarya. (2022). Analysing Cyberbullying Using Natural Language Processing by Understanding Jargon in Social Media. *Lecture Notes in Electrical Engineering*, 397–406. https://doi.org/10.1007/978-981-16-9012-9_32

[39] Maity K., Saha, S., & Bhattacharyya, P. (2022). Cyberbullying Detection in Code-Mixed Languages: Dataset and Techniques. *2022 26th International Conference on Pattern Recognition (ICPR)*. https://doi.org/10.1109/icpr56361.2022.9956390

[40] Verma, K., Milosevic, T., Cortis, K., & Davis, B. (n.d.). *Benchmarking Language Models for Cyberbullying Identification and Classification from Social-media texts*. https://aclanthology.org/2022.lateraisse-1.4.pdf

[41] Yi, P., & Zubiaga, A. (2022). Cyberbullying Detection across Social Media Platforms via Platform-Aware Adversarial Encoding. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*, 1430 1434. https://doi.org/10.1609/icwsm.v16i1.19401

[42] Behzadi, M., Harris, I. G., & Derakhshan, A. (2021). Rapid Cyber-bullying detection method using Compact BERT Models. *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*. https://doi.org/10.1109/icsc50631.2021.00042

[43] Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). https://doi.org/10.1609/icwsm.v12i1.14991

[44] Elsafoury, F., Katsigiannis, S., Pervez, Z., & Ramzan, N. (2021). When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection. *IEEE Access*, *9*, 103541–103563. https://doi.org/10.1109/access.2021.3098979

[45] Pericherla, S., & E, I. (2021). Cyberbullying detection on multi-modal data using pre-trained deep learning architectures. *Ingeniería Solidaria*, *17*(3), 1–20. https://doi.org/10.16925/2357-6014.2021.03.09

[46] Yadav, J., Kumar, D., & Chauhan, D. (2020). Cyberbullying Detection using Pre-Trained BERT Model. *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. https://doi.org/10.1109/icesc48915.2020.9155700

[47] Kamiri, J., & Mariga, G. (2021). Research Methods in Machine Learning: A Content Analysis, *Internal Journal of Computer and Information Technology (2279-0764, 10(2)*. https://doi.org/10/24203/ijcit.v10i2.79

[48] Kao, H.-T., Yan, S., Huang, D., Bartley, N., Hosseinmardi, H., & Ferrara, E. (2019). Understanding cyberbullying on Instagram and ask.fm via Social Role Detection. *Companion Proceedings of The 2019 World Wide Web Conference*, 183–188. https://doi.org/10.1145/3308560.3316505

[49] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., & Drame, M. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[50] Jacobs, G., Van Hee, C., & Hoste, V. (2022). Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?. *Natural Language Engineering*, *28*(2), 141-166.