IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 March 2023, accepted 27 April 2023, date of publication 10 May 2023, date of current version 8 June 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3275130*

RESEARCH ARTICLE

# Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches

**TEOH HWAI TENG AND KASTURI DEWI VARATHAN**
Department of Information Systems, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia
Corresponding author: Kasturi Dewi Varathan (kasturi@um.edu.my)

This work was supported by the Impact Oriented Interdisciplinary Research Grant Universiti Malaya under Project IIRG001A-19SAH.

**ABSTRACT** Information and Communication Technologies fueled social networking and facilitated communication. However, cyberbullying on the platform had detrimental ramifications. The user-dependent mechanisms like reporting, blocking, and removing bullying posts online is manual and ineffective. Bag-of-words text representation without metadata limited cyberbullying post text classification. This research developed an automatic system for cyberbullying detection with two approaches: Conventional Machine Learning and Transfer Learning. This research adopted AMiCA data encompassing significant amount of cyberbullying context and structured annotation process. Textual, sentiment and emotional, static and contextual word embeddings, psycholinguistics, term lists, and toxicity features were used in the conventional Machine Learning approach. This study was the first to use toxicity features to detect cyberbullying. This study is also the first to use the latest psycholinguistics features from the Linguistic Inquiry and Word (LIWC) 2022 tool, as well as Empath's lexicon, to detect cyberbullying. The contextual embeddings of ggeluBert, tnBert, and DistilBert have alike performance, however DistilBert embeddings were elected for higher F-measure. Textual features, DistilBert embeddings, and toxicity features that struck new benchmark were the top three unique features when fed individually. The model's performance was boosted to F-measure of 64.8% after feeding with a combination of textual, sentiment, DistilBert embeddings, psycholinguistics, and toxicity features to the Logistic Regression model that outperforms Linear SVC with faster training time and efficient handling of high-dimensionality features. Transfer Learning approach was by fine-tuning optimized version Pre-trained Language Models namely, DistilBert, DistilRoBerta, and Electra-small which were found to have speedier training computation than their base form. The fine-tuned DistilBert resulted with the highest F-measure of 72.42%, surpassing CML. Our research concluded that Transfer Learning was the best for uplifted performance and lesser effort as feature engineering and resampling was omitted.

**INDEX TERMS** Cyberbullying detection, DistilBert, machine learning, pre-trained language models (PLMs), transfer learning, toxicity features, AMiCa dataset, LIWC, empath.

## I. INTRODUCTION

Information and Communication Technologies (ICT) have become an integral part of everyone's life, evolving imperceptibly with time, catalyzing online communication between people. Communication has been just one button click with the widespread use of the online platform, facilitating the growth of social networking. ICT dominance has a dark side when people easily misuse technological advancement with abusive behaviors such as cyberbullying. Cyberbullying is the expanded form of direct or traditional bullying through electronic platforms [1], [2], [3], [4], [5], [6]. Social media becomes the virtual medium for bullying, shielding the bully's identity, making detecting cyberbullying a

The associate editor coordinating the review of this manuscript and approving it for publication was Valentina E. Balas.

complex and challenging mission to protect online communities. Cyberbullying cases increase with volumized Internet usage because it can be easily committed anonymously [7], leading to a grave public health concern that brings many negative impacts, such as mental, psychological, and social problems [8]. While cyberbullying victims tend to suffer from mental health problems such as depression, anxiety, loneliness, and anhedonia, some are reported to be committing self-injurious behavior and suicidal ideation [9].

Initially, the community implemented a manual approach to monitoring cyberbullying activities. Parent-Teacher Association started a good initiative from the Japanese school that formed Internet Patrol to help filter websites manually with inappropriate content, but it is impossible to handle the vast volume of data on the Internet within a short time without a computational approach [10], [11], [12]. Automating cyberbullying detection is paramount to facilitate the process, ensuring a safe environment within online social media. As the computational text analysis can effectively be adopted to examine the social and cultural phenomena [13], the primary focus of this research is to automate the detection of cyberbullying instances from the unruly post, deeming the problem as a text classification task with the help of state-of-the-art techniques using artificial intelligence and natural language processing knowledge. By natural language processing, text classification is frequently employed in identifying the category of a given corpus through several stages, such as text preprocessing, feature extraction, and the development of a classification model [14].

Social media companies have developed policies and mechanisms to maintain the regulation of social media platforms. However, the social media company was not performing well in tackling cyberbullying [15], [16]. The available mechanisms are usually user-dependent, requiring users to report content, block, or unfriend, a passive way of mitigating cyberbullying [17]. Although the implementation of algorithms with supervised machine learning works to detect cyberbullying events and helps to expunge posts that may contain foul words; however, the outcome is not as accurate as those reported by users [17]. Furthermore, metadata associated with the online platform and user information are not always available due to privacy protection [18], [19]. In that case, textual content posted by the online platform users is the base input for cyberbullying detection model [20]. The initial studies on automatic cyberbullying detection deemed the presence of "bad" words (insult and swear words) or profane terms to be one factor in making a post likely to be an act of cyberbullying. However, looking for a list of words to detect such events is not very effective because the words or sentences can be easily deformed or obfuscated in terms of spelling, and a consistent list update is required [21]. Using textual features such as the presence of "bad" words (insult, swear, profane word) in making a post to be an act of cyberbullying has its limitation since the explicit existence of these words is not always right to detect cyberbullying [22]. Extraction of additional features by expanding the usual bag-of-words text representation is needed to improve the performance of cyberbullying detection model [18].

Most studies disregarded the posts from victims and bystanders, confining the cyberbullying contexts to bullying posts [23]. To develop cyberbullying detection models, this research adopted an open dataset focused only on English posts, with widespread cyberbullying context from different roles, such as harasser, victim, and bystander, so that the model tackles all the possible contexts during cyberbullying interaction. Aside from that, this research predominantly worked on textual online platform posts, having text as the only input for deriving features during the feature engineering process. Other metadata associated with the post were unavailable and not within this research's scope. This research aims to use state-of-the-art natural language processing (NLP), conventional machine learning (CML), and transfer learning (TL) approaches to attain the task. The main objective is to develop models that detect unruly posts with cyberbullying traits on social media to protect users from participating in cyberbullying acts or becoming cyberbullying victims.

There were a few contributions made in this research. Using the CML approach, this research explores various features utilized from previous studies that aid the work in the feature engineering process, shedding light on the derivation of features from the text to be fed into the models, then identifying the best combination of features and models. Aside from some features studied in the literature review, a new feature, toxicity, was introduced in this research. Gada et al. [24] used datasets with multi-label toxicity to study the cyberbullying phenomenon but deemed the task as multi-label classification instead of binary cyberbullying classification. Vo et al. [25] fine-tuned the toxic-Bert transformer for cyberbullying detection. However, none of the studies utilized the architecture to craft toxicity features as part of the feature engineering process to feed into the conventional machine learning model. This research differs in the sense that the toxicity of text was extracted using the 'detoxify' framework. This research is the first to employ toxicity features as the input for the conventional machine learning model in cyberbullying detection. Other than static word embeddings, more variants of contextual word embeddings that have not been experimented with before in other studies, such as tnBert, mobileBert and ggeluBert, were explored. While previous studies adopted LIWC15 to extract psycholinguistics features, this research used the newly launched LIWC 2022 and Empath packages which were not used in previous cyberbullying detection studies. This research will show how to improve the preprocessing of text, the generation of features based on text input, and methods used to perform the detection model for cyberbullying events. Additionally, the popularity of Pre-trained Language Models (PLMs) motivates the introduction of the TL approach in the NLP community. Transfer learning approach was known to be robust to work with small and imbalanced text classification [26] and it was found to surpass conventional machine learning models by fine-tuning PLMs [27]. Also, the lighter transformer of PLMs was found to be more effective with a shorter fine-tuning process and lesser computational resources without much deviation in the performance compared to the larger model in text classification [28]. Thus, this research adopts the lighter, smaller and optimized versions of PLMs, such as DistilBert, DistilRoBerta, and Electra-small.

Cyberbullying is a community concern; thus, timely detection is crucial for prevention and reduction. The main contribution of this work is significant as natural language processing and supervised learning are adopted in automating the detection of cyberbullying events. This research explores the features and examines the methods for modeling the automatic detection of cyberbullying traces from textual content. These findings significantly improve the task performance for text classification of cyberbullying events. Moreover, the results encourage better solutions that social media platforms can adopt to combat cyberbullying. The expected outcome of this research is the development of classification models that can effectively detect cyberbullying and non-cyberbullying events from unruly posts by applying the knowledge of state of the art in data science. This work incorporates text preprocessing, feature engineering, model development using different classifiers by the CML approach, and fine-tuning the optimized version of PLMs from the Hugging Face community platform by the TL approach. The methodological steps applied to reach the research objective are expected to be significant.

The research is divided into five sections. Section I sets the stage for the remainder of this research by providing an overview of the study, the importance of automated cyberbullying detection, the motives, and the significance of using data science knowledge in this research study. Section II provides an overview of recent works on cyberbullying detection. Section III elaborates on the methodology and process flow with detailed information on the tools used and system requirements while conducting the research. Section IV presents the experimental results of conventional machine learning and transfer learning approaches. This chapter shows the performance comparison between approaches, benchmarking with previous studies, and further discusses the findings. Lastly, Section V concludes the findings by providing a summary of the work adopted, addresses the limitations encountered while conducting the research, and discusses the future direction.

## II. RELATED WORKS

This section is divided into seven subsections. The first three subsections present the background study of previous work on cyberbullying detection with conventional machine learning, deep learning, and transfer learning approaches. The multi-modal cyberbullying detection and studies that work on other cyberbullying-related classifications are discussed in the fourth and fifth subsections. Other cyberbullying-related studies are also briefly discussed in the following subsection. The last subsection comprehensively reviews the feature used in cyberbullying detection across the studies.

### A. CYBERBULLYING DETECTION WITH CONVENTIONAL MACHINE LEARNING APPROACH

Based on the literature, SVM was found to be the best for many of the cyberbullying detection studies [29], [30], [31], [32], [33], [34], [35], [36]. Some researchers chose SVM for cyberbullying detection model as it was proven to work for data with highly skewed distribution [37], [38]. Tracing back the research trend for cyberbullying detection, Yin et al. [36] were pioneers in working for automated cyberbullying detection, where the authors classified the online harassment posts from three social websites: Kongregate, Slashdot, and MySpace, by adopting supervised learning with linear kernel Support Vector Machine (SVM). The authors suggested using sentiment, contextual, and similarity features in addition to TFIDF, which effectively detects harassment posts, especially chat-style posts. Reynolds et al. adopted several models such as SVM with Sequential Minimal Optimization (SMO), tree-based JRIP, tree-based J48, and Instance-Based Learner (IBK) to detect the cyberbullying post, collected from Formspring.me, a forum for question and answer with anonymized users. Dinakar et al. [35] experimented with several conventional machine learning models such as J48, SVM, and JRIP on the YouTube comments data, labeled by three cyberbullying topics, namely sexuality, race, and intelligence. The authors developed binary classification models that took individual labels as target class and multiclass classification models that trained all labels. The authors found that the binary classification models that took individual labels as target class yielded better results in classifying each cyberbullying topic's labels than the multiclass classification models that train all labels as a whole. Muneer and Fati [39] experimented with different models for cyberbullying detection and compared the models from training time consumed and prediction performance. The authors found that multinomial naïve bayes model resulted with shortest training time, however random forest consume the most training time. In terms of prediction ability, Logistic Regression outperformed the other models and eventually yield the shortest prediction time when the data size became greater. Bozyiğit, et al. [20] found that AdaBoost was the best model with the highest accuracy in classifying both cyberbullying and non-cyberbullying posts, mainly because the model was trained by integrating several SVM models, but slowest in prediction compared with other models. The authors explained that the Random Forest model would be preferred for detecting cyberbullying posts as it has the highest recall metric. Thun et al. [40] identified the best features of the Random Forest model in detecting cyberbullying posts by generating the feature importance based on the Gini index. Random Forest was preferred by some studies as it reduced overfitting issue by averaging the trees with random features subsets in classifying the instances [41], [42].

### B. CYBERBULLYING DETECTION WITH DEEP LEARNING APPROACH

Recently, deep learning models for cyberbullying detection have become prominent. Zhao and Mao [43] extended the deep learning model, Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA), for cyberbullying detection with the capability to uncover the hidden feature embedded within the cyberbullying post and the ability to learn from a robust and discriminative text representation. The novel pronunciation-based convolutional neural network (PCNN) was introduced by Zhang et al. [44] that fed the phoneme codes of text on a CNN-based model, and it performed better than other neural network models as it was designed to handle the spelling error without changing the pronunciation. Kumar and Sachdeva [45] developed a hybrid deep learning framework, Bi-GRU-Attention-CapsNet (Bi-GAC), by coupling Bi-GRU self-attention encoder and capsule network in capturing the semantic representations and spatial information of the textual content in social media for cyberbullying detection model. Shriniket et al. [46] proposed an algorithm, CNN-Semi Trained GloVe model, by incorporating the semantic word embedding with CNN, and this model consumes lesser training time with high prediction accuracy for cyberbullying detection. Deep learning models have gained massive attention in cyberbullying detection studies as the models were found to have more optimistic results than conventional machine learning models. Agrawal and Awekar [47]'s study focused on cyberbullying detection on multiple social platforms rather than a standalone data source. The authors have conducted extensive experiments to study the performance of several conventional machine learning models such as Logistic Regression, Support Vector Machine, Random Forest, and Naive Bayes, Deep Neural Network models such as CNN, LSTM, BLSTM, and Attention-based BLSTM. These models were coupled with several word representations techniques for word embedding generation, such as BoW, GloVe, and SSWE. Regardless of having a significant amount of data from multiple sources, the deep learning models still beat the performance of the conventional machine learning models. Dadvar and Eckert [48] experimented with deep neural network-based models such as CNN, Long Short Term Memory (LSTM), Bidirectional LSTM (BLSTM), and attention-based BLSTM that transcend the conventional machine learning models applied by other studies using the same YouTube dataset. Rosa, et al. [49] pointed out that CNN, a hybrid CNN-LSTM, and a mixed CNN-LSTM-DNN were robust to class imbalance problems in cyberbullying datasets and performed better than SVM and Logistic Regression. Cheng, et al. [50] constructed a sophisticated Hierarchical Attention Network for Cyberbullying Detection (HANCD) framework, outperforming LSTM and CNN Hani et al. [51] compared the Neural Network (NN) model with the SVM model, and without any surprise, NN has better accuracy than SVM. The novel Deep Decision Tree, built by Yuvaraj et al. [52], processed the text input on the tree nodes made up of the DNN hidden layers and yielded higher accuracy in classifying cyberbullying posts.

### C. CYBERBULLYING DETECTION WITH TRANSFER LEARNING APPROACH

Adopting a transfer learning approach using language models is not as extensive as the conventional machine learning approach in cyberbullying detection studies. Some studies have recently implemented the transformer model, especially Bert, with promising results. Paul and Saha [53] developed a Bert-based cyberbullying detection model that yields optimistic accuracy. Besides Bert, Jacobs et al. fine-tuned RoBerta for the cyberbullying participant's role classification model, which resulted in the best performance. Elsafoury et al. [26] developed cyberbullying detection model by fine-tuning the language model, Bert, which outperformed the other deep learning models. While most existing studies focus on just one platform in experimenting cyberbullying detection model, Yi and Zubiaga [54] introduced a feasible framework for cross-platform cyberbullying detection, XP-CB, built on Bert and RoBerta with the promising result when experimented on different cross-platform configurations Verma et al. [55] fine-tuned the pre-trained transformer from HuggingFace's Transformer library, the Hate-BERT model that transcended the base Bert model, Bi-LSTM, and the traditional machine learning model, SVM. Even though not many preprocessing steps are required by adopting the pre-trained transformer model compared to the conventional machine learning model, Bhatia et al. [56] found that the fine-tuned Bert model was significantly boosted when employing preprocessed data with slang-abusive corpus.

### D. MULTI-MODAL CYBERBULLYING DETECTION

The engagement of Internet users nowadays using the social platform is not limited to text but also other multimedia content such as images, video, and audio, giving rise to cyberbullying detection based on multi-modalities. Among the accessible publications, few researchers tackle the cyberbullying classification based on multi-modalities [57], [58], [59], [60], [61]. Hosseinmardi, et al. [62] developed a cyberbullying detection model based on the video-based online social platform Vine. The work was the first that adopted textual, sentiment, user, social media, and video features represented as the labeled emotions and contents in literature. Soni and Singh [63] found that multimodal cyberbullying detection that accommodates textual, audio, and visual features yield better accuracy. Wang et al. [61] introduced new multi-modal cyberbullying detection (MMCD) framework that coupling BiLSTM developed with attention, hierarchical attention network (HAN) that handle word and comment and visual embedding to deal with various information from the multi-modal data such as comment texts image, video, and time. A pure image-based cyberbullying detection model was developed by Roy and Mali [64], adopting CNN-based VGG16 and InceptionV3 transfer learning approaches. For cyberbullying involving non-textual context, Vishwamitra et al. [65] identified a few determinate factors for cyberbullying contents that can be extracted from the image: the posture of the body, facial expression (emotion), gesture, presence of objects, and social factors (i.e., anti-LGBT symbols).

### E. CYBERBULLYING CLASSIFICATION (ROLE, SEVERITY LEVEL, TYPES)

The existing studies were limited to the binary text classification of cyberbullying. Based on the literature, a few researchers have worked on the cyberbullying participants' role classification based on textual patterns. Sui was the first study to work on cyberbullying participants' role identification model to classify Twitter posts into different roles: reinforcer, defender, outsider, assistant, reporter, and accuser Chatzakou et al. [66] and Chatzakou et al. [67] developed cyberbullying detection models based on the Twitter datasets labeled as aggressors, bullies, spammers, or normal with text-based features and Twitter metadata to study the characteristics of bullies and aggressors. However, the studies above did not recognize the role of victims and bystanders. Considering the importance of studying victim and bystander context, Jacobs et al. [23] developed cyberbullying role classification model that identifies whether the post exhibits context from the harasser, victims, or bystanders within cyberbullying episodes using the AMiCA dataset. With the same dataset, Rathnayake et al. [68] developed two Bert-based classification models: the bullying model and the defending model. The former model classifies posts from the harasser and bystander assistant, and the latter classifies posts from the victim and bystander defender. Other than role classification, Van Hee et al. [38] tackled the problem as multiclass text classification and developed a model that classifies the more delicate cyberbullying types such as threat or blackmail, sexual talk, insult, curse or exclusion, defense, defamation, and encouragement using the dataset from Ask.fm platform. Another kind of multiclass text classification was cyberbullying severity level classification. Sugandhi et al. [69] developed a cyberbullying severity classification model that determines the post's severity level (high, medium, and low).

### F. FEATURES USED IN AUTOMATED CYBERBULLYING DETECTION

This subsection discovers and describes the features used across previous studies: user features, social media features, textual features, sentiment and emotional features, word embeddings, psycholinguistics features, personality traits features, customized dictionary lists, and topic modeling. Except for user and social media features, the rest can be derived or crafted from the text input with the help of available packages, tools, or models explicitly designed to attain different tasks.

User features reveal any personal information of the post's owner, such as gender, age, race, marital status, etc. Dadvar et al. (2012) [76] proposed a gender-based cyberbullying detection model that incorporates the user's gender information as an essential input for the model development. Based on their historical comments, the authors extended the studies to feed the machine learning models with other user's personal information, such as age, activity history, and characteristics [70].

Social media features are metadata associated with the text post (i.e., the number of likes, comments, and shares) and the user's online social platform account details (i.e., the number of followers, friends, the total number of posts, and active hours). Rafiq et al. [71] include social media features such as the number of followers, followings, likes, comments, and views of the media post from the Vine platform to develop cyberbullying detection model. Al-Garadi et al. [72] crawled corpus from Twitter and utilized the metadata associated with each tweet, such as the social media feature (i.e., number of followers, followings, tweets, mentions, etc.) and user feature (i.e., gender, age) for cyberbullying detection. However, the literature found that not all studies fed user features and social media features in cyberbullying models. This information was not openly shared by some social media sites due to users' privacy, causing incomplete extraction of information unless shared by previous researchers [73].

Textual features are those that adopt statistical metrics to quantify the text pattern. In literature, the computation of text statistics is the most common approach for feature extraction on textual input, for example, computation of TFIDF, weighted TFIDF, Bag of Words (BoW), n-gram, skip-gram, count or proportion of profane words, first, second, and other personal pronouns, nouns, adjectives, upper case letters, emoji, emoticons, and punctuation [52], [69], [70], [74], [75], [76], [77]. In previous studies, the Bag of Words (BOW) was the most popular text representation feature [26]. Among which, word-level BOW was widely employed in previous studies [23], [37], [60], [78], [79], [80], [81]. Agrawal and Awekar [47] compared the effectiveness cof both word-level and character-level BOW and found that character-level BOW could yield better results than word-level BOW in cyberbullying detection studies. The other crafted features from text input in literature were sentiment feature, emotional feature, word embeddings, psycholinguistics feature, personality traits feature, topic modeling feature, and customized dictionary list feature.

Sentiment and emotional features capture the sentiment, subjectivity, polarity, and expression of emotion embedded in the text. Sentiment analysis presented that text polarity improved the cyberbullying detection model [82]. The Sentiment Informed Cyberbullying Detection (SICD) developed by Dani et al. [83] outperformed other baseline models such as LS, Lasso, and SVM, whereby the learning framework took the sentiment and user relationship information of the post into account. Emotional competencies play an essential role in cyberbullying, and emotion regulation can help capture the characteristics of both bullies and victims [84]. Yuvaraj, et al. [52] incorporated associated emotional features such as the presence of polite words, modal words, unknown words, count of blocked words that contains hate and insult manner, harmful description, aggression form, power difference, the targeted person (one person or more than one person), intention, repetition (once or more than once), sentiment on the racist aspect. Balakrishnan et al. [85] generated emotional features (i.e., anger, fear, joy, sadness, and surprise) using Indico API to identify the emotional expression hidden beneath each Twitter post, and the model turned out to improve performance after the inclusion of these emotional features.

Word embeddings can efficiently reduce feature space for text classification tasks [86]. Zhao et al. [32] proposed a new embedding representation called Embedding-enhanced Bag of Words (EBoW) that resulted from the pre-defined insulting word lists with assigned weights, combining the BoW and latent semantic features in developing the cyberbullying detection model. Agrawal and Awekar [47] work on datasets from multiple platforms that cover cyberbullying topics such as sexism, racism, and attacks by experimenting with deep learning models with different word representations such as GloVe embeddings and SSWE embeddings, where BiLSTM with attention performed the best for the experimental outcome. Wang et al. [79] experimented with several word embeddings such as Word2Vec, GloVe, fastText, Bert, DistilBert, and Sentence Bert (SBert) as the input for cyberbullying detection model. There was consistent outperformance by SBERT when the embeddings were used as the input across the classifiers, which was expected since it worked on the semantic textual similarity (STS) benchmark. The performance of static word embedding (i.e., Word2Vec, GloVe, fastText) was not as optimal as the contextual embeddings from the language models (i.e, RoBerta, XLNet, Albert) when coupled with classifiers for cyberbullying detection [87].

Psycholinguistic features are extracted based on the dictionary or database that reveals the psycholinguistic information from the text. Linguistic Inquiry and Word Count (LIWC) is a text analyzer that quantifies the text's relevance with different psychological aspects to reveal the psychological pattern. Recently, Pennebaker et al. [88]'s team launched the evolved version, LIWC 2022, with several updates compared to the previous version, LIWC15. Dictionaries' flexibility was enhanced, and the tool can now accommodate the pattern of text commonly used on online platforms [89]. The new version includes new categories and dictionaries updated for existing categories with improved psychometric features. Since LIWC 2022 has just been introduced, many researchers might still use the previous LIWC15 tool for cyberbullying detection tasks [23], [61], [90], [91], [92].

Big Five and Dark Triad personality traits are highly prevalent in traditional bullying and cyberbullying [93]. Balakrishnan et al. [42] was the first study that included the Big Five personalities (i.e., Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) and Dark Triad features (i.e., Machiavellianism, Narcissism, Psychopathy) in building the cyberbullying roles (i.e., bully, aggressors, spammer, normal) classification model. However, the IBM Watson Personality Insights API used for the personality traits features extraction was no longer accessible. Hence, no other studies have adopted the same approach so far.

Depending on the context, some studies prepared customized dictionary lists by assembling the terms or words from different sources. In expanding the feature vectors, some studies aggregated the unique list of words or dictionaries, such as insulting words, vulgar words, profane words, abusive words, and bad words, to form features that reflect either the presence or occurrence number of these terms [35], [36], [70], [76], [94], [95], [96]. Mahbub et al. [97] worked on cyberbullying detection that focused on sexual harassment by feeding the model with computed binary representation for the presence of swear words, malevolent words, negative words ahead of swear or malevolent words, and approach words based on a typical list (i.e., words that point to specific approach such as sexual).

Topic modeling was indeed a technique with unsupervised learning. However, some cyberbullying detection studies employed it for feature generation Nahar, et al. [33] improved the model's performance by using unsupervised learning topic modeling to form semantic features with the Latent Dirichlet Allocation (LDA) model and weighted TFIDF that scaled up the bullying feature using the same datasets from CAW 2.0 Workshop. The authors also presented a cyberbullying network graph model to determine the active cyber-bullies and cyber-victims via ranking models. Bigelow, et al. [98] proposed using Latent Semantic Indexing (LSI), which was based on developing a term by document matrix and usually for longer documents in cyberbullying detection. Van Hee et al. [37] trained the LDA and LSI topic models using the crawled corpus that contained seed words that reflect different types of cyberbullying, such as threats, defamation, insults, and defense; however, when fed individually to the cyberbullying model, the performance was underachieved.

As far as this research work was conducted, none of the studies incorporated all the derived features from text input in modeling with the conventional machine learning approach based on the literature for cyberbullying detection. It is an excellent initiative to work on the feature engineering process by including the features derived from the input of text and observing how the features perform when fed into the models. The advancement of word embeddings inspired further exploration of word embeddings, primarily contextual embeddings. More recent word embedding approaches that may work for the task have been introduced, such as BERT and its variants, which have not been explored further in cyberbullying detection. Also, the transfer learning approach was said to be effective in handling the small and imbalanced data used for hate speech detection studies [26].

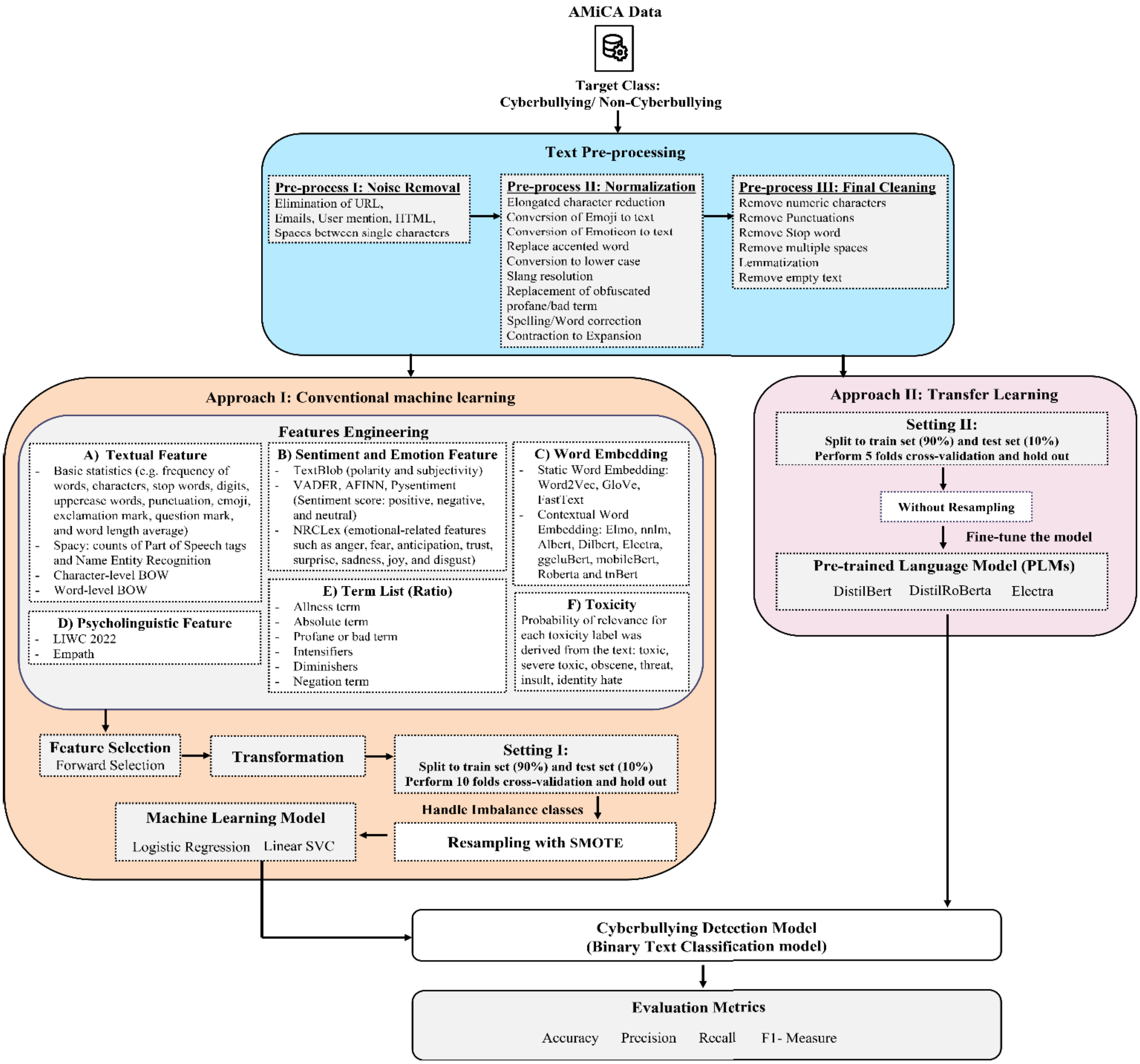

AMiCA Data
Target Class:
Cyberbullying/ Non-Cyberbullying
Text Pre-processing
Pre-process I: Noise Removal
Elimination of URL,
Emails, User mention, HTML,
Spaces between single characters
Pre-process II: Normalization
Elongated character reduction
Conversion of Emoji to text
Conversion of Emoticon to text
Replace accented word
Conversion to lower case
Slang resolution
Replacement of obfuscated
profane/bad term
Spelling/Word correction
Contraction to Expansion
Pre-process III: Final Cleaning
Remove numeric characters
Remove Punctuations
Remove Stop word
Remove multiple spaces
Lemmatization
Remove empty text
Approach I: Conventional machine learning
Features Engineering
A) Textual Feature
- Basic statistics (e.g. frequency of words, characters, stop words, digits, uppercase words, punctuation, emoji, exclamation mark, question mark, and word length average)
- Spacy: counts of Part of Speech tags and Name Entity Recognition
- Character-level BOW
- Word-level BOW
B) Sentiment and Emotion Feature
- TextBlob (polarity and subjectivity)
- VADER, AFINN, Pysentiment (Sentiment score: positive, negative, and neutral)
- NRCLex (emotional-related features such as anger, fear, anticipation, trust, surprise, sadness, joy, and disgust)
C) Word Embedding
- Static Word Embedding: Word2Vec, GloVe, FastText
- Contextual Word Embedding: Elmo, nnlm, Albert, Dilbert, Electra, ggeluBert, mobileBert, Roberta and tnBert
D) Psycholinguistic Feature
- LIWC 2022
- Empath
E) Term List (Ratio)
- Allness term
- Absolute term
- Profane or bad term
- Intensifiers
- Diminishers
- Negation term
F) Toxicity
Probability of relevance for each toxicity label was derived from the text: toxic, severe toxic, obscene, threat, insult, identity hate
Feature Selection
Forward Selection
Transformation
Setting I:
Split to train set (90%) and test set (10%)
Perform 10 folds cross-validation and hold out
Handle Imbalance classes
Resampling with SMOTE
Machine Learning Model
Logistic Regression
Linear SVC
Approach II: Transfer Learning
Setting II:
Split to train set (90%) and test set (10%)
Perform 5 folds cross-validation and hold out
Without Resampling
Fine-tune the model
Pre-trained Language Model (PLMs)
DistilBert
DistilRoBerta
Electra
Cyberbullying Detection Model
(Binary Text Classification model)
Evaluation Metrics
Accuracy
Precision
Recall
F1- Measure

**FIGURE 1.** Overview of research framework on the model development of cyberbullying detection.

However, it is not extensively adopted using the other optimized transformers of Bert in cyberbullying detection studies. The vital takeaways mentioned above suggested some direction to be included in this research.

## III. METHODOLOGY

The framework for cyberbullying detection (binary text classification) using conventional machine learning and transfer learning approaches is shown in Figure 1. The text classification task has six main stages: collecting the corpus, text preprocessing, feature extraction, feature selection, model development, and performance evaluation. Since this research used a secondary dataset, the first stage of corpus collection was replaced with the data preparation after acquiring the data. The framework starts with data preparation and text preprocessing, then branches into two paths to attain the task simultaneously by adopting conventional machine learning and transfer learning approaches, and ends with the model's evaluation. The conventional machine learning and transfer learning approaches with Pre-trained Language Models (PLMs) were adopted. While adopting the conventional machine learning approach, additional work in feature engineering was required to transform the text input into measurable features before feeding it to the model. The following sub-sections further elaborate on the workflows of each stage of the proposed method.

**TABLE 1.** Original class distribution used for cyberbullying detection.

| Text Classification Type | Classes | Posts | % |
|---|---|---|---|
| Binary | Cyberbullying | 5,375 | 4.7 |
| | Non-cyberbullying | 108,319 | 95.3 |

Note: Original total number of posts is 113,694

### A. DATASET

The dataset used in this research was crawled and collected by Van Hee et al. [37] from Ask.fm platform. ASK.fm is very popular among adolescents and has increasingly been used for cyberbullying studies [99]. The data's crawling period ranged from April to October 2013 under the ownership of the Automatic Monitoring for Cyberspace Applications (AMiCA) project in Belgium, which was first introduced in 2018. The original data has two languages: English and Dutch, whereby only the English corpus was used for this research. Table 1 shows the class distribution of the data used for cyberbullying detection. There were other publicly available cyberbullying-related datasets available. Upon comparison, AMiCA dataset was adopted for this research due to following criteria:

- It was an open dataset with significant size, with more recent corpus collection period.
- Annotators have reliable expertise backgrounds. They were made up of four trained linguists in English and Dutch while other open datasets were either annotated manually [36] or by crowdsourcing [100] or by Mechanical Turk [101].
- While most open datasets were lacking information about the annotation process as outlined by Rosa et al. [18], AMiCA has technical guideline[1] with detailed information was prepared for annotators as the main guideline to practice when labeling the data samples, providing a structured annotation process.
- More variation of cyberbullying topics were covered in the corpus, such as curses, defamation, defense, insult, sexual, and threats.

### B. TOOLS AND RESOURCES

The computation works in this research were executed by Python in version 3.9. Python is a versatile and open-source programming language that entails various accessible libraries. The scripts were documented in Jupyter Notebook, a web-based interactive programming platform. The source code is shared in the GitHub repository[2] for the reproducibility of work. The operations were executed by a machine equipped with an AMD Ryzen7 5800H Series processor, GPU with NVIDIA GeForce GTX 1650, 8 cores of CPUs, and 16GB of RAM. For works related to deep learning, high-tier Graphic Processing Units (GPU) is necessary to speed up the training of models. Paperspace Gradient Notebooks[3] was mainly utilized for transfer learning works as it offers complete access to JupyterLab, higher tier GPUs, higher RAM, and availability. It has become an appealing alternative to support machine requirements for running deep learning works with a free basic account. The Pro-plan subscription was required to support the transfer learning tasks in this research. The system selection is as follows: GPU with RTX 5000 having 16 GB RAM of GPU power, CUDA of version 11.0, and 30 GB RAM. The other paid tool was Linguistics Inquiry and Word Count (LIWC) of version 2022, a popular tool for natural language processing widely adopted by researchers.

### C. DATA PREPARATION AND TEXT PREPROCESSING

For any application relating to Natural Language Processing works, the raw text data need to be cleaned as the cleaned textual data are the elementary input to feed into any text application [102]. The AMiCA dataset was shared in Brat Repositories' stand-off document format, which requires further data wrangling. The BratReader[4] code was revamped to parse and integrate the text and annotation label from the Brat Repositories document. The blueprints built for the text preprocessing pipeline for the text data are shown in Figure 1. A python package containing the text preprocessing modules utilized in this research was created, available in the GitHub repository.[5] The preprocessing steps are divided into three parts. The first part involves the preliminary step to clean and remove noise from the text data, such as URLs, emails, username mention, HTML elements, multiple spaces, newline symbols, and the symbol that precedes each post as part of the indicator used during the annotation process. Space between consecutive single characters (e.g., 'W H A T') was removed, and the characters were concatenated.

The second part involves text normalization to return the base form of text. Regular expression was designed to reduce the elongated characters (e.g., 'youuuuuuuu') and tackle those with redundant symbols. Subsequently, the emojis and emoticons embedded in the text were transformed into the text to preserve the emotional expression with the help of the 'emot' package. The counts of emojis and emoticons for each post were computed. Accented characters were normalized to the standard English alphabet, and all characters were converted to lowercase for better word embedding representation. Any slang terms were resolved by referring to the slang term list used in online chat,[6] text messaging,[7] and social platforms.[8] Cleaning online posts becomes a big hurdle due to spelling mistakes, slang terms, and modified words commonly seen online because the words are deliberately changed when posting the content to avoid being

[1] https://www.lt3.ugent.be/media/uploads/publications/2015/Guidelines_Cyberbullying_TechnicalReport_1.pdf

[2] https://github.com/HwaiTengTeoh/NLP_TextClassification_Cyberbullying_Detection

[3] https://www.paperspace.com/

[4] https://github.com/clips/bratreader

[5] https://github.com/HwaiTengTeoh/preprocess_text

[6] https://slangit.com/terms/online_chat

[7] https://slangit.com/terms/text_messaging

[8] https://slangit.com/terms/social_media

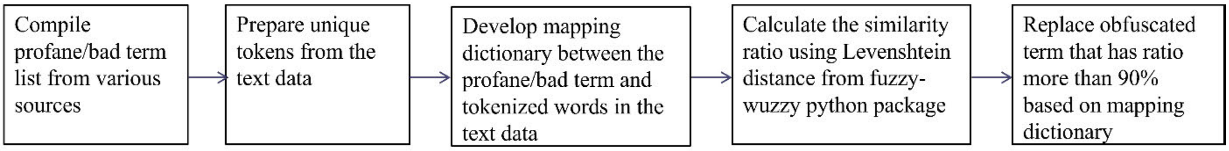

Compile profane/bad term list from various sources
Prepare unique tokens from the text data
Develop mapping dictionary between the profane/bad term and tokenized words in the text data
Calculate the similarity ratio using Levenshtein distance from fuzzy-wuzzy python package
Replace obfuscated term that has ratio more than 90% based on mapping dictionary

**FIGURE 2.** **Process flow of replacing obfuscated profane and bad terms in text preprocessing.**

flagged by the online platform's filtering mechanism. Hence, a robust preprocessing pipeline was included to handle such variations to replace the obfuscated profane term, as shown in Figure 2. The general idea is to compute Levenshtein distance from the 'FuzzyWuzzy' library to match strings to determine the similarity of string patterns. The rule was set to replace the term with the matched term in the mapping dictionary if it was found to precede the 90% ratio. The threshold was experimented with and found suitable for handling most cases. The spelling checker, LanguageTool, an open-source grammar and spelling checking tool, was used to help correct the spelling of each text word. To connect the LanguageTool's server, the associated package, namely 'language-tool-python', was installed in the Python environment. In conjunction with that, the contractions in the text were expanded using the 'pycontraction' package that works robustly based on the context.

The final part involves removing numeric, punctuation, and multiple spaces. This research retained stopwords in the text. Some words in the stopword lists, such as the negation term 'not' and pronouns, help understand the context of cyberbullying [79]. The next step was to perform lemmatization to convert the word to the base form except for pronouns and keep the original form for the word with a lemma resulting in 'be'. After several experiments, the proposed sequence was the best arrangement for text preprocessing. The sequence of the steps should not interfere, as early removal of any will disrupt the flow of the text preprocessing process. For example, early punctuation removal may delete part of the emoticon structure, and the removal of numeric might modify the slang term that incorporates numbers such as '2day' (resolve to 'today') and '2morrow' (resolve to 'tomorrow'). The last step is to remove any empty text, which results in 112,247 posts after text preprocessing.

### D. FEATURE ENGINEERING

Feature engineering can help improve machine learning models' performance by discovering significant features or deriving new features from existing ones [103]. Based on the reviewed literature, a few features constructed from the text were justified: textual feature, sentiment and emotion feature, word embeddings, psycholinguistics feature, personality traits feature, topic modeling, and customized dictionary list. This research used the above features except for personality traits and topic modeling. The first exclusion was due to the deprecation of the IBM Watson Personality API.[9] The second exclusion was due to the employment of unsupervised techniques, which is not within the scope of this study.

#### 1) TEXTUAL FEATURES

The pattern in the text were captured and represented statistically by generating basic statistics, including frequency of words, characters, stop words, digits, uppercase words, punctuation, emoji, exclamation mark, question mark, and word length average. The counts of each Part of Speech (POS) tag and Named Entity Recognition (NER) label using spacy were also included. The 'CountVectorizer' library was applied to generate BOW. Character-level inputs of BOW can better capture rare words and are robust to variation of typos, which are the usual scenario in the online social platform [104]. This research employed word-level and character-level Bag of Words (BOW) with up to quadruple grams, presuming these feature combinations would help characterize the text pattern better and improve the performance of machine learning models. Each unique n-gram of word and character made up the attribute for training the model.

#### 2) SENTIMENT AND EMOTION FEATURES

With the assistance of packages available in Python, several sentiment and emotional-related features were extracted from the text. The TextBlob package analyzed the polarity and subjectivity of text. The sentiment score: positive, negative, and neutral, were generated from different tools, namely VADER by Hutto and Gilbert [105], AFINN by Nielsen [106], and Pysentiment,[10] given that these tools were developed upon different dictionaries. NRCLex package was adopted to generate metrics of eight elementary emotional-related features from the text: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. The package was developed based on the National Research Council Canada (NRC) affect lexicon introduced by Mohammad and Turney [107].

#### 3) STATIC WORD EMBEDDING

Though BOW is popularly used to represent text characteristics, it still lacks contextual information as the unique word in texts is deemed an individual unit without considering the semantic meaning [108]. Word embeddings represent words

[9] https://www.ibm.com/docs/SSTTDS_11.0.0/com.ibm.ace.icp.doc/localconn_ibmwatsonpi.html

[10] https://nickderobertis.github.io/pysentiment/api/pysentiment2.html

into vectors that extract information for words with similar meanings in vector space.

Word2Vec, developed by Mikolov et al. [109] from Google, implements a continuous bag of words (CBOW) to predict the word corresponding to the context taken from each word. Even though this embedding is not robust to rare words, it is still widely used for various text classification tasks as it can capture words with semantic meaning. The pre-trained models from the 'spacy' package extracted the embeddings with 300 dimensions.

Unlike word2vec, GloVe is an unsupervised learning model that takes account of the frequency of words' co-occurrences when building word embeddings with word vectors that relate to the co-occurrence probability of words in the corpus [110]. For comparison purposes, this research employed pre-trained word vectors trained from tokens in Common Crawl, Wikipedia, and Twitter which come with several dimensions (100, 200). The word vectors are publicly available in the repository owned by Standford NLP.[11] Hence, the GloVe can handle out-of-vocabulary or rare words better since the word can be divided into n-grams characters to obtain the embeddings.

On the other hand, FastText was developed by Facebook, built on top of the word2vec model that encodes words in n-gram characters. This research uses the word vectors with one million sizes trained on the corpus from Wikipedia 2017 site, UMBC web base, and news site [111].

### 4) CONTEXTUAL WORD EMBEDDING

Contextual word embeddings represent words based on the context and give promising results in various natural language processing tasks, including text classification [112]. There are two ways of utilizing the pre-trained model: fine-tuning or taking the outcome obtained as a feature to solve a particular task [113]. Inspired by the authors, this research recognized the benefits of extracting embeddings from PLMs. Due to the rich library of PLMs available in the TensorFlow hub, this research explored several PLMs to generate the word embeddings from the contextual language models.

Elmo was proposed by Peters et al. [114] with the word vectors computed by a two-layer bidirectional language model in left-to-right and right-to-left contexts. Unlike static embeddings, Elmo can capture the context of the word from the sentence by generating different embeddings for the same word but with a different context. However, Elmo is shallowly bidirectional as it cannot simultaneously consider left and right contexts. The mean of the Elmo vectors for the constituent words from the post was computed to obtain the Elmo embedding. Figure 3 (a) shows the process flow of extracting Elmo embeddings. Nnlm embeddings are the token-based text embeddings trained on English Google News 200B corpus. Similarly, the sentence embeddings were computed by averaging the word vectors.

The release of Bert by Devlin et al. [115] from Google brings a new benchmark to the natural language processing community. Bert was designed as a purely bidirectional model that effectively extracts information from right and left contexts with Masked Language Modeling (MLM) implementation, randomly hides the tokens, and replaces them with a mask [MASK]. The unmasked words surrounding the masked word are used for prediction. Several transformer-based embeddings models of Bert variants were proposed to accommodate the needs of different tasks: Albert [116], Dilbert [117], Electra [118], ggeluBert [119] mobile-Bert [120], Roberta [121] and tnBert [122]. The contextual word embeddings derived from Bert and its variants were extracted the same way as illustrated in Figure 3 (b).

### 5) PSYCHOLINGUISTICS FEATURES

This work is the first to adopt the new version of the LIWC 2022 cyberbullying detection tool with 117 attributes. In addition, this research also employed the Empath library developed by Fast et al. [123], a text analysis tool similar to LIWC, to derive categories from a small number of seed terms. Empath provides text analysis for about 200 categories constructed from different web datasets.

### 6) TERM LISTS

The following term list was compiled to form six feature categories, with each category as one feature. Similar features were adopted by Van Hee et al. [37], Jacobs et al. [23], and Ali et al. [78] by binarizing the feature for any presence of a term from the separate lists. The terms ratio was computed, as it yielded a better outcome than binarization and counts after experiments.

- **Allness term.** An individual tends to communicate extremely when he or she feels emotional using allness terms [124]. Osgood and Walker [125] deduced the high usage of allness terms (i.e., always, never, forever, no one, and no more) when an individual was experiencing emotional affect when speaking or writing. The allness term was compiled from Tytko and Augstkalns [126]'s study.
- **Absolute term.** Absolute terms are words that should not be modified. They were consolidated from the resources: i) List of absolute,[12] ii) Grammar: Absolute Word site[13] and Absolute adjectives.[14]
- **Profane or bad term.** Post with cyberbullying traits are often associated with hate and embedded with the profane word [40], [75], [127]. The list of terms was assembled from various sources: i) Offensive and profane term list,[15] ii) bad word list from Google archive,[16] iii) Swear and curse word dictionary,[17] iv) dirty, naughty,

[11] https://github.com/stanfordnlp/GloVe

[12] http://nomistakespublishing.com/writing-resources/list-of-absolutes/

[13] https://kddidit.com/2015/04/20/grammar-absolute-words/

[14] https://kathysteinemann.com/Musings/absolute/

[15] https://www.cs.cmu.edu/∼ biglou/resources/bad-words.txt

[16] https://code.google.com/archive/p/badwordslist/downloads

[17] https://www.noswearing.com/dictionary/

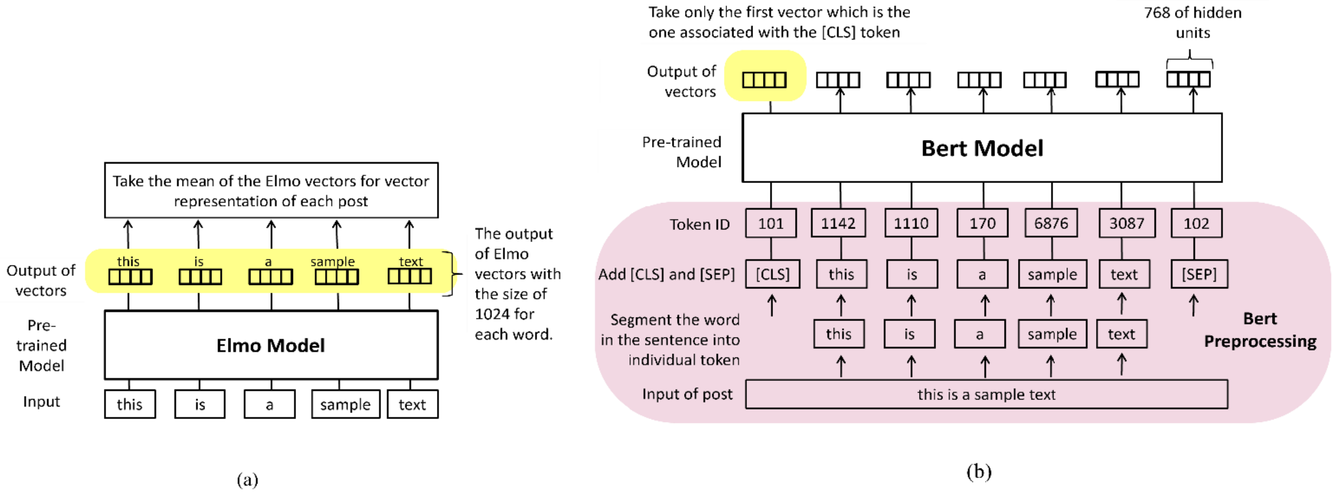


**FIGURE 3.** Process flow for generating (a) Elmo and nnlm sentence vector embeddings and (b) Word embeddings from bert model and its variant.

obscene, and bad word list from Github repository.[18] After resorting, a list of the profane or bad terms was collected, also used in obfuscated profane term replacement.

Intensifiers, negations, and diminishers are valence shifters that can alter the text sentiment orientation, reversing the initial polarity of words [128].

- **Intensifiers.** In the case of intensifiers, the intensity of the whole expression increased. From a grammar perspective, adjectives are deemed natural intensifiers [129]. The list of intensifiers (i.e., too, so, quite, extremely, and fantastically) was assembled from the study by Strohm [130] and the GitHub repository[19] by manual collection, and the One Minute English blog[20]
- **Diminishers.** The presence of the diminisher term (i.e., little, rarely, partially) reduces the intensity of the whole expression and weakens the strength of the word that comes after it. The list of diminishers terms was taken from the study by Strohm [130].
- **Negation term.** Sommar and Wielondek [131] observed that statement with negative polarity is often embedded with negated positive words. The presence of negation terms (i.e., not and no) contradicted the expression's meaning and changed its polarity. The same resources for getting the list of intensifier terms were accessed to compile the list of negation terms.

#### 7) TOXICITY FEATURES

The toxicity levels of text: Toxic, severe toxic, obscene, threat, insult, identity hate were extracted using the detoxify library developed by Laura Hanu from Unitary. It is a library trained upon the framework to classify toxic comments of three Jigsaw challenges: Toxic comment classification, Unintended Bias in Toxic comments, and Multilingual toxic comment classification. The architecture of the work was also available in the HuggingFace Transformer library. This library benefited the research community and was incorporated by different studies to attain other tasks [132], [133]. The probability of relevance for each toxicity label was derived from the text: toxic, severe toxic, obscene, threat, insult, identity hate, making up six attributes germane toxicity of the post.

### E. FEATURE SELECTION

After the feature engineering process, the features were selected by the forward selection method to form the different combinations of features permutated by feature types. Overall, this research consists of around 620k attributes. Table 2 shows the list of features implemented in this research.

### F. DATA TRANSFORMATION

An extra transformation step was introduced before feeding the features constructed into the conventional machine learning model. The features were scaled individually by the 'MaxAbsScaler' function from Python's scikit-learn library, with the maximum absolute value of each attribute assigned as 1. This scaler was chosen as it preserves the sparsity of the feature without shifting or centering the data.

### G. DATA SPLITTING

The 'StratifiedKFold' method from the scikit-learn package was used for cross-validation, whereas the 'train_test_split' method was used for the hold-out method. '1127' was set as the random number seed in this research. The hold-out method splits the data into train and test sets by setting 10% of the data as the test set. In developing a conventional machine learning model, this research performed a ten-fold

[18] https://github.com/chucknorris-io/swear-words/blob/master/en

[19] https://github.com/zengyan-97/Sentiment-Lexicon/blob/master/intensifier.txt

[20] https://oneminuteenglish.org/en/list-intensifiers/

**TABLE 2.** Feature groupings for cyberbullying detection.

| Label | Feature grouping | Attributes | Attributes number |
|---|---|---|---|
| A | Textual | Word-level BOW | 611,604 |
| | | Character-level BOW | 10,297 |
| | | Handcrafted textual statistics<br>- Frequency of words, characters, stop words, digits, uppercase words, punctuation, emoji, exclamation mark, question mark, and word length average<br>- Counts of Part of Speech (POS) tag<br>- Counts of Named Entity Recognition (NER) label | 32 |
| B | Sentiment and Emotional | TextBlob’s sentiment polarity and subjectivity | 2 |
| | | Vader’s negative score, neutral score, positive score, and compound score | 4 |
| | | General Inquirer’s positive score, negative score, polarity score, and subjectivity score | 5 |
| | | AFINN's sentiment | 1 |
| | | NRCLEX’s fear score, anger score, anticipation score, trust score, surprise score, positive score, negative score, sadness score, disgust score, joy score | 10 |
| C | Word Embedding | Static word embeddings | |
| | | - Word2Vec embeddings | 300 |
| | | - GloVe embeddings Wikipedia | 100 |
| | | - GloVe embeddings Common42B | 300 |
| | | - GloVe embeddings Common840B | 300 |
| | | - GloVe embeddings Twitter100 | 100 |
| | | - GloVe embeddings Twitter200 | 200 |
| | | - GloVe embeddings Twitter25 | 25 |
| | | - GloVe embeddings Twitter50 | 50 |
| | | - FastText embeddings | 300 |
| | | Contextual word embeddings | |
| | | - Albert embeddings | 300 |
| | | - Bert embeddings | 768 |
| | | - DistilBert embeddings | 768 |
| | | - Electra-small embeddings | 768 |
| | | - Elmo embeddings | 1,024 |
| | | - ggeluBert embeddings | 768 |
| | | - mobileBert embeddings | 512 |
| | | - nnlm embeddings | 128 |
| | | - RoBerta embeddings | 1,024 |
| | | - tnBert embeddings | 768 |
| D | Psycho-linguistics | LIWC 2022 | 117 |
| | | Empath | 200 |
| E | Term Lists (Ratio) | Allness term (Ratio) | 1 |
| | | Absolute term (Ratio) | 1 |
| | | Profane or bad term (Ratio) | 1 |
| | | Intensifiers (Ratio) | 1 |
| | | Diminishers (Ratio) | 1 |
| | | Negation (Ratio) | 1 |
| F | Toxicity | Relevance probability of six toxicity level<br>- Toxic, severe toxic, obscene, threat, insult, identity hate | 6 |

cross-validation method in which 10% of the data was used as the test set in each fold. Due to heavy computational requirements, the five folds cross-validation method was performed to fine-tune the PLMs with a similar proportion of test data in each fold.

### H. RESAMPLING WITH SMOTE

The original training set had a highly imbalanced class distribution, with about 95% non-cyberbullying and 5% cyberbullying instances. In dealing with an imbalanced dataset, this research applied algometric resampling methods, SMOTE, to mitigate the class imbalance problem. Synthetic Minority Over-sampling Technique (SMOTE), introduced by Chawla et al. [134], is an oversampling method that creates samples by interpolation, where new data points are created within the range of available data points.

**TABLE 3.** Class distribution of training data after resampling with SMOTE for cyberbullying detection model.

| Target class | Before SMOTE (original) | After SMOTE (resampled) |
|---|---|---|
| Cyberbullying (0) | 4,837 | 14,427 |
| Non-cyberbullying (1) | 96,185 | 96,185 |

The training data is resampled with SMOTE by setting the resampled ratio of minority class samples over the majority class samples as 0.15, increasing the cyberbullying instance up to 13%. The ratio was selected since it produced the best performance in the experiment run. The new distribution of classes for cyberbullying detection after SMOTE resampling is shown in Table 3.

**TABLE 4.** Hyperparameters used in fine-tuning the PLMs for cyberbullying detection.

| Hyperparameter | DistilBert | DistilRoBerta | Electra-small |
|---|---|---|---|
| Repository path | distilbert-base-uncased | distilroberta-base | google/electra-small-discriminator |
| Batch size for training | 8 | 8 | 8 |
| Batch size for testing | 8 | 8 | 8 |
| Maximum sequence length | 512 | 514 | 512 |
| Activation function | gelu | gelu | gelu |
| Learning rate | 0.00005 | 0.00005 | 0.00005 |
| Dropout probability | 0.1 | 0.1 | 0.1 |

## I. MODEL DEVELOPMENT

This research considered two approaches for the model development of the text classification tasks: i) conventional machine learning and ii) transfer learning approach using Pre-trained Language Models (PLMs).

### 1) CONVENTIONAL MACHINE LEARNING APPROACH

This research considered two conventional machine learning models: Support Vector Machine and Logistic Regression. These models were selected due to their frequent usage in text classification models [135], [136], and both are suited for classification tasks with large and high dimensionality attributes [137].

Logistic regression classifies the categories of text based on logistic function. The model evaluates the coefficients for each attribute input and predicts the text class it belongs [138]. In addition, the model assumed the absence of a linear relationship between the target label and attributes [139]. Different Logistic Regression types were used based on the number of classes available in the target variable. For cyberbullying detection task, this research applied binary logistic regression.

SVM finds a suitable hyperplane that optimally separates two different data classes away from each other [137]. The linear kernel of SVM works well with text classification, and the model takes the support vectors that fall at the boundary between two classes to determine the maximized margin for the separation of classes [140]. It can also use multi-class classification by training the model with one-vs-rest models. This research applied the Linear SVC classification model for cyberbullying detection task.

### 2) TRANSFER LEARNING APPROACH

The PLMs based on deep learning encourage transfer learning, a process of reusing or acquiring knowledge from the model trained on specific tasks to solve other tasks [141]. Table 4 summarizes the default hyperparameters settings applied when fine-tuning the PLMs.

The plain text was further preprocessed so that the input format fit the language model, as illustrated in Figure 4, which employed a transfer learning approach to fine-tune the PLMs for downstream tasks on cyberbullying detection. The preprocessing step encompassed the following sub-steps:

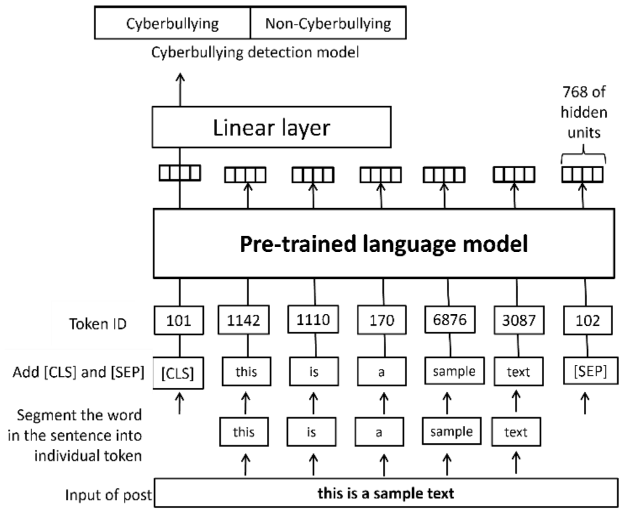


**FIGURE 4.** A transfer learning approach to fine-tune PLMs for cyberbullying detection.

- Firstly, the text was tokenized to generate a token for each word.
- Next, the [CLS] token was added at the start of the sentence.
- Then, the [SEP] token was appended at the end of the sentence.
- Convert the token to its unique id.

Each token was mapped with its corresponding embeddings, resulting in a 768 vector representation. The [CLS] token embeddings were passed through the linear layer for class prediction to attain cyberbullying detection from the given post. All three pre-trained models: DistilBert, DistilRoberta, and Electra-small, have specific tokenizer classes associated with each of them. The pre-trained models and tokenizer classes were publicly accessible in the HuggingFace Transformers library [142].

With the transfer learning approach, this research fine-tuned the PLMs: DistilBert, DistilRoBerta, and Electra-small. According to Sanh et al. [117], DistilBert is the distilled version of the Bert-base model trained by distillation knowledge techniques. Due to the significant data size, this research decided to fine-tune DistilBert since the model's size is reduced by 40%, and the training speed is increased by 60% compared to Bert, which requires lesser training resources.

The architecture of DistilRoBerta was developed with six layers, 768 dimensions, 12 heads, 82 million parameters, and pre-trained on OpenWebTextCorpus [117]. Instead of adopting Roberta, the distilled version of it, DistilRoBerta, was employed in this research. The speed is double faster as the training data used is four times lesser with minor usage of parameters compared to Roberta.

Instead of the masked language modeling (MLM) technique, Electra adopts the Replace Token Detection (RTD) technique that replaces the input tokens with possible alternatives sampled from the small generator network [118]. The

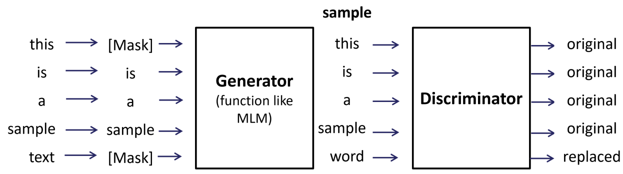


**FIGURE 5.** Architecture of electra-small model that works in generator-discriminator model.

discriminative model predicts if the generator will replace the token. This approach can overcome the limitation of the masking approach by the Bert models [87]. The architecture of Electra is presented in Figure 5. For this research, the optimized version of Electra-small was employed.

## IV. RESULTS AND DISCUSSION

The principle of selecting the best model and feature combinations was based on the highest F-measure of cyberbullying instances for cyberbullying detection. Instead of individually considering the precision and recall scores, the F-measure provided a way to capture precision and recall into a single measure.

### A. APPROACH I: CONVENTIONAL MACHINE LEARNING APPROACH FOR CYBERBULLYING DETECTION

In implementing the conventional machine learning approach, the combination of models and features that best accomplished the task of detecting cyberbullying posts was determined. This research experimented with Logistic Regression and Linear SVC models by feeding the proposed features to train the English corpus of Ask.fm posts. The experimentation for the conventional machine learning approach for cyberbullying detection was conducted under the original data size and resampled data using SMOTE, with ten-fold cross-validation and hold-out methods.

#### 1) INDIVIDUAL FEATURE GROUPING

As the concern for detecting cyberbullying instances appeared to be more prominent, the decision should be based on how well the model and features used in detecting the cyberbullying posts. It was inaccurate to deduce which model trained with the individual feature was performing better as the proportion of cyberbullying and non-cyberbullying classes was highly imbalanced. In that case, the experiments run on training data resampled with SMOTE increased the proportion of cyberbullying classes in the training data. The adjustments were intended to make the models learn more about cyberbullying posts. As expected, the performance metrics of the individual feature groupings were increased. Table 5 shows the performance evaluation metrics results for the positive class of cyberbullying detection model developed by Logistic Regression and Linear SVC, trained with the individual feature grouping under the original sample and resampled data with SMOTE.

Under original training samples, the overall performance of the Logistic Regression model fed with the individual feature grouping was not satisfying in predicting the interested class, cyberbullying. The Logistic Regression model had a better ability to classify the cyberbullying class after SMOTE resampling. Among the umbrella grouping of features, Textual features (A) formed with word-level and character-level n-grams vectorizers with the other crafted textual-related statistics attributes appeared to be the most powerful feature for both models. The highest cross-validated F-measure achieved by the Logistic Regression was 58.29% using the Textual features (A). Moreover, it achieved the highest hold-out F-measure, 59.91% using the Textual features (A). Furthermore, not much difference was found in the cross-validated and hold-out results, indicating that the trained models generalized well on the testing data.

Various types of word embeddings have been experimented with, and only the one with the best performance was incorporated for feature combination in the following experiments. Exceptionally for mobileBert and RoBerta embeddings, the other contextual word embeddings outperformed the static word embeddings (word2vec embedding, GloVe embedding, and fasttext embedding). Both mobilBert and RoBerta embeddings performed poorly in detecting posts with cyberbullying traits. Since the transformer used for the Roberta model accommodated cased text, the input of uncased text in generating the embeddings might affect the performance. Furthermore, word embeddings derived from mobileBert appeared unfit to classify the cyberbullying post since the precision, recall, and F-measure were nearly zero. Compared to the static word embeddings, the GloVe model trained with Twitter corpus, resulting in 200 dimensions of vectors, produced the highest F-measure for detecting the cyberbullying class. The outcome indicated that word embeddings extracted from a model trained using an online platform corpus, i.e., Twitter, better fit the linguistics style for AMiCA posts. As for the word embeddings formed from the variants of Bert, embeddings of DistilBert. ggeluBert and tnBert achieved almost similar high F-measure scores to the base transformer. As a highlight, DistilBert embeddings yielded the highest F-measure and Recall metrics. Hence, the DistilBert embeddings were retained to combine with other features in the following experiments.

Psycholinguistics (D), Term Lists (Ratio) (E), Sentiment, and Emotion (B) features performed poorly when each was fed individually to train the models. Their metrics scores achieved with Logistic Regression were slightly better than Linear SVC. The performance metrics of the Linear SVC model developed individually with these individual groups of features were less than satisfactory, as presented earlier using the original samples, as the performance was just minimally improved after applying SMOTE resampling. The F-measure metric resulting from cross-validation using Logistic Regression trained with Sentiment and Emotion features (B) was improved from 6.22% to 25.26% after resampling with SMOTE. On the flip side, the Toxicity features (F) achieved

**TABLE 5. Performance evaluation metrics of individual feature grouping for cyberbullying detection using logistic regression and linear SVC (Cyberbullying class).**

| | | | Original | | | | | | SMOTE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Cyberbullying class (positive class) | | | | | | Cyberbullying class (positive class) | | | | | |
| | | | Cross-validation (10 folds) | | | Hold-out | | | Cross-validation (10 folds) | | | Hold-out | | |
| **M** | **Cat.** | **Feature** | **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| Logistic Regression | Standalone Grouping | **Textual** | 77.54 | 36.89 | **49.96** | 77.61 | 37.36 | **50.44** | 69.14 | 50.44 | **58.29** | 71.28 | 51.67 | **59.91** |
| | | Sentiment and Emotion | 53.32 | 3.31 | 6.22 | 41.67 | 2.79 | 5.23 | 31.91 | 20.93 | 25.26 | 30.75 | 21.38 | 25.22 |
| | | Psycholinguistic | 57.72 | 13.10 | 21.32 | 59.50 | 13.38 | 21.85 | 46.94 | 35.35 | 40.31 | 45.00 | 33.46 | 38.38 |
| | | Term Lists (Ratio) | 25.12 | 2.29 | 4.19 | 25.64 | 1.86 | 3.47 | 29.97 | 9.75 | 14.70 | 34.36 | 10.41 | 15.98 |
| | | **Toxicity** | 62.66 | 35.33 | **45.16** | 61.76 | 35.13 | **44.79** | 48.12 | 57.60 | **52.41** | 49.06 | 58.18 | **53.23** |
| | Static Word Embeddings | Word2Vec Embedding | 60.38 | 18.27 | 28.03 | 59.76 | 18.77 | 28.57 | 45.14 | 37.84 | 41.16 | 47.23 | 39.59 | 43.07 |
| | | GloVe Embedding Wikipedia | 49.31 | 7.59 | 13.12 | 48.10 | 7.06 | 12.32 | 35.90 | 20.04 | 25.71 | 33.44 | 18.96 | 24.20 |
| | | GloVe Embedding Common42B | 60.39 | 17.71 | 27.37 | 54.60 | 16.54 | 25.39 | 45.53 | 37.58 | 41.16 | 47.31 | 37.55 | 41.87 |
| | | GloVe Embedding Common840B | 60.51 | 18.38 | 28.17 | 58.90 | 17.84 | 27.39 | 44.91 | 37.69 | 40.97 | 47.32 | 41.08 | 43.98 |
| | | GloVe Embedding Twitter100 | 58.41 | 16.19 | 25.33 | 60.00 | 16.73 | 26.16 | 43.40 | 34.44 | 38.39 | 44.24 | 34.94 | 39.04 |
| | | GloVe Embedding Twitter200 | 58.67 | 18.46 | 28.06 | 60.98 | 18.59 | 28.49 | 45.46 | 38.60 | 41.74 | 48.68 | 41.08 | 44.56 |
| | | GloVe Embedding Twitter25 | 49.42 | 7.20 | 12.56 | 45.35 | 7.25 | 12.50 | 33.49 | 19.22 | 24.39 | 33.76 | 19.70 | 24.88 |
| | | GloVe Embedding Twitter50 | 56.78 | 12.26 | 20.15 | 54.62 | 12.08 | 19.79 | 40.76 | 28.56 | 33.56 | 40.05 | 28.44 | 33.26 |
| | | FastText Embedding | 50.40 | 1.90 | 3.65 | 57.14 | 2.97 | 5.65 | 27.80 | 4.95 | 8.40 | 36.78 | 5.95 | 10.24 |
| | Contextual Word Embeddings | Albert Embedding | 69.47 | 21.28 | 32.56 | 72.06 | 18.22 | 29.08 | 44.50 | 39.07 | 41.55 | 43.90 | 40.15 | 41.94 |
| | | Bert Embedding | 69.61 | 30.98 | 42.82 | 73.56 | 28.44 | 41.02 | 49.19 | 49.88 | 49.50 | 51.09 | 47.96 | 49.47 |
| | | **DistilBert Embedding** | 71.05 | 37.34 | **48.88** | 72.40 | 37.55 | **49.45** | 52.29 | 56.22 | **54.13** | 50.95 | 54.83 | **52.82** |
| | | Electra-small Embedding | 69.61 | 30.98 | 42.82 | 73.56 | 28.44 | 41.02 | 49.19 | 49.88 | 49.50 | 51.09 | 47.96 | 49.47 |
| | | Elmo Embedding | 63.10 | 27.66 | 38.43 | 63.41 | 29.00 | 39.80 | 52.84 | 48.39 | 50.49 | 53.00 | 47.58 | 50.15 |
| | | ggeluBert Embedding | 70.28 | 33.02 | 44.86 | 68.75 | 30.67 | 42.42 | 51.47 | 51.78 | 51.60 | 51.98 | 51.30 | 51.64 |
| | | mobileBert Embedding | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 17.42 | 0.33 | 0.66 | 44.44 | 0.74 | 1.46 |
| | | nnlm Embedding | 69.57 | 22.33 | 33.78 | 67.90 | 20.45 | 31.43 | 51.11 | 47.42 | 49.18 | 49.81 | 48.88 | 49.34 |
| | | RoBerta Embedding | 57.01 | 9.84 | 16.77 | 53.75 | 7.99 | 13.92 | 43.38 | 28.87 | 34.63 | 43.54 | 28.81 | 34.68 |
| | | tnBert Embedding | 70.72 | 37.15 | 48.66 | 68.09 | 35.69 | 46.83 | 51.82 | 54.92 | 53.29 | 51.95 | 54.46 | 53.18 |
| Linear SVC | Standalone Grouping | **Textual** | 67.02 | 49.08 | **56.64** | 68.29 | 52.04 | **59.07** | 58.54 | 56.45 | **57.45** | 61.63 | 59.11 | **60.34** |
| | | Sentiment and Emotion | 46.67 | 0.19 | 0.37 | 66.67 | 0.37 | 0.74 | 39.22 | 18.34 | 24.98 | 37.02 | 18.03 | 24.25 |
| | | Psycholinguistic | 57.88 | 9.15 | 15.78 | 65.17 | 10.78 | 18.50 | 48.42 | 30.31 | 37.25 | 48.41 | 31.04 | 37.83 |
| | | Term Lists (Ratio) | 24.34 | 2.16 | 3.96 | 23.68 | 1.67 | 3.13 | 29.08 | 8.50 | 13.15 | 34.01 | 9.29 | 14.60 |
| | | **Toxicity** | 63.00 | 34.38 | **44.47** | 63.39 | 34.76 | **44.90** | 49.13 | 55.70 | **52.19** | 50.08 | 56.51 | **53.10** |
| | Static Word Embeddings | Word2Vec Embedding | 65.18 | 12.86 | 21.45 | 65.09 | 12.83 | 21.43 | 47.77 | 32.63 | 38.75 | 49.32 | 33.46 | 39.87 |
| | | GloVe Embedding Wikipedia | 55.53 | 3.65 | 6.83 | 66.67 | 5.20 | 9.66 | 38.63 | 15.18 | 21.77 | 37.07 | 14.13 | 20.46 |
| | | GloVe Embedding Common42B | 65.70 | 12.58 | 21.09 | 66.02 | 12.64 | 21.22 | 47.92 | 31.83 | 38.23 | 49.42 | 31.78 | 38.69 |
| | | GloVe Embedding Common840B | 64.95 | 12.89 | 21.49 | 65.42 | 13.01 | 21.71 | 47.49 | 32.33 | 38.46 | 48.90 | 32.90 | 39.33 |
| | | GloVe Embedding Twitter100 | 61.55 | 9.99 | 17.17 | 61.36 | 10.04 | 17.25 | 47.02 | 28.86 | 35.74 | 50.47 | 29.74 | 37.43 |
| | | GloVe Embedding Twitter200 | 64.99 | 12.60 | 21.08 | 66.67 | 12.64 | 21.25 | 48.77 | 33.67 | 39.83 | 51.91 | 35.32 | 42.04 |
| | | GloVe Embedding Twitter25 | 54.49 | 2.31 | 4.42 | 52.17 | 2.23 | 4.28 | 36.81 | 14.08 | 20.35 | 39.62 | 15.61 | 22.40 |
| | | GloVe Embedding Twitter50 | 62.68 | 6.98 | 12.53 | 58.46 | 7.06 | 12.60 | 43.63 | 23.13 | 30.20 | 43.92 | 24.16 | 31.18 |
| | | FastText Embedding | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 38.88 | 2.83 | 5.27 | 46.67 | 3.90 | 7.20 |
| | Contextual Word Embeddings | Albert Embedding | 76.33 | 18.03 | 28.95 | 80.18 | 16.54 | 27.43 | 48.72 | 46.75 | 46.84 | 53.37 | 38.29 | 44.59 |
| | | Bert Embedding | 77.67 | 23.27 | 35.76 | 79.75 | 23.42 | 36.21 | 51.39 | 49.26 | 50.30 | 51.48 | 45.35 | 48.22 |
| | | **DistilBert Embedding** | 75.40 | 32.89 | **45.76** | 75.34 | 31.23 | **44.15** | 53.67 | 56.05 | **54.80** | 53.38 | 54.28 | **53.82** |
| | | Electra-small Embedding | 77.67 | 23.27 | 35.76 | 79.75 | 23.42 | 36.21 | 51.39 | 49.26 | 50.30 | 51.48 | 45.35 | 48.22 |
| | | Elmo Embedding | 66.13 | 24.33 | 35.54 | 68.93 | 26.39 | 38.17 | 52.14 | 46.66 | 49.21 | 52.55 | 45.91 | 49.01 |
| | | ggeluBert Embedding | 75.99 | 28.56 | 41.47 | 78.80 | 26.95 | 40.17 | 52.49 | 50.90 | 51.64 | 52.47 | 51.30 | 51.88 |
| | | mobileBert Embedding | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | | nnlm Embedding | 75.44 | 15.16 | 25.23 | 75.24 | 14.68 | 24.57 | 51.75 | 45.93 | 48.65 | 51.11 | 47.21 | 49.08 |
| | | RoBerta Embedding | 56.08 | 4.45 | 8.23 | 55.26 | 3.90 | 7.29 | 44.20 | 26.03 | 32.73 | 42.05 | 23.61 | 30.24 |
| | | tnBert Embedding | 77.25 | 31.76 | 44.96 | 77.88 | 31.41 | 44.77 | 53.21 | 54.40 | 53.76 | 52.46 | 53.53 | 52.99 |

Note:
- M = Model, Cat. = Category of Feature, P = Precision, R = Recall and F = F-measure
- Bold values are the highest value across models and categories of feature

competitive performance in cyberbullying detection, with cross-validated F-measure scores of 52.41% for the Logistic Regression model. Though both models provided distinct values for each evaluation metric, the performance ranking of the individual feature grouping in detecting cyberbullying posts appeared to be consistent. Again, this finding denoted the importance of features used to build the model, and a proper model selection would enhance the performance.

## 2) FEATURES COMBINATION

To further boost the performance of each model, the experiments were extended to train different combinations of features. Table 6 shows the performance evaluation metrics for cyberbullying classes with the top eight combinations of features using Logistic Regression and Linear SVC with and without SMOTE resampling. When putting the features in different sets of combinations, the performance of the classification models was significantly boosted, especially in the case of Logistic Regression. It was clear that Textual features (A) and Toxicity features (F) contributed the most to the classification model since they appeared in all combination sets.

The inclusion of DistilBert embeddings (C) further enhanced the performance of the models. Although the

**TABLE 6.** Performance evaluation metrics of top 8 features combination for cyberbullying detection using logistic regression and linear SVC (Cyberbullying class).

| | | Original | | | | | | SMOTE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Cyberbullying class (positive class) | | | | | | Cyberbullying class (positive class) | | | | | |
| | | Cross-validation (10 folds) | | | Hold-out | | | Cross-validation (10 folds) | | | Hold-out | | |
| **M** | **Feature combination** | **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| Logistic Regression | A + C + D + F | 75.11 | 50.25 | 60.16 | 75.93 | 49.26 | 59.75 | 67.22 | 58.60 | **62.56** | 69.09 | 58.18 | 63.17 |
| | A + C + E + F | 75.01 | 49.79 | 59.78 | 75.49 | 50.37 | 60.42 | 66.86 | 57.82 | 61.95 | 66.46 | 58.92 | 62.46 |
| | A + C + F | 75.13 | 50.68 | 60.44 | 76.69 | 50.74 | 61.07 | 67.03 | 58.21 | 62.28 | 67.02 | 58.18 | 62.29 |
| | A + B + C + F | 74.81 | 50.81 | 60.46 | 76.19 | 50.56 | 60.78 | 66.85 | 58.36 | 62.28 | 67.96 | 58.74 | 63.01 |
| | A + C + D + E + F | 75.04 | 50.01 | 59.97 | 75.97 | 51.12 | 61.11 | 66.47 | 58.60 | 62.25 | 69.66 | 57.62 | 63.07 |
| | A + B + C + D + E + F | 75.56 | 50.59 | 60.54 | 75.27 | 50.93 | 60.75 | 66.40 | 58.42 | 62.12 | 69.45 | 58.74 | 63.65 |
| | **A + B + C + D + F** | 75.12 | 51.16 | **60.81** | 76.40 | 50.56 | 60.85 | 66.25 | 58.36 | 62.02 | 70.13 | 60.22 | **64.80** |
| | **A + B + C + E + F** | 75.31 | 50.83 | 60.64 | 76.97 | 50.93 | **61.30** | 66.82 | 57.84 | 61.97 | 67.96 | 58.74 | 63.01 |
| Linear SVC | A + C + D + F | 67.58 | 53.00 | 59.36 | 70.77 | 54.46 | 61.55 | 63.34 | 55.85 | 59.33 | 66.16 | 56.32 | 60.84 |
| | A + C + E + F | 67.68 | 53.41 | 59.67 | 70.17 | 54.65 | 61.44 | 62.63 | 55.89 | 59.04 | 65.69 | 58.36 | 61.81 |
| | A + C + F | 67.92 | 53.49 | 59.81 | 69.78 | 54.09 | 60.94 | 63.08 | 56.45 | 59.55 | 65.54 | 57.62 | 61.33 |
| | **A + B + C + F** | 68.05 | 53.69 | **59.99** | 69.95 | 54.09 | 61.01 | 62.79 | 56.26 | **59.31** | 66.81 | 57.62 | **61.88** |
| | A + C + D + E + F | 67.59 | 53.11 | 59.44 | 70.63 | 54.09 | 61.26 | 63.13 | 55.64 | 59.12 | 66.59 | 57.43 | 61.68 |
| | A + B + C + D + E + F | 67.78 | 53.13 | 59.53 | 71.32 | 54.09 | 61.52 | 62.93 | 55.65 | 59.03 | 66.81 | 57.25 | 61.66 |
| | **A + B + C + D + F** | 67.86 | 53.13 | 59.56 | 71.74 | 54.28 | **61.80** | 63.24 | 55.78 | 59.24 | 65.86 | 55.58 | 60.28 |
| | A + B + C + E + F | 67.88 | 53.51 | 59.80 | 69.64 | 53.72 | 60.65 | 62.51 | 56.09 | 59.09 | 65.96 | 57.99 | 61.72 |

Note:
- M = Model, P = Precision, R = Recall and F = F-measure
- Bold values are the highest value across resampling methods, models, and combination of features

Psycholinguistics features (D), Term Lists (Ratio) features (E), and Sentiment and Emotional features (B) performed poorly on their own, there were minor additive impacts on the performance of the models when these features were integrated with other feature groups.

From the perspective of the hold-out testing, the current best system of Logistic Regression with SMOTE was reported to incorporate a combination of Textual features (A), Sentiment and Emotion features (B), DistilBert embeddings (C), Psycholinguistics (D), and Toxicity features (F) with the F-measure of 64.8%. The cross-validated test result yielded F-measure of 62.56% with the same set of features but excluding Sentiment and Emotion features (B). The best combination of features that worked well with Linear SVC after resampling was reported to be Textual features (A), Sentiment and Emotion features (B), DistilBert embeddings (C), and Toxicity features (F) for cross-validated and hold-out test results, with F-measure of 59.3% and 61.9% respectively. The F-measures scores were not improved much compared with Linear SVC trained with original data samples.

While the performance of Logistic Regression appeared to be boosted after being resampled with SMOTE, there were just little inconspicuous changes in the metrics scores of the Linear SVC model. The cross-validated F-measure result for the Linear SVC model did not surpass the best system of Logistic Regression. This result deduced that the feature combination and Logistic Regression worked better with SMOTE resampling than Linear SVC. The outperformance of the Logistic Regression after SMOTE resampling was one of the significant findings in this research. The performance of the Logistic Regression model was significantly improved after feeding with different combinations of features, with the F-measure hitting at least 59%. The performance of recall metrics traded off with better precision scores of over 75% using Logistic Regression. The best feature combination in the cross-validated result for Logistic Regression encompassed: Textual features (A), Sentiment and Emotion features (B), DistilBert Embeddings (C), Psycholinguistics features (D), and Toxicity features (F), whereas the best feature combination in hold-out result encompassed: Textual features (A), Sentiment and Emotion features (B), DistilBert Embeddings (C), Term Lists (Ratio) features (E) and Toxicity features (F).

**TABLE 7.** Comparison of epoch training for fine-tuning of optimized PLM versus base PLM.

| Optimized PLM | Iteration/second | Base PLM | Iteration/second |
|---|---|---|---|
| DistilBert | 4.71it/s | Bert | 2.75 it/s |
| DistilRoBerta | 4.83it/s | RoBerta | 2.78 it/s |
| Electra-small | 9.33it/s | Electra-base | 2.80 it/s |

### B. APPROACH II: TRANSFER LEARNING APPROACH FOR CYBERBULLYING DETECTION

As discussed earlier, the pre-trained language models (PLMs) exempted the tedious processes of developing features for text classification tasks from scratch. Table 7 compares the execution of epoch training by iteration per second for the optimized PLMs and their respective base form. Since more extended epoch training was required for the base PLMs during initial experiments, this research opted to work merely with the optimized PLM such as DistilBert, DistilRoBerta, and Electra-small. The transfer learning approach was found robust to the data imbalance issue in the reviewed literature, so PLMs were fine-tuned with the original data size. The work implementation adopted the default hyperparameters setting. The only optimization performed was the number

**TABLE 8.** Performance evaluation metrics for cyberbullying detection by fine-tuning PLMs at different epoch numbers (Cyberbullying class).

| | | Cyberbullying class (positive class) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Cross-validation (5 folds) | | | Hold out | | |
| **PLM** | **Epoch Number** | **P** | **R** | **F** | **P** | **R** | **F** |
| DistilBert | 1 | 75.18 | 58.07 | 64.96 | 74.52 | 72.30 | 73.40 |
| | 2 | 76.51 | 58.03 | 65.50 | 76.00 | 70.63 | 73.22 |
| | 3 | 77.06 | 60.04 | 67.19 | 76.46 | 68.22 | 72.10 |
| | **4** | 75.60 | 61.71 | **67.90** | 73.89 | 71.00 | 72.42 |
| | 5 | 76.25 | 58.88 | 66.35 | 72.41 | 72.68 | 72.54 |
| | 6 | 73.84 | 60.67 | 66.55 | 76.37 | 67.29 | 71.54 |
| | 7 | 74.22 | 59.78 | 65.74 | 77.27 | 66.36 | 71.40 |
| | 8 | 74.12 | 61.41 | 67.02 | 78.27 | 65.61 | 71.39 |
| DistilRoBerta | 1 | 73.14 | 55.50 | 62.91 | 76.67 | 61.71 | 68.38 |
| | 2 | 71.32 | 59.78 | 64.40 | 73.65 | 60.78 | 66.60 |
| | **3** | 72.90 | 62.86 | **67.24** | 73.16 | 66.36 | 69.59 |
| | 4 | 74.66 | 60.37 | 66.61 | 81.46 | 57.99 | 67.75 |
| | 5 | 76.76 | 56.17 | 64.49 | 79.38 | 61.52 | 69.32 |
| | 6 | 75.72 | 58.44 | 65.67 | 81.17 | 56.88 | 66.89 |
| | 7 | 74.79 | 60.22 | 66.66 | 78.38 | 61.34 | 68.82 |
| | 8 | 75.63 | 58.74 | 66.03 | 78.92 | 59.85 | 68.08 |
| Electra-small | 1 | 75.68 | 47.29 | 56.07 | 59.78 | 49.44 | 54.12 |
| | 2 | 78.02 | 39.44 | 48.67 | 57.03 | 52.04 | 54.42 |
| | 3 | 74.05 | 55.61 | 63.41 | 64.10 | 49.44 | 55.82 |
| | **4** | 74.17 | 56.91 | **64.32** | 60.21 | 53.16 | 56.47 |
| | 5 | 69.41 | 59.44 | 63.95 | 65.51 | 49.07 | 56.11 |
| | 6 | 74.27 | 56.10 | 63.87 | 73.72 | 45.35 | 56.16 |
| | 7 | 73.40 | 56.88 | 64.05 | 68.15 | 51.30 | 58.54 |
| | 8 | 69.85 | 59.63 | 64.25 | 68.05 | 51.86 | 58.86 |

Note:
- PLM = Pre-trained Language Model, P = Precision, R = Recall and F = F-measure
- Bold values are the highest value across epochs and PLMs

of epochs used to fine-tune the PLMs by running up to eight epochs with five-fold cross-validation. Then, the epoch number was determined based on the highest F-measure of cross-validation runs in detecting cyberbullying class. Table 8 presents promising results for the metrics of cyberbullying detection tasks developed by fine-tuning the DistilBert, DistilRoBerta, and Electra-small across eight epochs. The results show the credibility of the PLMs in distinguishing cyberbullying posts even though the distribution of classes was imbalanced. At four epochs, the fine-tuned DistilBert resulted in the highest overall F-measure metric of 83.22% and 85.53% under five-fold cross-validation and hold-out testing. When looking at the prediction for cyberbullying class, the highest F-measure metric of 67.24% and 69.59% were obtained at three epochs under five-fold cross-validation and hold-out testing, respectively. The fine-tuned Electra-small resulted in the highest F-measure metric of 64.32% and 56.47% for cyberbullying class at four epochs under five-fold cross-validation and hold-out testing, respectively.

### C. COMPARISON BETWEEN APPROACHES FOR CYBERBULLYING DETECTION

Table 9 shows the comparative hold-out testing's performance metrics using conventional machine learning and transfer learning approaches based on the model's binary classes and overall performance. As for the conventional machine learning approach, more computational time was required when training models with more dimensions, such as the inclusion of character-level features [143]. Logistic regression outperformed Linear SVC with higher precision, recall, and F-measure. By inspecting the training and testing time of the model, Logistic Regression was found to be more efficient than linear SVC, especially when the dimensions of features increase. Although the best combination feature sets (A + B + C + D + F) fed into Logistic Regression were greater than Linear SVC (A + B + C + F), it still took less computation time than linear SVC. The duration for developing the models with SMOTE resampling is gauged and recorded in Table 10. The results proved that Logistic Regression appeared to be more efficient when handling high dimensionality of features than linear SVC, which was in line with Thangaraj and Sivakami [144], who also found Logistic regression to train faster and better performance than linear SVC. Also, resampling with SMOTE helped mitigate the class imbalance problem, improving the model's performance in classifying cyberbullying posts.

DistilBert and DistilRoBerta easily outperformed the conventional machine learning models without resampling. It was known that Bert's transformer model was more extensive and consumed more computational resources. This research adopted that optimized version of Bert, such as DistilBert and DistilRoBerta, using transfer learning approach, marked new milestones for the performance of the cyberbullying detection model in classifying the AMiCA posts. Electra-small's hold-out metric scores were unexpectedly worse than the conventional machine learning models. By referring to the recall metrics, just 29.00% and 33.64% of cyberbullying posts were misclassified as non-cyberbullying posts by the DistilBert and DistilRoBerta models. DistilBert was the best system for cyberbullying detection tasks, resulting in higher predictive power overall with fewer false positives and false negatives. Almost

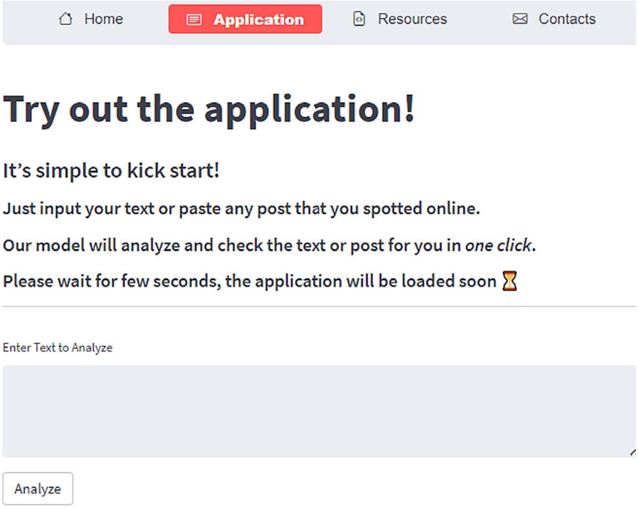
**TABLE 9.** Best performance evaluation metrics for cyberbullying detection.

| | | | | Cyberbullying class | | | Non-Cyberbullying class | | | Overall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Approach | M/PLM | RS | Best Feature Combination | P | R | F | P | R | F | A | MP | MR | MF |
| Conventional ML | **Logistic Regression** | SMOTE | A + B + C + D + F | **70.13** | **60.22** | **64.80** | **98.01** | **98.71** | **98.36** | **96.86** | **84.07** | **79.47** | **81.58** |
| | Linear SVC | SMOTE | A + B + C + F | 66.81 | 57.62 | 61.88 | 97.88 | 98.56 | 98.22 | 96.60 | 82.35 | 78.09 | 80.05 |
| Transfer Learning | **DistilBert** | Original | Not applicable | **73.89** | **71.00** | **72.42** | **98.54** | **98.74** | **98.64** | **97.41** | **86.22** | **84.87** | **85.53** |
| | Electra-small | Original | Not applicable | 60.21 | 53.16 | 56.47 | 97.66 | 98.23 | 97.94 | 96.07 | 78.93 | 75.70 | 77.20 |
| | DistilRoBerta | Original | Not applicable | 73.16 | 66.36 | 69.59 | 98.31 | 98.77 | 98.54 | 97.22 | 85.74 | 82.57 | 84.07 |

Note:
- M = Model, PLM = Pre-trained Language Model, RS = Resampling, A = Accuracy, MP = Macro Precision, MR = Macro Recall, MF = Macro F-measure, P = Precision, R = Recall and F = F-measure
- Bold values are the highest values obtained by the approach

**TABLE 10.** Time taken for development of cyberbullying detection model with logistic regression and linear SVC.

| | | | Cross Validation | | | Hold Out | |
|---|---|---|---|---|---|---|---|
| Model | Feature Combinations | RS | Number of Folds | Second/ iteration | Duration of Time | Second/ iteration | Duration of Time |
| Logistic Regression | A + B + C + D + F | SMOTE | 10 | 1189.41 | 19 minutes | 55.12 | 55 seconds |
| Logistic Regression | A + B + C + F | SMOTE | 10 | 1166.73 | 19 minutes | 55.65 | 56 seconds |
| Linear SVC | A + B + C + D + F | SMOTE | 10 | 1753.14 | 29 minutes | 88.57 | 88 seconds |
| Linear SVC | A + B + C + F | SMOTE | 10 | 1570.92 | 23 minutes | 87.00 | 85 seconds |



Try out the application!

It's simple to kick start!

Just input your text or paste any post that you spotted online.

Our model will analyze and check the text or post for you in *one click*.

Please wait for few seconds, the application will be loaded soon ⏳

**FIGURE 6.** Screenshot of cyberbullying checker application developed with the fine-tuned DistilBert.

half of the cyberbullying posts are classified as non-cyberbullying posts for Electra-small. Figure 6 shows the screenshot of the Cyberbullying Checker Application developed Using the best system to help to detect posts that exhibit cyberbullying traits automatically. The Cyberbullying Checker Application can be assessed from the application website: https://hwaitengteoh-nlp-textclassification-cyb-cb-detection-app-vm9xk8.streamlit.app/

### D. BENCHMARKING WITH PREVIOUS STUDIES FOR CYBERBULLYING DETECTION

Table 11 consolidates all the best performance metric scores on the cyberbullying class that the previous researcher adopted with AMiCA corpus data for cyberbullying detection tasks. The best-performing model was the cost-sensitive SVM developed by Van Hee et al. [37], with the highest F-measure of 64.26% resulting from ten-fold cross-validation and the corresponding F-measure of 63.69% from hold-out testing. Hence, their research was set as the first benchmark for cyberbullying detection model. The authors then applied cascading ensemble model for the cyberbullying detection task in their subsequent research, but the hold-out F-measure achieved was merely 60.02%, which was lower than the formal research [23]. Despite that, the true positive rate of the cascading model was significantly improved to 70.06%. However, there was a trade-off with lower precision scores, resulting in no improvement in the F-measure score.

Rathnayake et al. [68]'s result was eliminated for comparison as the researchers were applying pre-trained models for hate speech detection tasks, and they adopted the model to run the hold-out test directly without fine-tuning. The class distribution for their testing set was curated to include all the cyberbullying posts, making up one-third of the testing set.

Their work application revealed similar linguistic traits between hate speech and cyberbullying corpus, which can be a new direction for future studies. Albeit the performance metrics of the training sets achieved by Ali et al. [145] were promising, the models were overfitted as they failed to generalize in the hold-out test set.

The best three models in this research outperformed the former benchmark result of the hold-out testing and created a new benchmark for cyberbullying detection using AMiCA data. With the TL approach, the DistilBert model fine-tuned in this research outperformed the previous best system using cost-sensitive SVM by an increment of 8.73% and 3.64% of the F-measure metric for the cross-validation and hold-out test results, respectively.

**TABLE 11.** Comparison of cyberbullying class's metrics with previous studies using AMiCA dataset for cyberbullying detection.

| | | | | Cross-validation (10 folds) | | | | Hold Out | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Previous studies** | **M/PLM** | **RS** | **Best Feature Combination** | **A** | **P** | **R** | **F** | **A** | **P** | **R** | **F** |
| Van Hee, et al. [37] | SVM | Cost-Sensitive | Character n-grams (A),<br>Subjectivity lexicons (B),<br>Term Lists (E),<br>Topic model | 96.97 | 73.32 | 57.19 | 64.26 | 97.21 | 74.13 | 55.82 | 63.69 |
| Jacobs, et al. [23] | Cascading ensemble | Random Downsampling | NA | NA | NA | NA | NA | NA | 52.49 | 70.06 | 60.02 |
| Ali, et al. [145] | DT | None | NA | NA | NA | NA | NA | 94.42 | 42.33 | 30.24 | 35.28 |
| Ali, et al. [145] | SVM | None | NA | NA | NA | NA | NA | 93.73 | 38.14 | 39.64 | 38.87 |
| Ali, et al. [78] | SVM | None | Word TFIDF (A),<br>Character TFIDF (A),<br>Word Embedding (C) | NA | NA | NA | NA | 96.57 | 75.23 | 44.86 | 56.21 |
| **This research** | **DistilBert** | **Original** | **NA** | **97.22** | **75.60** | **61.71** | **67.90** | **97.41** | 73.89 | **71.00** | **72.42** |
| **This research** | **DistilRoBerta** | **Original** | **NA** | **97.10** | 72.90 | **62.86** | **67.24** | **97.22** | 73.16 | **66.36** | **69.59** |
| **This research** | **Logistic Regression** | **SMOTE** | **A + B + C + D + F** | 96.58 | 66.25 | **58.36** | 62.02 | 96.86 | 70.13 | **60.22** | **64.80** |

Note:
- M = Model, PLM = Pre-trained Language Model, A = Accuracy, P = Precision, R = Recall and F = F-measure
- Bold values are the values achieved by the best three models developed in this research

**TABLE 12.** Examples of misclassified posts by the best cyberbullying detection model using DistilBert.

| No | Post | True label | Predicted label |
|---|---|---|---|
| 1 | ew | CB | NCB |
| 2 | jew | CB | NCB |
| 3 | school gay | CB | NCB |
| 4 | post a picture of your little undie duh | CB | NCB |
| 5 | i am surprised you have friend you always try and copy them with everything they wear and buy do they not get annoy at you | CB | NCB |

Note:
- CB: Cyberbullying, NCB: Non-cyberbullying
- Pre-processed text of the post was displayed

Another significant finding was the introduction to adopt Logistic Regression with SMOTE resampling that incorporated Textual features (A), Sentiment and Emotion features (B), DistilBert embeddings (C), Psycholinguistics (D), and Toxicity features (F). It was found to train faster than Linear SVC, saving time in training the model using conventional machine learning approach.

### E. MISCLASSIFIED POSTS

This section discusses the misclassified post for the cyberbullying detection task by taking the best hold-out result from the fine-tuned DistilBert model. The misclassification was probably due to the intuitive cyberbullying context embedded within the text's linguistics. Table 12 displays the misclassified cyberbullying posts by the best cyberbullying detection model. Although DistilBert was found to perform optimistically in correctly detecting posts with cyberbullying traits, about 29% of them were still misclassified as non-cyberbullying. It was observed that the false negative was probably due to the lack of explicit cyberbullying context presented in the post because some misclassified cyberbullying posts were short and without many expressions. Some examples were presented in Posts 1 to 4 of Table 12.

Surprisingly, the short posts (embedded with sexualism and racism conversations were misclassified, for example, Post 2 to Post 4. Although the posts give a feeling of sexual and racial harassment, the model failed to classify it as cyberbullying due to the short post length and insufficient context of the entire conversation. The following example includes the post with an ironic tone and scornful context, as shown in Post 7, since Ironic-related context often has ambiguous interpretations. Not much research analyzes the relationship between irony with cyberbullying context.

## V. CONCLUSION AND FUTURE WORK

Cyberbullying is an unexpected ramification of technological advancement, which can bring destructive consequences to any Internet user. Automatic detection is essential for the prevention and reduction to curve out the act from spreading. Although the textual feature remains popular and primarily used to attain cyberbullying classification, this research was interested in exploring the features that could be crafted from text and shed light on the methodological steps to adopt textual features, sentiment and emotional features, embeddings, psycholinguistics features, term lists features, and toxicity features. The feature engineering process was part of the conventional machine learning approach. Feed-forward selection fed different feature groupings into the model to form a different combo.

Different cyberbullying detection models were developed using Logistic Regression and Linear SVC by conventional machine learning approach and by fine-tuning DistilBert, DistilRoBerta, and Electra-small via epochs training to attain the task, which was a form of binary text classification. The first approach required finding the best feature combination for cyberbullying detection. Logistic Regression and Linear SVC were trained under original data and data with SMOTE resampling. Textual features, DistilBert embeddings, and toxicity features were the primary features based on the experiments while fed into models individually based on their umbrella grouping, with F-measures of 58.29%, 54.13%, and 52.14%, respectively. Although the individual performance of sentiment and emotion features was just with F-measure of 40.31%, there was an additive effect on the model's performance when these features were combined. Contextual embeddings were found to perform better than static embeddings in representing the contextual meaning of the cyberbullying post. However, not all contextual embeddings produced competitive outcomes. Based on the experimentation results, embeddings derived from mobileBert and RoBerta performed poorly in the tasks. DistilBert, tnBert, and ggeluBert were performing well, but DistilBert embeddings were retained as the embeddings yielded slightly better performance among the three.

The experimentation result of the conventional machine learning approach showed that Logistic Regression worked better than Linear SVC in detecting cyberbullying posts after resampling with SMOTE. Logistic Regression appeared to work better with higher feature dimensionality than Linear SVC. When the dimensionality of features was getting higher, the performance of Linear SVC fallback and longer time was required for the training process. To our best understanding, this paper is the first to incorporate the Toxicity features, brand-new psycholinguistics features by Linguistic Inquiry and Word (LIWC) 2022 tool, and Empath's lexicon for derivation of psycholinguistics features as well as the rarely used contextual word embeddings such as tnBert, mobileBert and ggeluBert for the development of machine learning models. Our findings also opened the doors for other research to use the same lexicons from the brand-new LIWC 2022 and Empath as well as the toxicity features from 'detoxify' library as part of feature engineering process. Due to the outstanding features performance, this research's conventional machine learning models outperformed the benchmark of previous works under the same approach. This achievement proved that the proposed feature sets were a step forward for the conventional machine learning approach in working with cyberbullying detection.

The result was proven to be optimistic with these features. From the perspective of the hold-out experiment runs, the highest F-measure was obtained by coupling Logistic Regression with SMOTE resampling and a combination of textual features, sentiment and emotion features, DistilBert embeddings, psycholinguistics features, and toxicity features ($A + B + C + D + F$ ) was 64.8% for cyberbullying detection. This research also supported that toxicity detection should be incorporated into cyberbullying detection from the perspective of model evaluation.

This research has explored the adoption of transfer learning in building cyberbullying detection model. The optimized version of the Bert model was utilized, with the promising result obtained by DistilBert, DistilRoBerta, and Electra-small under default settings using the original data size. DistilBert and Electra-small were suitably trained with four epoch numbers, whereas DistilRoBerta was suitably trained with three epoch numbers. The experiment shows that DistilBert performed excellently in the text classification task since it achieved the highest F-measures under four epochs. Distilbert was the best model for cyberbullying detection, with the highest F-measure, 72.42%, in classifying posts with cyberbullying traits. Due to the outperformance of the model, DistilBert was also promoted for its fast-training time (which processes about four to five iterations per second when training each epoch), stable performance achievement in just four epochs, and it was robust to imbalance binary classes. Finally, a Cyberbullying Checker application was built using the best system.

When compared with transfer learning approach, using a conventional machine learning approach could be challenging since much effort was put into the feature engineering process. Lack of the requirement of domain expertise might hinder the feature engineering process of developing a task-specific text classification model. Aside from extensive feature engineering, conventional machine learning approach requires significant amount of labeled data. Unlike transfer learning approach, fine-tuned PLMs using smaller context-specific datasets even under imbalanced class distribution could yield with uplifted performance as experimented in this study. By adopting transfer learning approach, steps prior conventional machine learning model development such as resampling of data and feature engineering were skipped as transfer learning was known to be robust with imbalanced dataset and better generalization since the PLMs were pre-trained on diverse data that can accommodate downstream tasks. The proposed approach of transfer learning with lighter transfomers by fine-tuning the optimized version of PLMs (i.e DistilBert, DistilRoBerta, Electra-small) was proven to be performing better than conventional machine learning approach. Moreover, they were proven to be more feasible over their respective base PLMs (i.e Bert, RoBerta and Electra-base), with shorter computation time and optimistic performance. This finding proved that optimized PLMs could be considered since they are less computationally expensive.

In short, the main contributions of this research are different for both approaches. The findings of adopting conventional machine learning approach are as follows: (i) Introduction of toxicity features that boosted the conventional machine learning models, (ii) Utilization of LIWC 2022 and Empath library for psycholinguistics features, (iii) Exploration of contextual word embeddings from other Bert variants such as tnBert, mobileBert and ggeluBert that yielded competitive performance as DistilBert embeddings, with the later performed slightly better, (iv) Logistic Regression outperformed Linear SVC and dealt efficiently with multidimensional attributes with lesser training time. On the other hand, the findings of transfer learning approach are as follows: (i) The effort to fine-tune PLMs that made up of lighter transformers was reduced compared to base transformers, (ii) Fine-tuned DistilBert yielded the best performance for cyberbullying detection task though the dataset distribution was skewed, which denoted that transfer learning was robust to imbalanced dataset.

**TABLE 13. Performance Evaluation Metrics of Individual Feature Grouping for Cyberbullying Detection Using Logistic Regression and Linear SVC (Non-cyberbullying class).**

| | | | Original | | | | | | SMOTE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Non-cyberbullying class (negative class) | | | | | | Non-cyberbullying class (negative class) | | | | | |
| | | | Cross-validation (10 folds) | | | Hold-out | | | Cross-validation (10 folds) | | | Hold-out | | |
| M | Cat. | Feature | P | R | F | P | R | F | P | R | F | P | R | F |
| Logistic Regression | Standalone Grouping | Textual | 96.91 | 99.46 | 98.17 | 96.93 | 99.46 | 98.18 | 97.54 | 98.86 | 98.20 | 97.60 | 98.95 | 98.27 |
| | | Sentiment and Emotion | 95.36 | 99.85 | 97.55 | 95.33 | 99.80 | 97.51 | 96.09 | 97.75 | 96.91 | 96.10 | 97.58 | 96.83 |
| | | Psycholinguistic | 95.79 | 99.52 | 97.62 | 95.80 | 99.54 | 97.64 | 96.79 | 97.99 | 97.39 | 96.69 | 97.94 | 97.31 |
| | | Term Lists (Ratio) | 95.30 | 99.66 | 97.43 | 95.28 | 99.73 | 97.45 | 95.61 | 98.86 | 97.21 | 95.64 | 99.00 | 97.29 |
| | | Toxicity | 96.82 | 98.94 | 97.87 | 96.80 | 98.91 | 97.84 | 97.85 | 96.87 | 97.36 | 97.87 | 96.96 | 97.41 |
| | Static Word Embeddings | Word2Vec Embedding | 96.03 | 99.40 | 97.68 | 96.05 | 99.36 | 97.68 | 96.90 | 97.69 | 97.29 | 96.98 | 97.77 | 97.38 |
| | | GloVe Embedding Wikipedia | 95.54 | 99.61 | 97.54 | 95.51 | 99.62 | 97.52 | 96.07 | 98.20 | 97.12 | 96.01 | 98.10 | 97.04 |
| | | GloVe Embedding Common42B | 96.00 | 99.41 | 97.68 | 95.94 | 99.31 | 97.60 | 96.89 | 97.74 | 97.31 | 96.89 | 97.89 | 97.39 |
| | | GloVe Embedding Common840B | 96.03 | 99.40 | 97.69 | 96.00 | 99.37 | 97.66 | 96.89 | 97.67 | 97.28 | 97.05 | 97.70 | 97.37 |
| | | GloVe Embedding Twitter100 | 95.93 | 99.42 | 97.65 | 95.95 | 99.44 | 97.67 | 96.74 | 97.74 | 97.24 | 96.76 | 97.78 | 97.27 |
| | | GloVe Embedding Twitter200 | 96.04 | 99.35 | 97.66 | 96.04 | 99.40 | 97.69 | 96.94 | 97.67 | 97.30 | 97.06 | 97.82 | 97.44 |
| | | GloVe Embedding Twitter25 | 95.52 | 99.63 | 97.53 | 95.52 | 99.56 | 97.50 | 96.02 | 98.08 | 97.04 | 96.04 | 98.05 | 97.04 |
| | | GloVe Embedding Twitter50 | 95.75 | 99.53 | 97.61 | 95.74 | 99.49 | 97.58 | 96.46 | 97.92 | 97.18 | 96.45 | 97.86 | 97.15 |
| | | FastText Embedding | 95.29 | 99.90 | 97.54 | 95.34 | 99.89 | 97.56 | 95.41 | 99.36 | 97.34 | 95.46 | 99.49 | 97.43 |
| | Contextual Word Embeddings | Albert Embedding | 96.17 | 99.53 | 97.82 | 96.03 | 99.64 | 97.80 | 96.95 | 97.55 | 97.25 | 97.00 | 97.42 | 97.21 |
| | | Bert Embedding | 96.62 | 99.32 | 97.95 | 96.51 | 99.49 | 97.97 | 97.48 | 97.41 | 97.44 | 97.39 | 97.69 | 97.54 |
| | | DistilBert Embedding | 96.92 | 99.23 | 98.06 | 96.93 | 99.28 | 98.09 | 97.79 | 97.42 | 97.60 | 97.72 | 97.34 | 97.53 |
| | | Electra-small Embedding | 96.62 | 99.32 | 97.95 | 96.51 | 99.49 | 97.97 | 97.48 | 97.41 | 97.44 | 97.39 | 97.69 | 97.54 |
| | | Elmo Embedding | 96.46 | 99.19 | 97.81 | 96.52 | 99.16 | 97.82 | 97.42 | 97.82 | 97.62 | 97.37 | 97.88 | 97.62 |
| | | ggeluBert Embedding | 96.72 | 99.30 | 97.99 | 96.60 | 99.30 | 97.93 | 97.57 | 97.54 | 97.56 | 97.55 | 97.61 | 97.58 |
| | | mobileBert Embedding | 95.21 | 99.99 | 97.54 | 95.21 | 100.00 | 97.54 | 95.22 | 99.91 | 97.51 | 95.24 | 99.95 | 97.54 |
| | | nnlm Embedding | 96.22 | 99.51 | 97.84 | 96.13 | 99.51 | 97.79 | 97.37 | 97.72 | 97.54 | 97.43 | 97.52 | 97.47 |
| | | RoBerta Embedding | 95.65 | 99.63 | 97.60 | 95.56 | 99.65 | 97.56 | 96.48 | 98.10 | 97.28 | 96.48 | 98.12 | 97.29 |
| | | tnBert Embedding | 96.91 | 99.23 | 98.06 | 96.84 | 99.16 | 97.98 | 97.73 | 97.43 | 97.58 | 97.70 | 97.46 | 97.58 |
| Linear SVC | Standalone Grouping | Textual | 97.47 | 98.78 | 98.12 | 97.61 | 98.78 | 98.20 | 97.81 | 97.99 | 97.90 | 97.95 | 98.15 | 98.05 |
| | | Sentiment and Emotion | 95.22 | 99.99 | 97.54 | 95.22 | 99.99 | 97.55 | 96.00 | 98.57 | 97.27 | 95.98 | 98.46 | 97.20 |
| | | Psycholinguistic | 95.62 | 99.66 | 97.60 | 95.69 | 99.71 | 97.66 | 96.56 | 98.38 | 97.46 | 96.59 | 98.33 | 97.45 |
| | | Term Lists (Ratio) | 95.29 | 99.66 | 97.43 | 95.27 | 99.73 | 97.45 | 95.56 | 98.96 | 97.23 | 95.59 | 99.09 | 97.31 |
| | | Toxicity | 96.77 | 98.98 | 97.87 | 96.79 | 98.99 | 97.88 | 97.76 | 97.10 | 97.43 | 97.80 | 97.16 | 97.48 |
| | Static Word Embeddings | Word2Vec Embedding | 95.79 | 99.65 | 97.68 | 95.78 | 99.65 | 97.68 | 96.67 | 98.20 | 97.43 | 96.70 | 98.27 | 97.48 |
| | | GloVe Embedding Wikipedia | 95.37 | 99.86 | 97.56 | 95.44 | 99.87 | 97.60 | 95.86 | 98.79 | 97.30 | 95.81 | 98.79 | 97.28 |
| | | GloVe Embedding Common42B | 95.77 | 99.67 | 97.68 | 95.77 | 99.67 | 97.68 | 96.63 | 98.25 | 97.43 | 96.63 | 98.36 | 97.49 |
| | | GloVe Embedding Common840B | 95.79 | 99.65 | 97.68 | 95.79 | 99.65 | 97.68 | 96.65 | 98.20 | 97.42 | 96.68 | 98.27 | 97.47 |
| | | GloVe Embedding Twitter100 | 95.66 | 99.69 | 97.63 | 95.65 | 99.68 | 97.63 | 96.49 | 98.36 | 97.42 | 96.53 | 98.53 | 97.52 |
| | | GloVe Embedding Twitter200 | 95.78 | 99.66 | 97.68 | 95.77 | 99.68 | 97.69 | 96.72 | 98.22 | 97.46 | 96.80 | 98.35 | 97.57 |
| | | GloVe Embedding Twitter25 | 95.31 | 99.91 | 97.56 | 95.30 | 99.90 | 97.55 | 95.81 | 98.79 | 97.28 | 95.88 | 98.80 | 97.32 |
| | | GloVe Embedding Twitter50 | 95.52 | 99.79 | 97.61 | 95.52 | 99.75 | 97.59 | 96.22 | 98.50 | 97.35 | 96.27 | 98.45 | 97.34 |
| | | FastText Embedding | 95.21 | 100.00 | 97.55 | 95.21 | 100.00 | 97.54 | 95.33 | 99.77 | 97.50 | 95.38 | 99.78 | 97.53 |
| | Contextual Word Embeddings | Albert Embedding | 96.03 | 99.71 | 97.84 | 95.96 | 99.79 | 97.84 | 97.33 | 97.31 | 97.31 | 96.94 | 98.32 | 97.62 |
| | | Bert Embedding | 96.27 | 99.66 | 97.94 | 96.28 | 99.70 | 97.96 | 97.45 | 97.66 | 97.55 | 97.27 | 97.85 | 97.56 |
| | | DistilBert Embedding | 96.72 | 99.46 | 98.07 | 96.64 | 99.49 | 98.04 | 97.79 | 97.56 | 97.67 | 97.70 | 97.61 | 97.66 |
| | | Electra-small Embedding | 96.27 | 99.66 | 97.94 | 96.28 | 99.70 | 97.96 | 97.45 | 97.66 | 97.55 | 97.27 | 97.85 | 97.56 |
| | | Elmo Embedding | 96.31 | 99.38 | 97.82 | 96.41 | 99.40 | 97.88 | 97.33 | 97.84 | 97.59 | 97.29 | 97.91 | 97.60 |
| | | ggeluBert Embedding | 96.52 | 99.54 | 98.01 | 96.44 | 99.64 | 98.01 | 97.53 | 97.68 | 97.61 | 97.55 | 97.66 | 97.61 |
| | | mobileBert Embedding | 95.21 | 100.00 | 97.55 | 95.21 | 100.00 | 97.54 | 95.21 | 100.0 | 97.55 | 95.21 | 100.0 | 97.54 |
| | | nnlm Embedding | 95.90 | 99.75 | 97.79 | 95.87 | 99.76 | 97.78 | 97.30 | 97.85 | 97.57 | 97.35 | 97.73 | 97.54 |
| | | RoBerta Embedding | 95.41 | 99.82 | 97.56 | 95.38 | 99.84 | 97.56 | 96.36 | 98.35 | 97.34 | 96.24 | 98.36 | 97.29 |
| | | tnBert Embedding | 96.67 | 99.53 | 98.08 | 96.65 | 99.55 | 98.08 | 97.70 | 97.59 | 97.65 | 97.66 | 97.56 | 97.61 |

Note:

- M = Model, Cat. = Category of Feature, P = Precision, R = Recall and F = F-measure

The current study is not without limitations. This work is limited to binary text classification even though the cyberbullying corpus encompasses input from other roles within cyberbullying episodes, but it does not help us to determine who posted the cyberbullying post. Furthermore, the features of Big Five and Dark Triad were excluded as the former IBM Watson's Personality traits API has been deprecated, and no reliable alternative was suitable for performing the feature extraction. Topic modeling features were not used as well since it is an unsupervised method to identify a text cluster like the corpus, and the outcome is not guaranteed. Due to the expensive computation, the PLMs were fine-tuned within eight to ten epoch numbers under the default setting without other hyperparameter tunings. Though DistilBert was giving excellent performance overall for cyberbullying detection, some cyberbullying posts were misclassified due to their own ambiguous meaning. The other misclassification cases contained sarcastic context, confusing the model in classifying the post to the correct category. There was a fallback for the models developed in this research as they are designed to classify only English textual posts. However, in real cases, a mixture of languages would be used.

Because this study has some limitations, future research can explore different directions. Firstly, this work can be

**TABLE 14.** Performance evaluation metrics of top 8 features combination for cyberbullying detection using logistic regression and linear SVC (non-cyberbullying class).

| | | Original | | | | | | SMOTE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Non-cyberbullying class (negative class) | | | | | | Non-cyberbullying class (negative class) | | | | | |
| | | Cross-validation (10 folds) | | | Hold-out | | | Cross-validation (10 folds) | | | Hold-out | | |
| M | Feature combination | P | R | F | P | R | F | P | R | F | P | R | F |
| Logistic Regression | A + C + D + F | 97.54 | 99.16 | 98.34 | 97.49 | 99.21 | 98.34 | 97.93 | 98.56 | 98.24 | 97.91 | 98.69 | 98.30 |
| | A + C + E + F | 97.52 | 99.16 | 98.33 | 97.54 | 99.18 | 98.35 | 97.89 | 98.55 | 98.22 | 97.94 | 98.50 | 98.22 |
| | A + C + F | 97.56 | 99.15 | 98.35 | 97.56 | 99.22 | 98.39 | 97.91 | 98.56 | 98.23 | 97.91 | 98.56 | 98.23 |
| | A + B + C + F | 97.57 | 99.14 | 98.34 | 97.55 | 99.20 | 98.37 | 97.92 | 98.54 | 98.23 | 97.94 | 98.61 | 98.27 |
| | A + C + D + E + F | 97.53 | 99.16 | 98.34 | 97.58 | 99.19 | 98.38 | 97.93 | 98.51 | 98.22 | 97.88 | 98.74 | 98.31 |
| | A + B + C + D + E + F | 97.56 | 99.17 | 98.36 | 97.57 | 99.16 | 98.36 | 97.92 | 98.51 | 98.21 | 97.94 | 98.70 | 98.32 |
| | A + B + C + D + F | 97.58 | 99.14 | 98.36 | 97.55 | 99.21 | 98.38 | 97.92 | 98.50 | 98.21 | 98.01 | 98.71 | 98.36 |
| | A + B + C + E + F | 97.57 | 99.16 | 98.36 | 97.57 | 99.23 | 98.39 | 97.89 | 98.55 | 98.22 | 97.94 | 98.61 | 98.27 |
| Linear SVC | A + C + D + F | 97.66 | 98.72 | 98.19 | 97.73 | 98.87 | 98.30 | 97.79 | 98.37 | 98.08 | 97.82 | 98.55 | 98.18 |
| | A + C + E + F | 97.68 | 98.71 | 98.19 | 97.74 | 98.83 | 98.28 | 97.79 | 98.32 | 98.06 | 97.92 | 98.47 | 98.19 |
| | A + C + F | 97.69 | 98.73 | 98.20 | 97.71 | 98.82 | 98.26 | 97.82 | 98.34 | 98.08 | 97.88 | 98.47 | 98.18 |
| | A + B + C + F | 97.70 | 98.73 | 98.21 | 97.71 | 98.83 | 98.27 | 97.81 | 98.32 | 98.06 | 97.88 | 98.56 | 98.22 |
| | A + C + D + E + F | 97.67 | 98.72 | 98.19 | 97.72 | 98.87 | 98.29 | 97.78 | 98.36 | 98.07 | 97.87 | 98.55 | 98.21 |
| | A + B + C + D + E + F | 97.67 | 98.73 | 98.19 | 97.72 | 98.91 | 98.31 | 97.78 | 98.35 | 98.06 | 97.86 | 98.57 | 98.21 |
| | A + B + C + D + F | 97.67 | 98.73 | 98.20 | 97.73 | 98.92 | 98.32 | 97.79 | 98.37 | 98.08 | 97.78 | 98.55 | 98.16 |
| | A + B + C + E + F | 97.69 | 98.72 | 98.20 | 97.70 | 98.82 | 98.26 | 97.80 | 98.30 | 98.05 | 97.90 | 98.49 | 98.19 |

Note:
- M = Model, Cat. = Category of Feature, P = Precision, R = Recall and F = F-measure

**TABLE 15.** Performance evaluation metrics for cyberbullying detection by fine-tuning PLMS at different epoch numbers.

| | | Non-cyberbullying class | | | | | |
|---|---|---|---|---|---|---|---|
| | | Cross-validation (5 folds) | | | Hold out | | |
| PLM | Epoch Number | P | R | F | P | R | F |
| DistillBert | 1 | 97.92 | 99.04 | 98.47 | 98.61 | 98.76 | 98.68 |
| | 2 | 97.92 | 99.11 | 98.51 | 98.53 | 98.88 | 98.70 |
| | 3 | 98.01 | 99.12 | 98.56 | 98.41 | 98.94 | 98.67 |
| | **4** | 98.09 | 99.01 | 98.55 | 98.54 | 98.74 | 98.64 |
| | 5 | 97.95 | 99.08 | 98.51 | 98.62 | 98.61 | 98.62 |
| | 6 | 98.04 | 98.93 | 98.48 | 98.36 | 98.95 | 98.66 |
| | 7 | 98.00 | 98.96 | 98.48 | 98.32 | 99.02 | 98.67 |
| | 8 | 98.08 | 98.93 | 98.50 | 98.28 | 99.08 | 98.68 |
| DistilRoBerta | 1 | 97.79 | 98.94 | 98.36 | 98.09 | 99.05 | 98.57 |
| | 2 | 97.99 | 98.75 | 98.37 | 98.04 | 98.91 | 98.47 |
| | **3** | 98.15 | 98.83 | 98.48 | 98.31 | 98.77 | 98.54 |
| | 4 | 98.03 | 98.98 | 98.50 | 97.92 | 99.34 | 98.62 |
| | 5 | 97.83 | 99.14 | 98.48 | 98.08 | 99.20 | 98.64 |
| | 6 | 97.93 | 99.06 | 98.49 | 97.86 | 99.34 | 98.59 |
| | 7 | 98.02 | 98.99 | 98.50 | 98.07 | 99.15 | 98.61 |
| | 8 | 97.95 | 99.06 | 98.50 | 98.00 | 99.20 | 98.60 |
| Electra-small | 1 | 97.40 | 99.14 | 98.26 | 97.48 | 98.33 | 97.90 |
| | 2 | 97.03 | 99.34 | 98.17 | 97.60 | 98.03 | 97.81 |
| | 3 | 97.79 | 99.01 | 98.40 | 97.48 | 98.61 | 98.04 |
| | **4** | 97.86 | 99.01 | 98.43 | 97.66 | 98.23 | 97.94 |
| | 5 | 97.97 | 98.67 | 98.32 | 97.47 | 98.70 | 98.08 |
| | 6 | 97.82 | 99.03 | 98.42 | 97.30 | 99.19 | 98.23 |
| | 7 | 97.85 | 98.96 | 98.41 | 97.58 | 98.79 | 98.18 |
| | 8 | 97.98 | 98.68 | 98.33 | 97.61 | 98.77 | 98.19 |

Note:
- PLM = Pre-trained Language Model, P = Precision, R = Recall and F = F-measure
- Highlighted and the bold number denotes the selected epochs number

extended to classify the participant roles such as harassers, victims, bystanders, and non-bullies. Regarding personality traits, MBTI could serve as the alternative to capture personality traits information from the text; however, a further literature review was necessary to identify the relationship between MBTI and cyberbullying perpetration. Besides that, since there was an inevitable overlap of cyberbullying posts with sarcastic context, cyberbullying detection can be extended to incorporate ironic detection. Since the conversation was handled as an individual post, the research could be extended to account for the relationship between posts to capture the interaction between users within cyberbullying episodes. Lastly, the work on cyberbullying detection should be extended to cover multilingual settings and analyze the context from other metadata such as memes, images, and videos. Features derived from other media, such as image, video, time, and network embeddings, were rarely used among the researchers.

**TABLE 16.** Performance evaluation metrics of top 8 features combination for cyberbullying detection using logistic regression and linear SVC (overall).

| | | | Cross-validation (10 folds) | | | | Hold Out | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| RS | M | Feature combination | A | MP | MR | MF | A | MP | MR | MF |
| Original | Logistic Regression | A + C + D + F | 96.82 | 86.32 | 74.70 | 79.25 | 96.82 | 86.71 | 74.24 | 79.05 |
| | | A + C + E + F | 96.80 | 86.26 | 74.47 | 79.05 | 96.84 | 86.52 | 74.77 | 79.39 |
| | | A + C + F | 96.83 | 86.34 | 74.91 | 79.39 | 96.90 | 87.12 | 74.98 | 79.73 |
| | | A + B + C + F | 96.82 | 86.19 | 74.97 | 79.40 | 96.87 | 86.87 | 74.88 | 79.58 |
| | | A + C + D + E + F | 96.81 | 86.29 | 74.58 | 79.15 | 96.88 | 86.77 | 75.15 | 79.74 |
| | | A + B + C + D + E + F | 96.85 | 86.56 | 74.88 | 79.45 | 96.85 | 86.42 | 75.04 | 79.56 |
| | | A + B + C + D + F | 96.85 | 86.35 | 75.15 | 79.58 | 96.88 | 86.98 | 74.89 | 79.61 |
| | | A + B + C + E + F | 96.84 | 86.44 | 74.99 | 79.50 | 96.92 | 87.27 | 75.08 | 79.85 |
| | Linear SVC | A + C + D + F | 96.53 | 82.62 | 75.86 | 78.78 | 96.74 | 84.25 | 76.66 | 79.93 |
| | | A + C + E + F | 96.54 | 82.68 | 76.06 | 78.93 | 96.71 | 83.95 | 76.74 | 79.86 |
| | | A + C + F | 96.56 | 82.81 | 76.11 | 79.01 | 96.68 | 83.75 | 76.46 | 79.60 |
| | | A + B + C + F | 96.57 | 82.87 | 76.21 | 79.10 | 96.69 | 83.83 | 76.46 | 79.64 |
| | | A + C + D + E + F | 96.53 | 82.63 | 75.92 | 78.81 | 96.72 | 84.17 | 76.48 | 79.78 |
| | | A + B + C + D + E + F | 96.54 | 82.72 | 75.93 | 78.86 | 96.76 | 84.52 | 76.50 | 79.91 |
| | | A + B + C + D + F | 96.55 | 82.76 | 75.93 | 78.88 | 96.78 | 84.74 | 76.60 | 80.06 |
| | | A + B + C + E + F | 96.56 | 82.79 | 76.11 | 79.00 | 96.66 | 83.67 | 76.27 | 79.45 |
| SMOTE | Logistic Regression | A + C + D + F | 96.64 | 82.58 | 78.58 | 80.40 | 96.75 | 83.50 | 78.43 | 80.73 |
| | | A + C + E + F | 96.60 | 82.38 | 78.19 | 80.09 | 96.61 | 82.20 | 78.71 | 80.34 |
| | | A + C + F | 96.62 | 82.47 | 78.38 | 80.25 | 96.62 | 82.47 | 78.37 | 80.26 |
| | | A + B + C + F | 96.61 | 82.39 | 78.45 | 80.25 | 96.69 | 82.95 | 78.67 | 80.64 |
| | | A + C + D + E + F | 96.60 | 82.20 | 78.56 | 80.23 | 96.77 | 83.77 | 78.18 | 80.69 |
| | | A + B + C + D + E + F | 96.59 | 82.16 | 78.46 | 80.17 | 96.78 | 83.69 | 78.72 | 80.98 |
| | | A + B + C + D + F | 96.58 | 82.09 | 78.43 | 80.12 | 96.86 | 84.07 | 79.47 | 81.58 |
| | | A + B + C + E + F | 96.60 | 82.36 | 78.20 | 80.10 | 96.69 | 82.95 | 78.67 | 80.64 |
| | Linear SVC | A + C + D + F | 96.34 | 80.57 | 77.11 | 78.70 | 96.53 | 81.99 | 77.43 | 79.51 |
| | | A + C + E + F | 96.29 | 80.21 | 77.10 | 78.55 | 96.54 | 81.80 | 78.41 | 80.00 |
| | | A + C + F | 96.33 | 80.45 | 77.39 | 78.81 | 96.52 | 81.71 | 78.05 | 79.75 |
| | | A + B + C + F | 96.31 | 80.30 | 77.29 | 78.69 | 96.60 | 82.35 | 78.09 | 80.05 |
| | | A + C + D + E + F | 96.32 | 80.46 | 77.00 | 78.59 | 96.58 | 82.23 | 77.99 | 79.94 |
| | | A + B + C + D + E + F | 96.30 | 80.36 | 77.00 | 78.55 | 96.59 | 82.34 | 77.91 | 79.94 |
| | | A + B + C + D + F | 96.33 | 80.51 | 77.07 | 78.66 | 96.49 | 81.82 | 77.06 | 79.22 |
| | | A + B + C + E + F | 96.28 | 80.15 | 77.20 | 78.57 | 96.55 | 81.93 | 78.24 | 79.96 |

Note:
- RS = Resampling, M = Model, Cat. = Category of Feature, A = Accuracy, MP = Macro Precision, MR = Macro Recall and MF = Macro F-measure

**TABLE 17.** Performance evaluation metrics for cyberbullying detection by fine-tuning PLMS at different epoch numbers (overall).

| | | Cross-validation (5 folds) | | | | Hold out | | | |
|---|---|---|---|---|---|---|---|---|---|
| PLM | Epoch Number | A | MP | MR | MF | A | MP | MR | MF |
| DistilBert | 1 | 97.07 | 86.55 | 78.55 | 81.72 | 97.49 | 86.56 | 85.53 | 86.04 |
| | 2 | 97.15 | 87.21 | 78.57 | 82.01 | 97.52 | 87.26 | 84.75 | 85.96 |
| | 3 | 97.25 | 87.54 | 79.58 | 82.88 | 97.47 | 87.43 | 83.58 | 85.39 |
| | 4 | 97.22 | 86.84 | 80.36 | 83.22 | 97.41 | 86.22 | 84.87 | 85.53 |
| | 5 | 97.15 | 87.10 | 78.98 | 82.43 | 97.36 | 85.52 | 85.64 | 85.58 |
| | 6 | 97.09 | 85.94 | 79.80 | 82.52 | 97.43 | 87.37 | 83.12 | 85.10 |
| | 7 | 97.08 | 86.11 | 79.37 | 82.11 | 97.45 | 87.80 | 82.69 | 85.03 |
| | 8 | 97.13 | 86.10 | 80.17 | 82.76 | 97.48 | 88.28 | 82.35 | 85.03 |
| DistilRoBerta | 1 | 96.86 | 85.46 | 77.22 | 80.63 | 97.27 | 87.38 | 80.38 | 83.48 |
| | 2 | 96.88 | 84.66 | 79.26 | 81.38 | 97.08 | 85.85 | 79.84 | 82.54 |
| | 3 | 97.10 | 85.52 | 80.84 | 82.86 | 97.22 | 85.74 | 82.57 | 84.07 |
| | 4 | 97.13 | 86.34 | 79.68 | 82.55 | 97.35 | 89.69 | 78.66 | 83.19 |
| | 5 | 97.08 | 87.29 | 77.66 | 81.49 | 97.39 | 88.73 | 80.36 | 83.98 |
| | 6 | 97.11 | 86.83 | 78.75 | 82.08 | 97.30 | 89.51 | 78.11 | 82.74 |
| | 7 | 97.13 | 86.40 | 79.60 | 82.58 | 97.34 | 88.23 | 80.24 | 83.72 |
| | 8 | 97.13 | 86.79 | 78.90 | 82.27 | 97.31 | 88.46 | 79.52 | 83.34 |
| Electra-small | 1 | 96.66 | 86.54 | 73.21 | 77.17 | 95.98 | 78.63 | 73.88 | 76.01 |
| | 2 | 96.47 | 87.52 | 69.39 | 73.42 | 95.82 | 77.31 | 75.04 | 76.12 |
| | 3 | 96.93 | 85.92 | 77.31 | 80.90 | 96.25 | 80.79 | 74.02 | 76.93 |
| | 4 | 96.99 | 86.01 | 77.96 | 81.38 | 96.07 | 78.93 | 75.70 | 77.20 |
| | 5 | 96.79 | 83.69 | 79.05 | 81.13 | 96.32 | 81.49 | 73.89 | 77.10 |
| | 6 | 96.97 | 86.05 | 77.56 | 81.14 | 96.61 | 85.51 | 72.27 | 77.20 |
| | 7 | 96.95 | 85.63 | 77.92 | 81.23 | 96.52 | 82.86 | 75.05 | 78.36 |
| | 8 | 96.81 | 83.92 | 79.15 | 81.29 | 96.53 | 82.83 | 75.32 | 78.52 |

Note:
- PLM = Pre-trained Language Model, A = Accuracy, MP = Macro Precision, MR = Macro Recall, MF = Macro F-measure,
- Bold values are the highest value across epochs and PLMs

In wrapping up the research, the cyberbullying detection model developed in this research gave valuable contributions to future research that can have close supervision on the cyberbullying interaction from the textual posts. The proposed features for conventional machine learning effectively trained the model for task-specific text classification

**TABLE 18.** Performance evaluation metrics of individual feature grouping for cyberbullying detection using logistic regression and linear SVC (overall).

| | | | Original | | | | | | | | SMOTE | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Cross-validation (10 folds) | | | | Hold-out | | | | Cross-validation (10 folds) | | | | Hold-out | | | |
| **M** | **Cat.** | **Feature** | **A** | **MP** | **MR** | **MF** | **A** | **MP** | **MR** | **MF** | **A** | **MP** | **MR** | **MF** | **A** | **MP** | **MR** | **MF** |
| Logistic Regression | Standalone Grouping | **Textual** | 96.46 | 87.23 | 68.18 | **74.06** | 96.48 | 87.27 | 68.41 | **74.31** | 96.55 | 83.34 | 74.65 | **78.25** | 96.69 | 84.44 | 75.31 | **79.09** |
| | | Sentiment and Emotion | 95.23 | 74.34 | 51.58 | 51.89 | 95.15 | 68.50 | 51.30 | 51.37 | 94.07 | 64.00 | 59.34 | 61.09 | 93.92 | 63.43 | 59.48 | 61.03 |
| | | Psycholinguistic | 95.38 | 76.76 | 56.31 | 59.47 | 95.41 | 77.65 | 56.46 | 59.74 | 94.99 | 71.86 | 66.67 | 68.85 | 94.85 | 70.85 | 65.70 | 67.85 |
| | | Term Lists (Ratio) | 95.00 | 60.21 | 50.97 | 50.81 | 95.04 | 60.46 | 50.79 | 50.46 | 94.59 | 62.79 | 54.30 | 55.95 | 94.75 | 65.00 | 54.70 | 56.63 |
| | | **Toxicity** | 95.90 | 79.74 | 67.14 | **71.51** | 95.85 | 79.28 | 67.02 | **71.31** | 94.99 | 72.98 | 77.24 | **74.88** | 95.10 | 73.47 | 77.57 | **75.32** |
| | Static Word Embeddings | Word2Vec Embedding | 95.51 | 78.21 | 58.83 | 62.86 | 95.50 | 77.91 | 59.07 | 63.12 | 94.82 | 71.02 | 67.77 | 69.22 | 94.98 | 72.11 | 68.68 | 70.23 |
| | | GloVe Embedding Wikipedia | 95.21 | 72.43 | 53.60 | 55.33 | 95.18 | 71.81 | 53.34 | 54.92 | 94.46 | 65.98 | 59.12 | 61.41 | 94.31 | 64.72 | 58.53 | 60.62 |
| | | GloVe Embedding Common42B | 95.50 | 78.20 | 58.56 | 62.52 | 95.34 | 75.27 | 57.93 | 61.49 | 94.86 | 71.21 | 67.66 | 69.23 | 95.00 | 72.10 | 67.72 | 69.63 |
| | | GloVe Embedding Common840B | 95.52 | 78.27 | 58.89 | 62.93 | 95.47 | 77.45 | 58.61 | 62.52 | 94.80 | 70.90 | 67.68 | 69.13 | 94.98 | 72.19 | 69.39 | 70.68 |
| | | GloVe Embedding Twitter100 | 95.44 | 77.17 | 57.80 | 61.49 | 95.47 | 77.98 | 58.08 | 61.91 | 94.71 | 70.07 | 66.09 | 67.81 | 94.77 | 70.50 | 66.36 | 68.16 |
| | | GloVe Embedding Twitter200 | 95.47 | 77.35 | 58.90 | 62.86 | 95.53 | 78.51 | 58.99 | 63.09 | 94.84 | 71.20 | 68.14 | 69.52 | 95.10 | 72.87 | 69.45 | 71.00 |
| | | GloVe Embedding Twitter25 | 95.20 | 72.47 | 53.41 | 55.04 | 95.14 | 70.43 | 53.40 | 55.00 | 94.31 | 64.76 | 58.65 | 60.72 | 94.30 | 64.90 | 58.88 | 60.96 |
| | | GloVe Embedding Twitter50 | 95.35 | 76.27 | 55.90 | 58.88 | 95.31 | 75.18 | 55.79 | 58.68 | 94.60 | 68.61 | 63.24 | 65.37 | 94.53 | 68.25 | 63.15 | 65.20 |
| | | FastText Embedding | 95.21 | 72.84 | 50.90 | 50.60 | 95.24 | 76.24 | 51.43 | 51.61 | 94.83 | 61.60 | 52.15 | 52.87 | 95.00 | 66.12 | 52.72 | 53.83 |
| | Contextual Word Embeddings | Albert Embedding | 95.78 | 82.82 | 60.41 | 65.19 | 95.74 | 84.05 | 58.93 | 63.44 | 94.75 | 70.73 | 68.31 | 69.40 | 94.67 | 70.45 | 68.78 | 69.57 |
| | | Bert Embedding | 96.04 | 83.12 | 65.15 | 70.39 | 96.08 | 85.03 | 63.96 | 69.50 | 95.13 | 73.33 | 73.64 | 73.47 | 95.31 | 74.24 | 72.82 | 73.51 |
| | | **DistilBert Embedding** | 96.27 | 83.98 | 68.29 | **73.47** | 96.32 | 84.67 | 68.41 | **73.77** | 95.45 | 75.04 | 76.82 | **75.87** | 95.31 | 74.33 | 76.09 | **75.17** |
| | | Electra-small Embedding | 96.04 | 83.12 | 65.15 | 70.39 | 96.08 | 85.03 | 63.96 | 69.50 | 95.13 | 73.33 | 73.64 | 73.47 | 95.31 | 74.24 | 72.82 | 73.51 |
| | | Elmo Embedding | 95.76 | 79.78 | 63.43 | 68.12 | 95.80 | 79.97 | 64.08 | 68.81 | 95.46 | 75.13 | 73.11 | 74.05 | 95.47 | 75.19 | 72.73 | 73.89 |
| | | ggeluBert Embedding | 96.12 | 83.50 | 66.16 | 71.42 | 96.01 | 82.68 | 64.98 | 70.17 | 95.35 | 74.52 | 74.66 | 74.58 | 95.39 | 74.76 | 74.46 | 74.61 |
| | | mobileBert Embedding | 95.21 | 47.61 | 50.00 | 48.77 | 95.21 | 47.60 | 50.00 | 48.77 | 95.14 | 56.32 | 50.12 | 49.08 | 95.20 | 69.84 | 50.35 | 49.50 |
| | | nnlm Embedding | 95.81 | 82.90 | 60.92 | 65.81 | 95.72 | 82.02 | 59.98 | 64.61 | 95.31 | 74.24 | 72.57 | 73.36 | 95.19 | 73.62 | 73.20 | 73.41 |
| | | RoBerta Embedding | 95.33 | 76.33 | 54.73 | 57.18 | 95.26 | 74.65 | 53.82 | 55.74 | 94.78 | 69.93 | 63.49 | 65.96 | 94.80 | 70.01 | 63.46 | 65.98 |
| | | tnBert Embedding | 95.23 | 74.34 | 51.58 | 51.89 | 96.12 | 82.46 | 67.42 | 72.41 | 95.39 | 74.77 | 76.17 | 75.43 | 95.40 | 74.83 | 75.96 | 75.38 |
| Linear SVC | Standalone Grouping | **Textual** | 96.40 | 82.25 | 73.93 | **77.38** | 96.54 | 82.95 | 75.41 | **78.63** | 96.00 | 78.18 | 77.22 | **77.67** | 96.28 | 79.79 | 78.63 | **79.19** |
| | | Sentiment and Emotion | 95.21 | 70.94 | 50.09 | 48.96 | 95.22 | 80.95 | 50.18 | 49.14 | 94.73 | 67.61 | 58.46 | 61.12 | 94.60 | 66.50 | 58.24 | 60.73 |
| | | Psycholinguistic | 95.33 | 76.75 | 54.41 | 56.69 | 95.45 | 80.43 | 55.25 | 58.08 | 95.12 | 72.49 | 64.34 | 67.36 | 95.11 | 72.50 | 64.69 | 67.64 |
| | | Term Lists (Ratio) | 94.99 | 59.82 | 50.91 | 50.69 | 95.03 | 59.48 | 50.70 | 50.29 | 94.63 | 62.32 | 53.73 | 55.19 | 94.79 | 64.80 | 54.19 | 55.96 |
| | | **Toxicity** | 95.89 | 79.89 | 66.68 | **71.17** | 95.91 | 80.09 | 66.87 | **71.39** | 95.12 | 73.44 | 76.40 | **74.81** | 95.22 | 73.94 | 76.84 | **75.29** |
| | Static Word Embeddings | Word2Vec Embedding | 95.50 | 80.48 | 56.26 | 59.57 | 95.49 | 80.44 | 56.24 | 59.55 | 95.06 | 72.22 | 65.42 | 68.09 | 95.16 | 73.01 | 65.86 | 68.67 |
| | | GloVe Embedding Wikipedia | 95.25 | 75.45 | 51.75 | 52.20 | 95.33 | 81.05 | 52.54 | 53.63 | 94.78 | 67.24 | 56.98 | 59.54 | 94.73 | 66.44 | 56.46 | 58.87 |
| | | GloVe Embedding Common42B | 95.50 | 80.74 | 56.12 | 59.39 | 95.50 | 80.90 | 56.16 | 59.45 | 95.07 | 72.28 | 65.04 | 67.83 | 95.17 | 73.02 | 65.07 | 68.09 |
| | | GloVe Embedding Common840B | 95.50 | 80.37 | 56.27 | 59.58 | 95.50 | 80.61 | 56.33 | 59.69 | 95.05 | 72.07 | 65.27 | 67.94 | 95.14 | 72.79 | 65.58 | 68.40 |
| | | GloVe Embedding Twitter100 | 95.39 | 78.60 | 54.84 | 57.40 | 95.39 | 78.51 | 54.86 | 57.44 | 95.04 | 71.75 | 63.61 | 66.58 | 95.23 | 73.50 | 64.14 | 67.47 |
| | | GloVe Embedding Twitter200 | 95.49 | 80.38 | 56.13 | 59.38 | 95.51 | 81.22 | 56.16 | 59.47 | 95.13 | 72.74 | 65.95 | 68.65 | 95.33 | 74.35 | 66.83 | 69.80 |
| | | GloVe Embedding Twitter25 | 95.23 | 74.90 | 51.11 | 50.99 | 95.22 | 73.74 | 51.06 | 50.91 | 94.73 | 66.31 | 56.44 | 58.81 | 94.82 | 67.75 | 57.21 | 59.86 |
| | | GloVe Embedding Twitter50 | 95.35 | 79.10 | 53.38 | 55.07 | 95.31 | 76.99 | 53.41 | 55.10 | 94.89 | 69.93 | 60.81 | 63.77 | 94.89 | 70.09 | 61.31 | 64.26 |
| | | FastText Embedding | 95.21 | 47.61 | 50.00 | 48.77 | 95.21 | 47.60 | 50.00 | 48.77 | 95.13 | 67.10 | 51.30 | 51.38 | 95.18 | 71.02 | 51.84 | 52.37 |
| | Contextual Word Embeddings | Albert Embedding | 95.80 | 86.18 | 58.87 | 63.39 | 95.80 | 88.07 | 58.17 | 62.63 | 94.89 | 73.02 | 72.03 | 72.08 | 95.44 | 75.15 | 68.30 | 71.11 |
| | | Bert Embedding | 96.00 | 86.97 | 61.47 | 66.85 | 96.04 | 88.01 | 61.56 | 67.08 | 95.34 | 74.42 | 73.46 | 73.93 | 95.33 | 74.37 | 71.60 | 72.89 |
| | | **DistilBert Embedding** | 96.27 | 86.06 | 66.18 | **71.92** | 96.21 | 85.99 | 65.36 | 71.10 | 95.58 | 75.73 | 76.81 | **76.24** | 95.54 | 75.54 | 75.94 | **75.74** |
| | | Electra-small Embedding | 96.00 | 86.97 | 61.47 | 66.85 | 96.04 | 88.01 | 61.56 | 67.08 | 95.34 | 74.42 | 73.46 | 73.93 | 95.33 | 74.37 | 71.60 | 72.89 |
| | | Elmo Embedding | 95.78 | 81.22 | 61.86 | 66.68 | 95.90 | 82.67 | 62.90 | 68.03 | 95.39 | 74.74 | 72.25 | 73.40 | 95.42 | 74.92 | 71.91 | 73.31 |
| | | ggeluBert Embedding | 96.14 | 86.26 | 64.05 | 69.74 | 96.15 | 87.62 | 63.29 | 69.09 | 95.44 | 75.01 | 74.29 | 74.62 | 95.44 | 75.01 | 74.48 | 74.74 |
| | | mobileBert Embedding | 95.21 | 47.61 | 50.00 | 48.77 | 95.21 | 47.60 | 50.00 | 48.77 | 95.21 | 47.61 | 50.00 | 48.77 | 95.21 | 47.60 | 50.00 | 48.77 |
| | | nnlm Embedding | 95.70 | 85.67 | 57.46 | 61.51 | 95.68 | 85.56 | 57.22 | 61.17 | 95.36 | 74.52 | 71.89 | 73.11 | 95.31 | 74.23 | 72.47 | 73.31 |
| | | RoBerta Embedding | 95.26 | 75.74 | 52.13 | 52.90 | 95.24 | 75.32 | 51.87 | 52.43 | 94.88 | 70.28 | 62.19 | 65.04 | 94.78 | 69.15 | 60.98 | 63.76 |
| | | tnBert Embedding | 96.28 | 86.96 | 65.64 | 71.52 | 96.29 | 87.26 | 65.48 | **71.42** | 95.52 | 75.46 | 75.99 | 75.70 | 95.45 | 75.06 | 75.54 | 75.30 |

Note:
- M = Model, Cat. = Category of Feature, A = Accuracy, MP = Macro Precision, MR = Macro Recall and MF = Macro F-measure
- Bold values are the highest value across resampling methods, models, and combination of features

for corpus relating to the study of cyberbullying episodes. The new benchmark achieved by Logistic Regression and the optimized version of PLMs signifies a step forward in cyberbullying detection.

## APPENDIX

See Tables 13–18.

## REFERENCES

[1] B. Cagirkan and G. Bilek, “Cyberbullying among Turkish high school students,” *Scandin. J. Psychol.*, vol. 62, no. 4, pp. 608–616, Aug. 2021, doi: 10.1111/sjop.12720.

[2] P. T. L. Chi, V. T. H. Lan, N. H. Ngan, and N. T. Linh, “Online time, experience of cyber bullying and practices to cope with it among high school students in Hanoi,” *Health Psychol. Open*, vol. 7, no. 1, Jan. 2020, Art. no. 205510292093574, doi: 10.1177/2055102920935747.

[3] A. López-Martínez, J. A. García-Díaz, R. Valencia-García, and A. Ruiz-Martínez, “CyberDect. A novel approach for cyberbullying detection on Twitter,” in *Proc. Int. Conf. Technol. Innov.*, Guayaquil, Ecuador: Springer, 2019, pp. 109–121, doi: 10.1007/978-3-030-34989-9_9.

[4] R. M. Kowalski and S. P. Limber, “Psychological, physical, and academic correlates of cyberbullying and traditional bullying,” *J. Adolescent Health*, vol. 53, no. 1, pp. S13–S20, Jul. 2013, doi: 10.1016/j.jadohealth.2012.09.018.

[5] Y.-C. Huang, “Comparison and contrast of piaget and Vygotsky’s theories,” in *Proc. Adv. Social Sci., Educ. Humanities Res.*, 2021, pp. 28–32, doi: 10.2991/assehr.k.210519.007.

[6] A. Anwar, D. M. H. Kee, and A. Ahmed, “Workplace cyberbullying and interpersonal deviance: Understanding the mediating effect of silence and emotional exhaustion,” *Cyberpsychol., Behav., Social Netw.*, vol. 23, no. 5, pp. 290–296, May 2020, doi: 10.1089/cyber.2019.0407.

[7] D. M. H. Kee, M. A. L. Al-Anesi, and S. A. L. Al-Anesi, “Cyberbullying on social media under the influence of COVID-19,” *Global Bus. Organizational Excellence*, vol. 41, no. 6, pp. 11–22, Sep. 2022, doi: 10.1002/joe.22175.

[8] I. Kwan, K. Dickson, M. Richardson, W. MacDowall, H. Burchett, C. Stansfield, G. Brunton, K. Sutcliffe, and J. Thomas, “Cyberbullying and children and young people’s mental health: A systematic map of systematic reviews,” *Cyberpsychol., Behav., Social Netw.*, vol. 23, no. 2, pp. 72–82, Feb. 2020, doi: 10.1089/cyber.2019.0370.

[9] R. Garett, L. R. Lord, and S. D. Young, “Associations between social media and cyberbullying: A review of the literature,” *mHealth*, vol. 2, p. 46, Dec. 2016, doi: 10.21037/mhealth.2016.12.01.

[10] M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, and K. Araki, “Automatic extraction of harmful sentence patterns with application in cyberbullying detection,” in *Proc. Lang. Technol. Conf.* Poznań, Poland: Springer, 2015, pp. 349–362, doi: 10.1007/978-3-319-93782-3_25.

[11] M. Ptaszynski, P. Lempa, F. Masui, Y. Kimura, R. Rzepka, K. Araki, M. Wroczynski, and G. Leliwa, “"Brute-force sentence pattern extortion from harmful messages for cyberbullying detection,"’ *J. Assoc. Inf. Syst.*, vol. 20, no. 8, pp. 1075–1127, 2019.

[12] M. O. Raza, M. Memon, S. Bhatti, and R. Bux, “Detecting cyberbullying in social commentary using supervised machine learning,” in *Proc. Future Inf. Commun. Conf.* Cham, Switzerland: Springer, 2020, pp. 621–630.

[13] D. Nguyen, M. Liakata, S. Dedeo, J. Eisenstein, D. Mimno, R. Tromble, and J. Winters, “How we do things with words: Analyzing text as social and cultural data,” *Frontiers Artif. Intell.*, vol. 3, p. 62, Aug. 2020, doi: 10.3389/frai.2020.00062.

[14] J. Cai, J. Li, W. Li, and J. Wang, "Deeplearning model used in text classification," in *Proc. 15th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2018, pp. 123–126, doi: 10.1109/ICCWAMTIP.2018.8632592.

[15] N. Tiku and C. Newton. *Twitter CEO: We Suck at Dealing With Abuse*. Verge. Accessed: Aug. 17, 2022. [Online]. Available: https://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the

[16] D. Noever, "Machine learning suites for online toxicity detection," 2018, *arXiv:1810.01869*.

[17] D. G. Krutka, S. Manca, S. M. Galvin, C. Greenhow, M. J. Koehler, and E. Askari, "Teaching 'against' social media: Confronting problems of profit in the curriculum," *Teachers College Rec., Voice Scholarship Educ.*, vol. 121, no. 14, pp. 1–42, Dec. 2019, doi: 10.1177/016146811912101410.

[18] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. M. V. Simão, and I. Trancoso, "Automatic cyberbullying detection: A systematic review," *Comput. Hum. Behav.*, vol. 93, pp. 333–345, Apr. 2019, doi: 10.1016/j.chb.2018.12.021.

[19] S. Bharti, A. K. Yadav, M. Kumar, and D. Yadav, "Cyberbullying detection from tweets using deep learning," *Kybernetes*, vol. 51, no. 9, pp. 2695–2711, Sep. 2022.

[20] A. Bozyiğit, S. Utku, and E. Nasibov, "Cyberbullying detection: Utilizing social media features," *Expert Syst. Appl.*, vol. 179, Oct. 2021, Art. no. 115001, doi: 10.1016/j.eswa.2021.115001.

[21] H.-S. Lee, H.-R. Lee, J.-U. Park, and Y.-S. Han, "An abusive text detection system based on enhanced abusive and non-abusive word lists," *Decis. Support Syst.*, vol. 113, pp. 22–31, Sep. 2018, doi: 10.1016/j.dss.2018.06.009.

[22] Y. Fang, S. Yang, B. Zhao, and C. Huang, "Cyberbullying detection in social networks using bi-GRU with self-attention mechanism," *Information*, vol. 12, no. 4, p. 171, Apr. 2021, doi: 10.3390/info12040171.

[23] G. Jacobs, C. Van Hee, and V. Hoste, "Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?" *Natural Lang. Eng.*, vol. 28, no. 2, pp. 141–166, Mar. 2022, doi: 10.1017/S135132492000056X.

[24] M. Gada, K. Damania, and S. Sankhe, "Cyberbullying detection using LSTM-CNN architecture and its applications," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Jan. 2021, pp. 1–6, doi: 10.1109/ICCCI50826.2021.9402412.

[25] H. H.-P. Vo, H. Trung Tran, and S. T. Luu, "Automatically detecting cyberbullying comments on online game forums," in *Proc. RIVF Int. Conf. Comput. Commun. Technol. (RIVF)*, Aug. 2021, pp. 1–5, doi: 10.1109/RIVF51545.2021.9642116.

[26] F. Elsafoury, S. Katsigiannis, Z. Pervez, and N. Ramzan, "When the timeline meets the pipeline: A survey on automated cyberbullying detection," *IEEE Access*, vol. 9, pp. 103541–103563, 2021, doi: 10.1109/ACCESS.2021.3098979.

[27] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2018, pp. 328–339.

[28] R. Silva Barbon and A. T. Akabane, "Towards transfer learning techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for automatic text classification from different languages: A case study," *Sensors*, vol. 22, no. 21, p. 8184, Oct. 2022, doi: 10.3390/s22218184.

[29] J. Eronen, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, "Exploring the potential of feature density in estimating machine learning classifier performance with application to cyberbullying detection," 2022, *arXiv:2206.01949*.

[30] J. Bhagya and P. S. Deepthi, *Cyberbullying Detection on Social Media Using SVM* (Inventive Systems and Control). Singapore: Springer, 2021, pp. 17–27, doi: 10.1007/978-981-16-1395-1_2.

[31] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," *Proc. Comput. Sci.*, vol. 181, pp. 605–611, Jan. 2021, doi: 10.1016/j.procs.2021.01.207.

[32] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proc. 17th Int. Conf. Distrib. Comput. Netw.*, Jan. 2016, pp. 1–6, doi: 10.1145/2833312.2849567.

[33] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," *Commun. Inf. Sci. Manage. Eng.*, vol. 3, no. 5, p. 238, 2013.

[34] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, pp. 1–30, Sep. 2012.

[35] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proc. Int. AAAI Conf. Web Social Media*, Barcelona, Spain, 2011, pp. 11–17. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14209

[36] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," in *Proc. Content Anal. Web*, vol. 2. Madrid, Spain, 2009, pp. 1–7. [Online]. Available: https://www.academia.edu/download/47631616/Yin_etal_CAW2009.pdf

[37] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste, "Automatic detection of cyberbullying in social media text," *PLoS ONE*, vol. 13, no. 10, Oct. 2018, Art. no. e0203794, doi: 10.1371/journal.pone.0203794.

[38] C. Van Hee, "Detection and fine-grained classification of cyberbullying events," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP)*, 2015, pp. 672–680.

[39] A. Muneer and S. M. Fati, "A comparative analysis of machine learning techniques for cyberbullying detection on Twitter," *Future Internet*, vol. 12, no. 11, p. 187, Oct. 2020, doi: 10.3390/fi12110187.

[40] L. J. Thun, P. L. Teh, and C.-B. Cheng, "CyberAid: Are your children safe from cyberbullying?" *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4099–4108, Jul. 2022, doi: 10.1016/j.jksuci.2021.03.001.

[41] D. Chatzakou, I. Leontiadis, J. Blackburn, E. D. Cristofaro, G. Stringhini, A. Vakali, and N. Kourtellis, "Detecting cyberbullying and cyberaggression in social media," *ACM Trans. Web*, vol. 13, no. 3, pp. 1–51, Aug. 2019.

[42] V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, "Cyberbullying detection on Twitter using big five and dark triad features," *Personality Individual Differences*, vol. 141, pp. 252–257, Apr. 2019, doi: 10.1016/j.paid.2019.01.024.

[43] R. Zhao and K. Mao, "Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 328–339, Jul. 2017, doi: 10.1109/TAFFC.2016.2531682.

[44] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J. P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth, and E. Dillon, "Cyberbullying detection with a pronunciation based convolutional neural network," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 740–745, doi: 10.1109/ICMLA.2016.0132.

[45] A. Kumar and N. Sachdeva, "A bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media," *World Wide Web*, vol. 25, no. 4, pp. 1537–1550, Jul. 2022, doi: 10.1007/s11280-021-00920-4.

[46] K. Shriniket, P. Vidyarthi, S. Udyavara, R. Manohar, and N. Shruthi, "A time optimised model for cyberbullying detection," *Int. Res. J. Modernization Eng., Technol. Sci.*, vol. 4, no. 7, pp. 808–815, 2022.

[47] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Proc. Eur. Conf. Inf. Retr. (ECIR)*, Grenoble, France: Springer, 2018, pp. 141–153, doi: 10.1007/978-3-319-76941-7_11.

[48] M. Dadvar and K. Eckert, "Cyberbullying detection in social networks using deep learning based models; a reproducibility study," 2018, *arXiv:1812.08046*.

[49] H. Rosa, D. Matos, R. Ribeiro, L. Coheur, and J. P. Carvalho, "A 'deeper' look at detecting cyberbullying in social networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8, doi: 10.1109/IJCNN.2018.8489211.

[50] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, "Hierarchical attention networks for cyberbullying detection on the Instagram social network," in *Proc. SIAM Int. Conf. Data Mining*, Calgary, ALB, Canada: SIAM, 2019, pp. 235–243.

[51] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, "Social media cyberbullying detection using machine learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, pp. 703–707, 2019, doi: 10.14569/IJACSA.2019.0100587.

[52] N. Yuvaraj, V. Chang, B. Gobinathan, A. Pinagapani, S. Kannan, G. Dhiman, and A. R. Rajan, "Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification," *Comput. Electr. Eng.*, vol. 92, Jun. 2021, Art. no. 107186, doi: 10.1016/j.compeleceng.2021.107186.

[53] S. Paul and S. Saha, "CyberBERT: BERT for cyberbullying identification," *Multimedia Syst.*, vol. 28, pp. 1897–1904, Nov. 2020, doi: 10.1007/s00530-020-00710-4.

[54] P. Yi and A. Zubiaga, "Cyberbullying detection across social media platforms via platform-aware adversarial encoding," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 16. Atlanta, GA, USA, 2022, pp. 1430–1434. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/19401

[55] K. Verma, T. Milosevic, K. Cortis, and B. Davis, "Benchmarking language models for cyberbullying identification and classification from social-media texts," in *Proc. 1st Workshop Lang. Technol. Resour. Fair, Inclusive, Safe Soc. Within 13th Lang. Resour. Eval. Conf.*, Marseille, France, 2022, pp. 26–31. [Online]. Available: https://aclanthology.org/2022.lateraisse-1.4

[56] B. Bhatia, A. Verma, and R. Katarya, "Analysing cyberbullying using natural language processing by understanding jargon in social media," in *Sustainable Advanced Computing*. Cham, Switzerland: Springer, 2022, pp. 397–406.

[57] J. Qiu, M. Moh, and T.-S. Moh, "Multi-modal detection of cyberbullying on Twitter," in *Proc. ACM Southeast Conf.*, Apr. 2022, pp. 9–16, doi: 10.1145/3476883.3520222.

[58] A. Kumar and N. Sachdeva, "Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network," *Multimedia Syst.*, vol. 28, pp. 2043–2052, Feb. 2021, doi: 10.1007/s00530-020-00747-5.

[59] K. Kumari, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, "Towards cyberbullying-free social media in smart cities: A unified multi-modal approach," *Soft Comput.*, vol. 24, no. 15, pp. 11059–11070, Aug. 2020, doi: 10.1007/s00500-019-04550-x.

[60] M. F. López-Vizcaíno, F. J. Nóvoa, V. Carneiro, and F. Cacheda, "Early detection of cyberbullying on social media networks," *Future Gener. Comput. Syst.*, vol. 118, pp. 219–229, May 2021, doi: 10.1016/j.future.2021.01.006.

[61] K. Wang, Q. Xiong, C. Wu, M. Gao, and Y. Yu, "Multi-modal cyberbullying detection on social networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9206663.

[62] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Prediction of cyberbullying incidents in a media-based social network," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2016, pp. 186–192, doi: 10.1109/ASONAM.2016.7752233.

[63] D. Soni and V. K. Singh, "See no evil, hear no evil: Audio-visual-textual cyberbullying detection," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, 2018, pp. 1–26, doi: 10.1145/3274433.

[64] P. K. Roy and F. U. Mali, "Cyberbullying detection using deep transfer learning," *Complex Intell. Syst.*, vol. 8, pp. 5449–5467, May 2022, doi: 10.1007/s40747-022-00772-z.

[65] N. Vishwamitra, H. Hu, F. Luo, and L. Cheng, "Towards understanding and detecting cyberbullying in real-world images," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, Jan. 2021, pp. 1–18, doi: 10.14722/ndss.2021.24260.

[66] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Detecting aggressors and bullies on Twitter," in *Proc. 26th Int. Conf. World Wide Web Companion WWW Companion*, 2017, pp. 767–768, doi: 10.1145/3041021.3054211.

[67] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on Twitter," in *Proc. ACM Web Sci. Conf.*, Jun. 2017, pp. 13–22, doi: 10.1145/3091478.3091487.

[68] G. Rathnayake, T. Atapattu, M. Herath, G. Zhang, and K. Falkner, "Enhancing the identification of cyberbullying through participant roles," in *Proc. 4th Workshop Online Abuse Harms*, 2020, pp. 89–94.

[69] R. Sugandhi, A. Pande, A. Agrawal, and H. Bhagat, "Automatic monitoring and prevention of cyberbullying," *Int. J. Comput. Appl.*, vol. 144, no. 8, pp. 17–19, Jun. 2016. [Online]. Available: https://pdfs.semanticscholar.org/eb09/

[70] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *Proc. Eur. Conf. Inf. Retr.* Moscow, Russia: Springer, 2013, pp. 693–696, doi: 10.1007/978-3-642-36973-5_62.

[71] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, "Careful what you share in six seconds: Detecting cyberbullying instances in vine," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2015, pp. 617–622, doi: 10.1145/2808797.2809381.

[72] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Comput. Hum. Behav.*, vol. 63, pp. 433–443, Oct. 2016, doi: 10.1016/j.chb.2016.05.051.

[73] W. N. Hamiza Wan Ali, M. Mohd, and F. Fauzi, "Cyberbullying detection: An overview," in *Proc. Cyber Resilience Conf. (CRC)*, Nov. 2018, pp. 1–3.

[74] M. Fortunatus, P. Anthony, and S. Charters, "Combining textual features to detect cyberbullying in social media posts," *Proc1 Comput. Sci.*, vol. 176, pp. 612–621, Jan. 2020, doi: 10.1016/j.procs.2020.08.063.

[75] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in *Proc. 3rd Int. Workshop Socially-Aware Multimedia*, Nov. 2014, pp. 3–6.

[76] M. Dadvar, F. D. Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proc. 12th Dutch-Belgian Inf. Retr. Workshop (DIR)*, Ghent, Belgium, 2012.

[77] B. S. Nandhini and J. I. Sheeba, "Cyberbullying detection and classification using information retrieval algorithm," in *Proc. Int. Conf. Adv. Res. Comput. Sci. Eng. Technol. (ICARCSET)*, Mar. 2015, pp. 1–5.

[78] W. N. H. W. Ali, M. Mohd, F. Fauzi, K. Shirai, and M. J. Mahamad Noor, "Implementation of Hyperparameter optimisation and over-sampling in detecting cyberbullying using machine learning approach," *Malaysian J. Comput. Sci.*, pp. 78–100, Dec. 2021. [Online]. Available: http://mojes.um.edu.my/index.php/MJCS/article/view/34401 and https://mojes.um.edu.my/index.php/MJCS/issue/view/1982

[79] J. Wang, K. Fu, and C.-T. Lu, "SOSNet: A graph convolutional network approach to fine-grained cyberbullying detection," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 1699–1708, doi: 10.1109/BigData50022.2020.9378065.

[80] A. Kumar, S. Nayak, and N. Chandra, "Empirical analysis of supervised machine learning techniques for cyberbullying detection," in *Proc. Int. Conf. Innov. Comput. Commun.* Ostrava, Czechia: Springer, 2019, pp. 223–230, doi: 10.1007/978-981-13-2354-6_24.

[81] E. Raisi and B. Huang, "Weakly supervised cyberbullying detection using co-trained ensembles of embedding models," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 479–486, doi: 10.1109/ASONAM.2018.8508240.

[82] J. O. Atoum, "Cyberbullying detection through sentiment analysis," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2020, pp. 292–297, doi: 10.1109/CSCI51800.2020.00056.

[83] H. Dani, J. Li, and H. Liu, "Sentiment informed cyberbullying detection in social media," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, vol. 10534. Skopje, Macedonia: Springer, 2017, pp. 52–67, doi: 10.1007/978-3-319-71249-9_4.

[84] N. Arató, A. N. Zsidó, K. Lénárd, and B. Lábadi, "Cybervictimization and cyberbullying: The role of socio-emotional skills," *Frontiers Psychiatry*, vol. 11, p. 248, Apr. 2020, doi: 10.3389/fpsyt.2020.00248.

[85] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using Twitter users' psychological features and machine learning," *Comput. Secur.*, vol. 90, Mar. 2020, Art. no. 101710, doi: 10.1016/j.cose.2019.101710.

[86] L. Ge and T. Moh, "Improving text classification with word embedding," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 1796–1805.

[87] S. Pericherla and E. Ilavarasan, "Performance analysis of word embeddings for cyberbullying detection," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 1085, no. 1, 2021, Art. no. 012008, doi: 10.1088/1757-899X/1085/1/012008.

[88] J. Pennebaker, R. Booth, R. Boyd, and M. Francis, *Linguistic Inquiry and Word Count: LIWC2015. 2015*. Austin, TX, USA: Pennebaker Conglomerates, 2015. [Online]. Available: www.LIWC.net

[89] R. L. Boyd, A. Ashokkumar, S. Seraj, and J. W. Pennebaker, *The Development and Psychometric Properties of LIWC-22*. Austin, TX, USA: Univ. Texas Austin, 2022. [Online]. Available: https://www.liwc.app

[90] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the Instagram social network," in *Proc. Int. Conf. Social Informat.*, vol. 9471. Beijing, China: Springer, 2015, pp. 49–66, doi: 10.1007/978-3-319-27433-1_4.

[91] L. Cheng, R. Guo, and H. Liu, "Robust cyberbullying detection with causal interpretation," in *Proc. Companion World Wide Web Conf.*, May 2019, pp. 169–175.

[92] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "XBully: Cyberbullying detection within a multi-modal context," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, Jan. 2019, pp. 339–347.

[93] M. van Geel, A. Goemans, F. Toprak, and P. Vedder, "Which personality traits are related to traditional bullying and cyberbullying? A study with the big five, dark triad and sadism," *Personality Individual Differences*, vol. 106, pp. 231–235, Feb. 2017, doi: 10.1016/j.paid.2016.10.063.

[94] J. Bayzick, A. Kontostathis, and L. Edwards, ''Detecting the presence of cyberbullying using computer software,'' in *Proc. 3rd Int. Web Sci. Conf.*, Koblenz, Germany, 2011, pp. 1–2.

[95] H. Sanchez and S. Kumar, ''Twitter bullying detection,'' in *Proc. NSDI*, vol. 12, 2011, p. 15.

[96] A. Kumar and N. Sachdeva, ''Cyberbullying checker: Online bully content detection using hybrid supervised learning,'' in *Proc. Int. Conf. Intell. Comput. Smart Commun.* Tehri, India: Springer, 2020, pp. 371–382, doi: 10.1007/978-981-15-0633-8_36.

[97] S. Mahbub, E. Pardede, and A. S. M. Kayes, ''Detection of harassment type of cyberbullying: A dictionary of approach words and its impact,'' *Secur. Commun. Netw.*, vol. 2021, pp. 1–12, Jun. 2021, doi: 10.1155/2021/5594175.

[98] J. L. Bigelow, A. E. (Kontostathis), and L. Edwards, ''Detecting cyberbullying using latent semantic indexing,'' in *Proc. 1st Int. Workshop Comput. Methods CyberSafety*, Oct. 2016, pp. 11–14.

[99] H.-T. Kao, S. Yan, D. Huang, N. Bartley, H. Hosseinmardi, and E. Ferrara, ''Understanding cyberbullying on Instagram and Ask.Fm via social role detection,'' in *Proc. Companion World Wide Web Conf.*, May 2019, pp. 183–188.

[100] E. Wulczyn, N. Thain, and L. Dixon, ''Ex machina: Personal attacks seen at scale,'' in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 1391–1399, doi: 10.1145/3038912.3052591.

[101] K. Reynolds, A. Kontostathis, and L. Edwards, ''Using machine learning to detect cyberbullying,'' in *Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops*, Dec. 2011, pp. 241–244, doi: 10.1109/ICMLA.2011.152.

[102] K. N. Singh, S. D. Devi, H. M. Devi, and A. K. Mahanta, ''A novel approach for dimension reduction using word embedding: An enhanced text classification approach,'' *Int. J. Inf. Manage. Data Insights*, vol. 2, no. 1, Apr. 2022, Art. no. 100061, doi: 10.1016/j.jjimei.2022.100061.

[103] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, and G. S. Choi, ''Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model,'' *IEEE Access*, vol. 9, pp. 78621–78634, 2021, doi: 10.1109/ACCESS.2021.3083638.

[104] Y. Xiao and K. Cho, ''Efficient character-level document classification by combining convolution and recurrent layers,'' 2016, *arXiv:1602.00367*.

[105] C. Hutto and E. Gilbert, ''VADER: A parsimonious rule-based model for sentiment analysis of social media text,'' in *Proc. Int. AAAI Conf. web Social Media*, vol. 8, no. 1, 2014, pp. 216–225.

[106] F. Nielsen, ''A new ANEW: Evaluation of a word list for sentiment analysis in microblogs,'' in *Proc. ESWC Workshop 'Making Sense Microposts', Big Things Come Small Packages*, vol. 718. Heraklion, Crete: CEUR Workshop Proceedings, 2011, pp. 93–98.

[107] S. Mohammad and P. Turney, ''Emotions evoked by common words and phrases: Using mechanical Turk to create an emotion lexicon,'' in *Proc. NAACL HLT Workshop Comput. Approaches Anal. Gener. Emotion Text*, Los Angeles, CA, USA, 2010, pp. 26–34.

[108] Q. Liu, J. Wang, D. Zhang, Y. Yang, and N. Wang, ''Text features extraction based on TF-IDF associating semantic,'' in *Proc. IEEE 4th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2018, pp. 2338–2343, doi: 10.1109/CompComm.2018.8780663.

[109] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, ''Distributed representations of words and phrases and their compositionality,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, doi: 10.48550/arXiv.1310.4546.

[110] J. Pennington, R. Socher, and C. Manning, ''Glove: Global vectors for word representation,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.

[111] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, ''Advances in pre-training distributed word representations,'' 2017, *arXiv:1712.09405*.

[112] S. S. Birunda and R. K. Devi, ''A review on word embedding techniques for text classification,'' in *Innovative Data Communication Technologies and Application*. Cham, Switzerland: Springer, 2021, pp. 267–281.

[113] H. Tanaka, H. Shinnou, R. Cao, J. Bai, and W. Ma, ''Document classification by word embeddings of BERT,'' in *Proc. Int. Conf. Pacific Assoc. Comput. Linguistics*. Hanoi, Vietnam: Springer, 2019, pp. 145–154, doi: 10.1007/978-981-15-6168-9_13.

[114] M. E. Peters, *Deep Contextualized Word Representations*. New Orleans, LA, USA: Association for Computational Linguistics, 2018, pp. 2227–2237, doi: 10.18653/v1/N18-1202.

[115] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' 2018, *arXiv:1810.04805*.

[116] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, ''ALBERT: A lite BERT for self-supervised learning of language representations,'' 2019, *arXiv:1909.11942*.

[117] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, ''DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter,'' 2019, *arXiv:1910.01108*.

[118] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, ''ELECTRA: Pre-training text encoders as discriminators rather than generators,'' 2020, *arXiv:2003.10555*.

[119] N. Shazeer, ''GLU variants improve transformer,'' 2020, *arXiv:2002.05202*.

[120] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, ''MobileBERT: A compact task-agnostic BERT for resource-limited devices,'' 2020, *arXiv:2004.02984*.

[121] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, ''RoBERTa: A robustly optimized BERT pretraining approach,'' 2019, *arXiv:1907.11692*.

[122] C. Roberts, A. Milsted, M. Ganahl, A. Zalcman, B. Fontaine, Y. Zou, J. Hidary, G. Vidal, and S. Leichenauer, ''TensorNetwork: A library for physics and machine learning,'' 2019, *arXiv:1905.01330*.

[123] E. Fast, B. Chen, and M. S. Bernstein, ''Empath: Understanding topic signals in large-scale text,'' in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, San Jose, CA, USA, 2016, pp. 4647–4657.

[124] A. M. Schoene and N. Dethlefs, ''Unsupervised suicide note classification,'' in *Proc. Workshop Issues Sentiment Discovery Opinion Mining Knowl. Discovery Data Mining (KDD)*, London, U.K., 2018, pp. 1–9.

[125] C. E. Osgood and E. G. Walker, ''Motivation and language behavior: A content analysis of suicide notes,'' *J. Abnormal Social Psychol.*, vol. 59, no. 1, pp. 58–67, Jul. 1959, doi: 10.1037/h0047078.

[126] T. Tytko and M. C. Augstkalns. (2020). *How Well Do We Know Ourselves? Identifying Suicide Markers in Online Communication: A Case Study of a Graduate Student's Writing*. [Online]. Available: http://hdl.handle.net/2142/109011

[127] P. L. Teh, C.-B. Cheng, and W. M. Chee, ''Identifying and categorising profane words in hate speech,'' in *Proc. 2nd Int. Conf. Compute Data Anal.*, Mar. 2018, pp. 65–69, doi: 10.1145/3193077.3193078.

[128] M. Alruily, ''Towards automatically extracting contextual valence shifters in reviews of Saudi universities,'' in *Proc. IEEE 9th Symp. Comput. Appl. Ind. Electron. (ISCAIE)*, Apr. 2019, pp. 278–281, doi: 10.1109/ISCAIE.2019.8743859.

[129] S. Cacchiani, ''Cognitive motivation in english complex intensifying adjectives,'' *Lexis*, no. 10, pp. 1–22, Jun. 2016, doi: 10.4000/lexis.1079.

[130] F. Strohm, ''The impact of intensifiers, diminishers and negations on emotion expressions,'' M.S. thesis, Univ. Stuttgart, Germany, 2017, doi: 10.18419/opus-9648.

[131] F. Sommar and M. Wielondek, ''Combining lexicon-and learning-based approaches for improved performance and convenience in sentiment classification,'' Ph.D. dissertation, KTH Roy. Inst. Technol., School Comput. Sci. Commun. (CSC), Stockholm, Swedan, 2015.

[132] P. G. Hoang, L. T. Nguyen, and K. Nguyen, ''UIT-E10dot3 at SemEval-2021 task 5: Toxic spans detection with named entity recognition and question-answering approaches,'' in *Proc. 15th Int. Workshop Semantic Eval. (SemEval-)*, 2021, pp. 919–926, doi: 10.18653/v1/2021.semeval-1.125.

[133] E. Wong, S. Santurkar, and A. Madry, ''Leveraging sparse linear layers for debuggable deep networks,'' in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11205–11216.

[134] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ''SMOTE: Synthetic minority over-sampling technique,'' *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[135] A. I. Kadhim, ''Survey on supervised machine learning techniques for automatic text classification,'' *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 273–292, Jun. 2019, doi: 10.1007/s10462-018-09677-1.

[136] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, ''Text classification algorithms: A survey,'' *Information*, vol. 10, no. 4, p. 150, Apr. 2019, doi: 10.3390/info10040150.

[137] M. Goudjil, M. Koudil, M. Bedda, and N. Ghoggali, ''A novel active learning method using SVM for text classification,'' *Int. J. Autom. Comput.*, vol. 15, no. 3, pp. 290–298, Jun. 2018, doi: 10.1007/s11633-015-0912-z.

[138] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A comparative analysis of logistic regression, random forest and KNN models for the text classification," *Augmented Hum. Res.*, vol. 5, no. 1, pp. 1–16, Dec. 2020, doi: 10.1007/s41133-020-00032-0.

[139] C. N. Kamath, S. S. Bukhari, and A. Dengel, "Comparative study between traditional machine learning and deep learning approaches for text classification," in *Proc. ACM Symp. Document Eng.*, Aug. 2018, pp. 1–11.

[140] A. K. Uysal, "On two-stage feature selection methods for text classification," *IEEE Access*, vol. 6, pp. 43233–43251, 2018, doi: 10.1109/ACCESS.2018.2863547.

[141] K. Taneja and J. Vashishtha, "Comparison of transfer learning and traditional machine learning approach for text classification," in *Proc. 9th Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2022, pp. 195–200, doi: 10.23919/INDIACom54597.2022.9763279.

[142] T. Wolf, "HuggingFace's transformers: State-of-the-art natural language processing," 2019, *arXiv:1910.03771*.

[143] Y. An, X. Tang, and B. Xie, "Sentiment analysis for short Chinese text based on character-level methods," in *Proc. 9th Int. Conf. Knowl. Smart Technol. (KST)*, Feb. 2017, pp. 78–82, doi: 10.1109/KST.2017.7886093.

[144] M. Thangaraj and M. Sivakami, "Text classification techniques: A literature review," *Interdiscipl. J. Inf., Knowl., Manage.*, vol. 13, pp. 117–135, Jan. 2018, doi: 10.28945/4066.

[145] W. N. H. W. Ali, M. Mohd, and F. Fauzi, "Cyberbullying predictive model: Implementation of machine learning approach," in *Proc. 5th Int. Conf. Inf. Retr. Knowl. Manage. (CAMP)*, Jun. 2021, pp. 65–69, doi: 10.1109/CAMP51653.2021.9497932.

**TEOH HWAI TENG** received the bachelor's degree (Hons.) in science, majoring in statistics from the Mathematical Department, Faculty of Science, Universiti Putra Malaysia, in 2018, and the master's degree (Hons.) in data science from the Faculty of Computer Science and Technology, University of Malaya, Malaysia. She was a part-time statistics tutor during her final year of studies. She was one of the presenters of Seminar Kebangsaan Institut Statistik Malaysia ke-12 (SKISM-XII) organized by the Institute of Statistics Malaysia (ISM), in 2018. She has over three years of experience in business intelligence, analysis, and visualization for global and regional reinsurance studies. She is skilled in data analytics, data visualization, machine learning, project management, and various programming languages. She is currently working at leading insurance company as a data scientist. Her research interests include machine learning and computational text analysis with natural language processing.

**KASTURI DEWI VARATHAN** received the Ph.D. degree in computer science from the National University of Malaysia. She was a Visiting Scientist with the Information Systems Research Group, Faculty of Informatics, University of Lugano, Switzerland, and a Research Fellow with the Institute of Visual Informatics, National University of Malaysia. She is currently an Associate Professor with the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. She has published in several high-ranked journals and conferences. Her research interests include data analytics, machine learning, and information retrieval. Her research on data analytics won several awards in international and national level competitions. Her research institutions too have gained from her expertise. She is also a recipient of the Prestigious Leadership in Innovation Fellowship Award by the U.K. and Malaysian Government.