IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 July 2024, accepted 23 July 2024, date of publication 30 July 2024, date of current version 9 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3435840

RESEARCH ARTICLE

A Chained Deep Learning Model for Fine-Grained Cyberbullying Detection With Bystander Dynamics

HAIFA SALEH ALFURAYJ[1,2], SYAHEERAH LEBAI LUTFI[2], (Member, IEEE), AND RAMESH PERUMAL[3], (Member, IEEE)

[1]College of Computer, Qassim University, Buraydah, Al Qassim 52571, Saudi Arabia
[2]School of Computer Sciences, Universiti Sains Malaysia, Gelugor, Penang 11800, Malaysia
[3]Intel Microelectronics (M) Sdn., Bhd., Bayan Lepas, Penang 11900, Malaysia

Corresponding author: Syaheerah Lebai Lutfi (Syaheerah@usm.my)

This work was supported in part by Universiti Sains Malaysia.

ABSTRACT Accurate detection of cyberbullying on Social Networking Sites (SNSs) is crucial for online safety, especially for individuals impacted by it. Cyberbullying language is often implicit, necessitating a comprehensive analysis of conversational context to determine intention and severity. Current studies on cyberbully detection predominantly focus on the main post, overlooking the valuable insights provided by bystander reactions. Moreover, confusion over the differences between cyber-aggression and cyberbullying can undermine the reliability of some of these studies. To ameliorate these issues, this paper addresses the gap in existing research by emphasizing the significance of bystander in fine-grained cyberbullying detection. We specifically investigate the influence of bystander for more precise identification of cyberbullying attacks. Our approach involves fine-tuning a pre-trained language model (BERT) to identify features associated with bystander, categorizing them into roles such as defender, instigator, neutral, and other. To enhance fine-grained cyberbullying detection, we propose a classifier chain that combines BERT's output with Long Short-Term Memory (LSTM) networks. Our experiments demonstrate that involving bystander roles increases the model's performance by 25.35%, achieving a final classification F1-score of 89%. Moreover, our approach accurately determines the level of aggression in cyberbullying incidents through considering the interdependency of bystander roles and cyberbullying classes labels, providing an innovative solution to the challenge of cyberbullying misclassification. In summary, our study highlights the significance of incorporating bystanders as a feature in cyberbullying detection and introduces a chained fine-grained model that surpasses conventional approaches, demonstrating promising outcomes in precisely classifying cyberbullying incidents.

INDEX TERMS Aggression, bystander, BERT, cyberbullying detection, chain, deep learning, LSTM.

I. INTRODUCTION

Recently, cyberbullying has become one of the most important areas of research and the focus of researchers due to its high prevalence rates of cyberbullying, with an average of 41% of U.S. adults have been harassed online at various reasons as reported in a Pew Research Center survey conducted in September 2021 [1]. Cyberbullying victimization has been associated with various adverse health outcomes. In [2], the authors conducted a meta-analysis of studies primarily focusing on adolescents and young adults to examine the relationships between cyberbullying victimization and psychosocial and behavioral variables. The findings revealed that individuals reporting high levels of cyberbullying victimization also tended to report high levels of stress, anxiety, depression, loneliness,

The associate editor coordinating the review of this manuscript and approving it for publication was Alba Amato.

conduct-emotional-somatic problems, and drug and alcohol use. In severe cases, there was also the potential for self-harm and suicidal ideation. Another study [3] on adolescents aimed to investigate the effect of cyberbullying victimization on suicide ideation. The results indicate that the relationship between cyberbullying victimization and self-harm thoughts is stronger among individuals experiencing emotional distress.

Cyberbullying is defined as frequent aggressive behavior carried out electronically by a person or a group of people, aimed at inflicting harm on a person who cannot easily fight back, creating a power imbalance in which the bully has power over the victim. Cyber-aggression, on the other hand, is defined as aggressive behavior intended to cause harm to a person (e.g., name-calling by an anonymous online user) [2], [4], [5], [6]. Using the above definitions, we explore whether there are distinguishing characteristics that differentiate cyberbullying instances from cyber-aggression, These characteristics include:(a)frequency: Cyber-aggression involves aggressive behavior intended to cause harm but may not occur frequently. In contrast, cyberbullying is characterized by repetitive and frequent aggression. (b)power Imbalance: Cyber-aggression may involve harmful actions without necessarily creating a power imbalance. Cyberbullying, on the other hand, is specifically defined as behavior that aims to create a power imbalance, with the bully having power over the victim. These characteristics highlight the nuanced differences in severity levels between cyber-aggression and cyberbullying, indicating that while all cyberbullying instances involve aggression, not all cyber-aggression qualifies as cyberbullying.

Within the dynamics of bullying, key actors include the perpetrator (bully), the victim, and bystanders. As highlighted by [7], traditional bullying research underscores the significant role of bystanders in bullying instances. Bystanders can assume different roles: *neutrals*, where they remain impartial or passive, acting as *instigators* who mimic the bully, or serving as *defenders* who support the victim. In this study, we add an additional bystander role referred to as *others*, which represents bystanders posting unrelated consent. The *others* category helps both annotators and the model handle ambiguous bystander roles.

To address the risks posed by cyberbullying, there has been a heightened focus on automated cyberbullying detection, resulting in the creation of numerous detection models. However, existing studies predominantly rely on standalone posts and treat cyberbullying detection as a binary classification task, distinguishing between "bullying" and "not bullying" posts, often neglecting bystanders involvement and the broader conversational context. Cyberbullying incidents can manifest varying levels of aggression, influenced by factors such as the bystanders involved [8]. Based on findings from the Pew Research Center, internet users encounter two levels of online harassment:less severe and more severe experiences [9]. Therefore, it is crucial to evaluate cyberbullying incidents by classifying them into low and high aggression levels to accurately depict their impact on victims. This classification would help inform Twitter's moderating policy to ensure that appropriate action occurs depending upon the level of aggression (bullying severity), ranging from issuing a warning to user blocking.

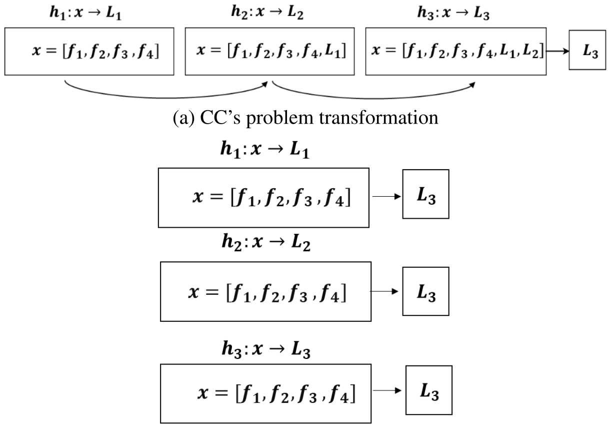
Despite bystanders being recognized as playing a key role in the dynamics of cyberbullying, researchers have not thoroughly explored how this feature influences the varying degrees of severity of cyberbullying incidents on victims. Therefore, it is essential to develop a method to detect cyberbullying according to fine-grained classifications: bullying with low aggression, bullying with high aggression, and aggression without bullying. This approach involves considering the entire conversation thread to capture the roles of bystanders and their influence on fine-grained cyberbullying detection.

A. RESEARCH PROBLEM

Examining the types of bystander roles involved in cyberbullying is crucial for the development of effective detection models. This is particularly important because cyberbullying involves real-world situations with multiple events initiated by a group. The classification of certain instances depends on interdependent events. Neglecting this interdependence may lead to inefficient detection results. Research suggests that using multi-label learning approaches, which explicitly consider label interdependence, generally results in better predictive performance [10], [11]. Therefore, label dependence is critical for achieving better performance in multi-label classification problems. Label dependence is integrated into the model through two sequential classification stages, where the output of the first stage feeds into the second one. This integration is important because in some instances, classifying the severity of bullying depends on identifying the bystander roles within the same thread.

Many earlier models for multi-label classification used the problem transformation method, either through Classifiers Chain (CC) or Binary Relevance (BR). BR is common due to its simplicity; it independently trains a binary classifier for each label, overlooking the explicit interaction among events. On the other hand, the CC method employs multiple classifiers, equal to the number of labels, with each one trained for a specific label. To classify a new instance, CC predicts the value of the first label, then uses this prediction along with the instance to predict the value of the next label. This process continues until the final label is predicted [11]. Fig. 1 illustrates how CC as shown in Fig. 1a considers label dependencies in the problem transformation method, comparing it to BR. The Classifier Chain (CC) model is widely adopted and popular for its ability to address label dependency, simplicity, and promising experimental results.

This research aims to understand the roles of bystanders in cyberbullying and assess how they impact the effectiveness of detecting cyberbullying with varying levels of aggression


$h_1: x \rightarrow L_1$
$h_2: x \rightarrow L_2$
$h_3: x \rightarrow L_3$
$x = [f_1, f_2, f_3, f_4]$
$x = [f_1, f_2, f_3, f_4, L_1]$
$x = [f_1, f_2, f_3, f_4, L_1, L_2]$
L_3
$h_1: x \rightarrow L_1$
$x = [f_1, f_2, f_3, f_4]$
L_3
$h_2: x \rightarrow L_2$
$x = [f_1, f_2, f_3, f_4]$
L_3
$h_3: x \rightarrow L_3$
$x = [f_1, f_2, f_3, f_4]$
L_3

(a) CC's problem transformation

(b) BR's problem transformation

FIGURE 1. The comparison of BR and CC problem transformation methods for each classifier h_j, which is trained to predict L_j.

(fine-grained). Unlike a simple binary classification, we use a more detailed approach to classify cyberbullying based on different aggression levels. Specifically, this is done by taking into account various bystander roles in a cyberbullying setting to determine the aggression level.

In this research, we enhance the fine-grained detection of cyberbullying by identifying different bystander roles in a Twitter thread using the CYBY23 dataset[1] [12]. While much of the existing research focuses on cyberbullying detection without considering bystander roles, the present study is the first to classify these roles to improve the fine-grained cyberbullying detection. Despite the limited size of the dataset, it is the only available session-based dataset that includes bystanders' roles label along with fine-grained bullying label, making it suitable for use with the proposed chained model. Fig. 2 illustrates the implementation of the proposed chained fine-grained cyberbullying detection model, which considers interdependent events involving bystanders.

The rest of this article is structured as follows:

- Related work is in Section II;
- Gap analysis is in Section III;
- Our proposed methodology is explained in Section IV;
- The experiments setting, and results are discussed in Section V and VI;
- Section VII is the conclusion.

B. RESEARCH OBJECTIVES AND CONTRIBUTIONS

This research has two main objectives:

1) To identify a feature associated with bystander roles and,
2) To develop a fine-grained cyberbullying detection model that uses Bystander Roles feature to learn a better semantic representation of the text that helps in detecting cyberbullying instances with varying levels of aggression.

The contributions of this work are:

1) The incorporation of the bystander as a feature in the introduction has notably enhanced the detection model, enabling the inclusion of fine-grained labels beyond binary categorization.
2) The methodological innovation in this study lies in the introduction of a novel chained architecture model, comprising two sequential classification layers,
3) The evaluation of our model involves a dual-pronged baseline comparison. Firstly, we showcase the effectiveness of our method by contrasting its performance, augmented with the bystander roles feature, against its performance in a baseline dataset consisting of individual tweets without this feature. Secondly, our evaluation extends to a comparative analysis with the outcomes reported in the existing literature for the most closely aligned state-of-the-art models.

II. RELATED WORKS

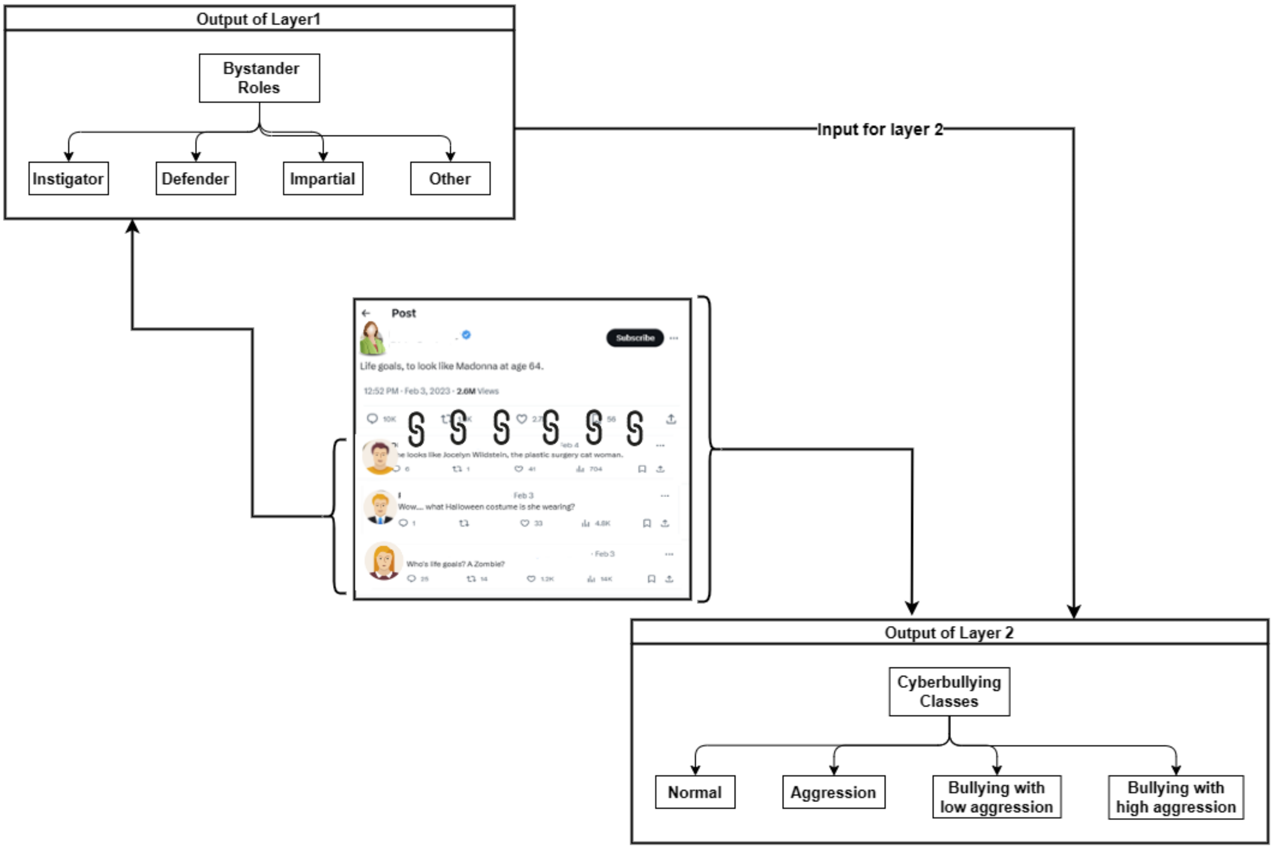
The increasing incidence of cyberbullying in Online Social Networks (OSNs), involving the sharing of text along with images, has led to an increasing research attention on developing methods for detecting cyberbullying in both textual and visual contexts.

In recent studies focusing on detecting cyberbullying through binary classification, the deep learning-based LSTM model outperforms the traditional machine learning classifiers in text cyberbullying detection by extracting the features using Term Frequency-Inverse Document Frequency (TF-IDF), which captures the frequency of the individual words [13]. Additionally, in [14], a soft-vote ensemble learning strategy was proposed to combine and refine the prediction results of the constructed multimodal feature with GIN feature transformation, and the embedded features using Bidirectional Encoder Representations from Transformers (BERT) and Vision Transformers (ViT). This approach enhances the integration of text-image features, while mitigating bias in data structure information. BERT is used for extracting text features, which have been shown in study by [15] to outperform other commonly used deep learning models in various cyberbullying datasets.

A few attempts to improve the detection of cyberbullying are closely related to our work, as summarized in Table 1. Some of the previous studies have utilized the inclusion of the entire contextual conversation with different methods to investigate the improvement of cyberbullying detection method with the conversation structure on Instagram, Vine and Twitter [16], [17], [18], [19].

A related study by Ziems et al. [19] emphasized the significance of incorporating the entire thread into their labeling scheme to accommodate the cyberbullying definition. This involves capturing the full context of each tweet, acquiring both the replies in the thread and the list of the most recent mentions of users. The process of data annotation is

[1] CYBY23 dataset is publicly available at (http://syaheerah.com/wp content/uploads/2023/07/CYBY23-Dataset.zip)



Life goals, to look like Madonna at age 64.

12:52 PM · Feb 3, 2023 · 2.6M Views

Feb 4
he looks like Jocelyn Wildstein, the plastic surgery cat woman.

Feb 3
Wow.... what Halloween costume is she wearing?

· Feb 3
Who's life goals? A Zombie?

FIGURE 2. Illustrations of the two sequential layers of the proposed chained fine-grained cyberbullying detection model.

TABLE 1. Comparison between state-of-the-art cyberbullying detection methods and the proposed method.

Study	participants bully & victim	Bystanders	Session-based	Fine grained bullying label	Approach
[19]	no	no	yes	no	single classification stage
[17]	no	yes (positive-negative commenters)	yes	no	single classification stage
[16]	no	no	yes	no	single classification stage
[20]	yes	yes (harasser-defender)	no	no	single classification stage
[18]	no	no	yes	yes (slight-medium-serious)	single classification stage
proposed*	no	yes (instigator-defender-impartial-other)	yes	yes (aggression-bullying with low aggression-bullying with high aggression)	two sequential classification stages

conducted by three different annotators who are provided with the entire tweet thread to binary label five key criteria for cyberbullying: aggressive language, repetition, harmful intent, peer visibility, and power imbalance. They achieved the optimal outcomes by adopting an approach based on AdaBoost using both labeled features and automatically labeled features such as social network features, textual features, user-based features, timeline features, and thread features. Despite the low precision and recall scores of their classifier's results, through their research, it is acknowledged that the criteria for cyberbullying can be subjective, relying on involving the whole social features surrounding the text.

In another approach, the authors [17] conducted a detailed analysis of labeled media sessions using (1) time and sentiment analysis using Python's NLTK library, (2) Inverse Document Frequency (IDF) analysis, and (3) text content analysis - of two labeled datasets. They aimed to distinguish between cyber-aggression and cyberbullying instances. While their study did not involve the utilization of a detection model, it sheds light on bystander as a significant factor to differentiate a cyberbullying session from a non-cyberbullying one.

Whereas other work has emphasized the importance of employing a session-based cyberbullying detection, which provides a more realistic representation of user interactions and captures the repetitive criteria of bullying behaviors [16]. They introduced a temporal graph-based cyberbullying model (TGBully) which consists of three models: (1) a semantic context modeling using bidirectional GRUs, (2) a temporal graph interaction learning model leverages graph attention network (GATs), and (3) classification model that aims to aggregates all user representations within the session into the session representation to classify a social media session into a bullying or non-bullying.

On the other hand, some of the previous research on cyberbullying detection has been performed in the area of efficient and accurate detection considering the inclusion of participants' roles. However, their research was mostly based on datasets that were limitedly labeled based on post level [20]. They proposed a detection model that considered different participant roles, focusing on detecting four specific categories: not bullying, harasser, victim, and bystander-defender, using a cyberbullying traces dataset. They conducted multiclass classification experiments using three distinct approaches: employing a traditional machine learning classifier, utilizing an ensemble classifier, and leveraging pre-trained language models based on transformers.

According to our search, we encountered a single study that challenges the common belief by positing that cyberbullying should not be treated as a binary issue. This study advocates for the adoption of a multi-class labeling approach in the detection of cyberbullying, aiming to identify different levels of bullying severity: slight, medium, and serious [18]. The researchers of this study have curated a dataset comprising Chinese-language cyberbullying instances, consisting of complete conversational dialogues that have been annotated with severity ratings for cyberbullying incidents falling within the aforementioned categories. The determination of cyberbullying severity for each dialogue is based on the assessment of three primary bullying criteria: intent, repetition, and aggression. They implement a hierarchical squashing-attention network (HSAN) model. The model is based on hierarchical text-encoding and attention methods and consists of four parts: The word encoder employed GRU and a word squashing- attention mechanism to estimate the weight of each hidden feature of a word. A sentence encoder employed the bidirectional GRU to obtain the hidden features of the sentence and an attention mechanism to estimate the attention weight of the hidden features of each sentence. Their approach involves examining cyberbullying samples obtained from interactive dialogues, which consist of a title, a primary piece of content, and a minimum of two response comments.

However, these studies treated each bullying sample independently, regardless of the bystander replies. This could lead to non-optimal classification accuracy or result in misclassified instances. To overcome these issues, we proposed a chained fine-grained cyberbullying detection model considering the impact of bystander replies. Our proposal addresses the limitations of existing studies by treating the detection of cyberbullying as a two-stage classification process, which captures the potential dependencies between the main posts and the replies from bystanders. This allows the proposed approach to distinguish the level of aggression in cyberbullying from aggression instances, thus enhancing cyberbullying detection performance.

III. GAP ANALYSIS

It is acknowledged that cyberbullying involves different participants, ranging from the bully, victim, to the bystander. Despite the fact that bystanders could influence the severity of bullying, the majority of research, until recently, has primarily focused on the victims and perpetrators of cyberbullying. However, a recent systematic review [21] indicates a growing recognition of the pivotal role of the bystanders in addressing the issue of cyberbullying.

Several social and education studies have emerged [22], [23] with the aim of investigating the influence of bystanders on cyberbullying. The actions taken by bystanders in such situations have significant implications for both the victims and society. Those who defend the victims by offering support can potentially mitigate the harmful effects of bullying. Conversely, instigators who align themselves with the perpetrators (bullies) may exacerbate the situation for the victim. Hence, the feature associated with bystander roles may significantly influence the severity of bullying.

However, the majority of existing research in the field of cyberbullying detection ignores the impact of bystanders and relies on individual tweets, leading to a limitation in their findings. The drawback of implementing this approach lies in its inability to detect the hidden meaning in indirect aggressive content. Additionally, it has limited ability to address cyberbullying instances involving explicit profanity or stereotyped words, sometimes used for joking purposes in the slang language used on social networking sites (SNS).

This proposed study involves inspecting the entire tweet thread within a comprehensive context for the detection of bystander roles and, consequently, the severity of bullying. Given that every tweet will be categorized for both the severity of bullying and the role of the bystander, this issue falls under the category of a multi-label, multi-class classification problem. In the realm of machine learning, this type of problem is typically addressed using a multi-label classifier such as binary relevance and the chained classifier. The former, mostly used in the context of multi-label cyberbullying classification, handles binary label classification concurrently and independently. The latter, as proposed in this paper, employs two sequential classification layers to capitalize on the correlations between the labels. This sequential classification method is appropriate for our problem since we have more than one label correlated in a certain sequence.

IV. PROPOSED METHODOLOGY

As shown in Fig. 4a, the methodology section consists of two main parts. First part is to present the main process of the used dataset. Next, it is proposed the core of the paper, which introduces the chained fine-grained cyberbullying detection model construction.

A. DATA COLLECTION

In order to construct representative models for cyberbullying detection, it is essential to have an appropriate dataset. The utilized corpus was constructed by collecting data from the social networking site Twitter, referred to as 'X' nowadays, where users can post tweets. On this platform, users can post tweets, and bystanders can engage with a tweet by replying to the post, forming a thread with the initial post and its corresponding replies.

The CYBY23 corpus provided by [12] for the fine-grained classification of cyberbullying threads. The Twitter API were used for corpus scraping using keyword searches and hashtags that are inherently controversial and could give rise to harassment content. For example, "hijab," "immigrant," "racism," "Nazi," "gypsy," "immigration," "bi**h," and "f**k" were used. Approximately 13,300 Tweets were collected within a 12-month timeframe spanning January 2022 until January 2023.

The manual labeling process is a time-consuming task. Thus, to test the performance considering the available resources, the annotation is done with a small portion of tweets including 150 tweet threads with approximately 1024 tweets. The annotation scheme is divided into three levels of different tasks for annotators as follow:

- Tweet level: rating the aggressiveness score with three-point scale (0-1-2) for the tweet post.
- Replies level: identifying the roles of the bystander through inspecting the replies.
- Thread level: ensuring the validity of their initial responses after reading the main post with the replies and examining the bystander roles; having a change of mind (CoM) chance for aggressiveness rating score.

For more comprehensive information, the details were provided in our previous research [12], [24].

As real Twitter users, the annotators were from different countries, cultures, religions, and age groups. The annotation is created using 15 Google Forms, each containing batches of 10 threads (initial post with its multi replies) with a different number of replies. Each form is annotated by five different annotators. The threads with low agreement were excluded, resulting in a reduction of tweets to 639. This ensures a reliable dataset and leads to a promising Fleiss' Kappa value of 0.6078.

The corpus includes two labels with multiple selections as follows:

- Fine-grained cyberbullying label: bullying with high aggression, bullying with low aggression, or aggression without indication of bullying;
- Bystander roles label: including the roles of the bystander that were identified as follows, and illustrated in Table 2:
  - *Instigators* who agree with the thread author of the main post topic.
  - *Defenders* who disagree with the thread topic and show defending manner.
  - *Impartials* who remain neutral or passive.
  - *Others* who post content not related to the thread topic.

TABLE 2. Example of annotations for bystanders' roles categories related to cyberbullying threads in CYBY23 datset.

Bystanders Roles	Samples of Bystander Replies
Defender	`Main Post:` F*ckin shill ass bitch! Just another douchebag not to trust. No surprise here #AMC #APE `@bystander` I disagree with Houston on this but we've got to stop calling everyone who disagrees with us shills. Especially when they clearly lay out their reasoning. I disagree with his analysis here and that's ok. It's unreasonable to assume that thoughtful human beings with always agree.
Instigator	`Main Post:` Bitch that's Marilyn Manson `@bystander` Yes, Dr., I would rather look like an alien than have one wrinkle
Impartial	`Main Post:` Bitch that's Marilyn Manson `@bystander` Yoda is looking thru the window behind her
Other	`Main Post:` olivia is suuuuuch a bitch had it out for zara because of tom and now tanyel because of kai she's so peak #LoveIsland `@bystander` Hon kong oil hat massage

TABLE 3. Statistic of bullying class label in CYBY23.

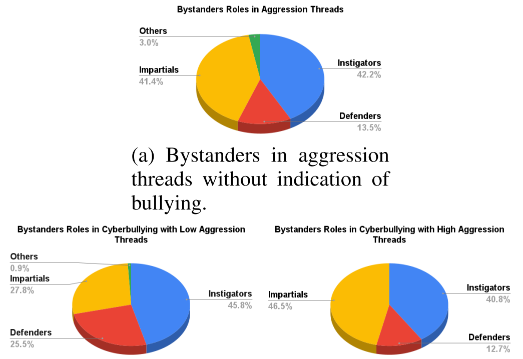
fine-grained cyberbullying label	Percentages
bullying with high aggression	11.61%
bullying with low aggression	54.46%
aggression without indication of bullying	33.93%

The lowest portion of threads are cyberbullying with high aggression class, comprising only 11.6% of the dataset. The remaining categories have varying percentage values, as indicated in Table 3. It is observed that instigators and impartial are the most frequent across all threads categories as illustrated in Fig. 3.

Given the limited instances in the CYBY23 dataset, we intend to use it for the pilot findings of our proposed method because it is the only available session-based dataset that includes bystanders roles label along with bullying label of different aggression levels. Its relevance to our proposed model lies in our aim to implement sequence classification of these labels to enhance fine-grained cyberbullying detection. '

B. PROPOSED MODEL

The proposed chained fine-grained cyberbullying detection model consists of two classification layers: bystander roles



(a) Bystanders in aggression threads without indication of bullying.

(b) Bystanders in cyberbullying with aggression threads

in(c) Bystanders in cyberbul-lowlying with high aggression threads

FIGURE 3. **The frequency distribution of bystander roles based on the aggression rate of the threads.**

TABLE 4. **Illustration of the transformation process of (x, y) in the proposed model, where y = [bystander-roles label, fine-grained cyberbullying label] and x = [reply_id, tweet text], each classifier h_i is trained to predict y_i.**

h: x –>	y
Pred 1: [reply_id, tweet text]	bystander-roles label
Pred 2: [reply_id, tweet text , **bystander-roles**]	fine-grained cyberbullying label

identification, and fine-grained cyberbullying detection. The two layers are executed sequentially to leverage the correlations between the bystander roles label and the fine-grained cyberbullying label, as shown in Fig. 4b and described in detail in the following subsections.

Layer 1: As illustrated in the top section of Fig. 4b, the first model is trained on the input features (reply_id, tweet text) and the bystander role labels. However, the fine-grained cyberbullying label is not taken into consideration. With the test set, model 1 is used to predict the bystanders roles label and updating the feature space by adding the predicted bystanders roles.

Layer 2: As illustrated in the lower section of Fig. 4b, the second model is trained on the input features, the predicted bystanders roles by the first model and the fine-grained cyberbullying label. With the test dataset, model 2 is used to predict the cyberbullying label. Table 4 illustrates the transformation process in the proposed model. The training and prediction procedures are outlined in the following algorithm.

Algorithm: the training and prediction processes in the sequential layers classification.

Input:
X: features set $X = (x_1, \ldots, x_i)$
Y: labels set $Y = (y_1, \ldots, y_i)$
F: classifiers set $F = (f_1, \ldots, f_i)$
Output:
X': Predicted labels set

```
ClassifierChain($F, Y, X$)
:for i =1 to len(Y)
    do training($f_i(X, y_i)$)
    do $y'_i < pred.(f_i(X, y_i))$
    $X' < -X' \cup y'_i$
    $X < -X \cup y'_i$            #Append pred. label to X
    $return \quad X'$
:End for
```

A chain of multi-classes classifiers (ClassifierChain) is formed, corresponding to the length of labels in the Y set. Each classifier f_i in the chain is trained for learning the y_i label given in the Y set. The classification process begins at classifier f_i and propagates along the chain, the i^{th}classifier predicts the i^{th} label, given the attribute space augmented by the predictions of the preceding classifiers in the chain. The function ClassifierChain is defined as follows:

$$\text{ClassifierChain}(F, Y, X) = \bigcup_{i=1}^{\text{Len}(Y)} \text{TrainAndPredict} \quad (f_i, \text{Update}(X), y_i) \tag{1}$$

$$\text{TrainAndPredict}(f_i, X, y_i) \quad , \text{ consists of Train, Predict, and Update functions defined as follows:} \tag{2}$$

$$\text{Train:} \quad Trainedf_i = \text{Train.}(f_i(X, y_i)) \tag{3}$$

$$\text{Predict:} \quad y'_i = \text{Pred.}(Trainedf_i(X, y_i)) \tag{4}$$

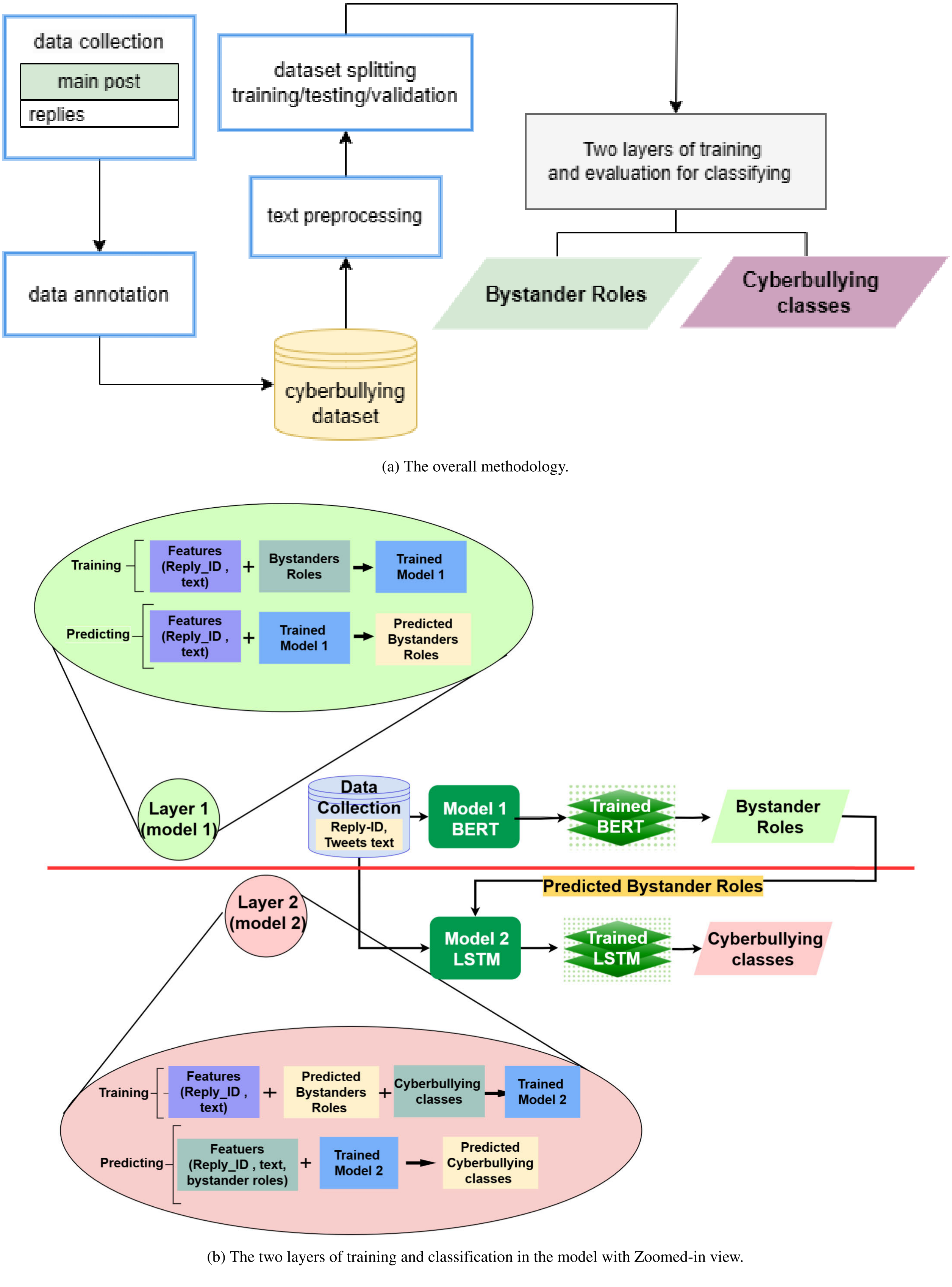
$$\text{Update:} \quad X = X \cup y'_i \tag{5}$$

The chaining method passes label information between classifiers, allowing classifier to take into account dependency and correlations between the labels.

A strong correlation between the labels will give the classifier more predictive power. In the proposed method, according to our labels, we arrange two models: the first is transfer learning approach using pre-trained BERT model to detect the bystander roles and the second is recurrent neural network (RNN) -Long Short-Term Memory (LSTM) model to perfume a fine-grained cyberbullying detection.

1) BYSTANDER ROLES MODEL

This section provides a methodological description of the transfer learning approach by implementing the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model to develop a bystander roles detection model. Our proposed chained model aims to detect cyberbullying considering the bystander roles. Implementing BERT serves as the first classifier in the chained cyberbullying detection model and its prediction is extracted as feature that feeds into the next classifier to solve a cyberbullying detection task. BERT is pre-trained on a large corpus of publicly available English text from the internet. It learns to predict words in a



(a) The overall methodology.

(b) The two layers of training and classification in the model with Zoomed-in view.

FIGURE 4. Flowchart of the chained fine-grained cyberbullying detection model.

sentence by considering the context of the surrounding words. It effectively could discern between distinct bystander roles by leveraging its transfer knowledge. Additionally, it exploits task-specific features by fine- tuning the weights of BERT's layers.

2) FINE-GRAINED CYBERBULLYING MODEL

Most studies have considered cyberbullying detection as a binary classification task, nor have they involved the bystander roles. However, this model is taken into account the entire conversational context to detect fine-grained bullying. Implementing LSTM serves as the second classifier in the chained cyberbullying detection model which incorporate the tweet texts along with the predicted bystander label from the preceding classifier to capture the interdependencies between these two labels.

LSTM is a type of recurrent neural network (RNN) designed for processing sequences of data, such as text. It's particularly effective for tasks like sentiment analysis. It allows the network to model dependencies of text over long sequences through a memory cell. This leads to addressing the exploding and vanishing gradient problem found in the standard RNN. Since we have three different inputs ($reply\_ID$, $tweettext$, $andbystanderroles$), three separated branches of the LSTM network are built with the same architecture to process the embedded inputs. The first network is used to model reply_id, the second network is used to model the tweet text, and the third network is used to model the bystander roles.

Each LSTM branch comprises four layers enhanced by the single tanh layer found in traditional RNNs. The key advantage to LSTMs is the special memory cell called "cell state". The LSTM has the ability to control the flow of information to the cell state, regulated by three main components: the forget gate, the input gate, and the output gate. The first step in each LSTM is to decide which information to discard from the memory cell C_{t-1}. This decision is made by "forget gate layer" that utilizes a *sigmoid* activation function.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \quad \text{"forget gate"} \tag{6}$$

The next step is to decide what new information should be added to the memory cell to update C_{t-1} by C_t. This is done by two layers: "input gate layer" and "tanh layer". A sigmoid layer is run to decide what parts of the cell state we're going to update.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \quad \text{"input gate"} \tag{7}$$
$$C'_t = \tanh(W_C[h_{t-1}, x_t] + b_C), \quad \text{"tanh layer"} \tag{8}$$

Finally, the "output gate layer", which involves a sigmoid function to regulate how much of the information in the memory cell should be used to generate the output. Cell state will cross tanhtanh to convert its value to vector (to push the values to be between −1 and 1) and multiply it by the sigmoid values. The multiplication output is the output information that will remain on the cell state for the next LSTM unit.

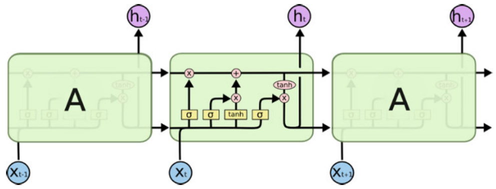


FIGURE 5. LSTM architecture [25].

$$C_t = f_t \cdot c_{t-1} + i_t \cdot \hat{C}_t, \quad \text{"cell state"} \tag{9}$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \quad \text{"output gate"} \tag{10}$$
$$h_t = o_t \cdot \tanh(C_t), \quad \text{"hidden state"} \tag{11}$$

- σ is the activation function that ranges from 0 to 1, to scale the range of data that can be completely removed, partially removed, or completely retained;
- C‘$_t$ is a "candidate" hidden state that is computed based on the current input x_t and the previous hidden state C_{t-1};
- it is the input gate that defines how much of the newly computed state for the current input x_t will pass through;
- h_{t-1} is the recurrent connection at the previous hidden layer and current hidden layer;
- W is the weight matrix connecting the inputs to the current hidden layer;
- C is the internal memory of the unit + C_{t-1};
- h_t is the output hidden state.

The structure of LSTM is shown in Fig. 5. It enables deep learning of each input through multiple steps, in order to grasp complicated relationships across long sequences of sequential data. Within each LSTM unit, decisions are made to determine the proportion of information to update, forget or retain to the subsequent step. Subsequently, the LSTMs outputs are concatenated to fed into a dense layer for classification. The final output of our fine-grained cyberbullying classification task is granted by integrating a softmax activation function with the dense layer.

3) CHAINED CLASSIFIER RATIONALE

There are different approaches to account for the interdependencies between labels, such as chained classifiers, Sequential Neural Networks (e.g., RNNs, LSTMs), and Hierarchical Models. However, addressing the research problem through a chained classifier seems more effective than employing another multi-label classifier. This preference arises from the fact that chained classifier approach accounts for the interdependencies between labels by updating the feature space at the end of each layer. This is crucial in our problem as it adds the bystander roles to the feature set, significantly enhancing the classification of bullying severity. This perspective aligns with studies in the field [26], [27], [28], [29], emphasizing the relevance of understanding label dependencies for accurate classification.

The proposed methodology employs advanced deep learning techniques to improve fine-grained cyberbullying detection by considering the interdependencies between different labels, such as bystander roles and cyberbullying classes. This is achieved through a chained classification approach integrating BERT and LSTM networks. The pre-trained BERT model excels at capturing deep, bidirectional context from text, generating embeddings that reflect the surrounding context within sentences. By fine-tuning BERT, its weights are adjusted to enhance the performance on the specific task. Following BERT, its predictions are fed into an LSTM model. The LSTM model is effective at handling sequential data, capturing long-term dependencies and patterns, which is crucial for understanding the context of cyberbullying threads. This combined approach enables us to achieve more accurate and fine-grained cyberbullying detection.

Recent studies in the detection of cyberbullying and toxic comment have shifted to implement transfer learning language models, like BERT, [15], [30], [31], [32], [33], [34]. The findings demonstrate that BERT outperforms other commonly used deep learning models on various cyberbullying-related datasets [15], [32]. Additionally, BERT models exhibit notably higher speed in detection, making them well-suited for real-time applications in cyberbullying detection [30], [33].

According to our search, we encountered a single study that challenges the common belief by positing that cyberbullying should not be treated as a binary issue. This study advocates for the adoption of a multi-class labeling approach in the detection of cyberbullying, aiming to identify different levels of bullying severity: slight, medium, and serious [18]. They implement a hierarchical squashing-attention network (HSAN) to determine the severity level of various cyberbullying incidents. Through evaluation, they discovered that their HSAN model outperformed various machine-learning and deep-learning models in accurately identifying the severity of cyberbullying. However, it's worth noting that they didn't include any RNN models in their comparative analysis. Our proposed fine-grained cyberbullying detection model is built using deep learning-based neural networks, specifically Long Short-Term Memory Recurrent Neural Networks (RNN-LSTM). RNN-LSTM has been shown to be effective in many fields and has been widely used in cyberbullying detection studies, proving successful at understanding sequences of words and interpreting their meaning [16], [32], [35].

V. EXPERIMENTS SETTING

In this section, we will evaluate the performance of the proposed chained model and the impact of bystander roles on severity of cyberbullying. All experiments were performed using CYBY23 dataset. The computing environment for the experimental implementations is as follows: 11th Gen Intel®Core™i7-1165G7 with four cores, 8 logical processors, and a processor speed of 2.80 GHz. The entire implementation process was implemented using Python programming on the PRO-Google Collaboratory platform using a T4 GPU engine with high RAM.

A. BYSTANDER ROLE MODEL (BERT)

We implement bert_base_cased which is the popular choice for natural language processing tasks. It consists of 12 layers, each layer is composed of multiple attention heads, making it a powerful model for processing and understanding text. Both lowercase and uppercase letters are preserved which is important for cyberbullying detection task where capitalization carries meaning. We used TensorFlow implementation of the BERT model that was provided by the Hugging Face Transformers library. The inputs reply_ids and tweet texts, and the output label named bystander role label are stored in xlsx file. We tokenized the inputs using the BERT tokenizer, ensuring that the tokenized sequence has a maximum length of 256 tokens. For tokenization, we defined the essential input layer for the BERT model with a shape of 256, indicating it expects input sequences of length 256. The optimal batch size was found to be 16. This input layer is used for preparing and feeding the reply_ids and tweet texts into the BERT model producing the embeddings layer. Two additional dense layers were added: intermediate layer and output layer. Intermediate layer is a dense layer with 512 units and a 'relu' activation function. This layer serves as an intermediary step between the BERT embeddings and the final output layer. The output layer is the final dense layer with 5 units according to our 5-class classification problem (bystander roles classes) and a softmax activation function to calculate the probability of classes. These additional layers are used to adapt the pre_trained BERT features for the task of bystander roles detection. We used AdamW (Adam optimizer with weight decay) as the optimization algorithm, with learning rates of 1e-5 and decay=1e-6. We fine-tuned the model for 5 epochs, which yielded the best results. The fine-tuning process took around 1 hours per epoch. Therefore, with 5 epochs, the total fine-tuning time was approximately 5 hours.

B. FINE-GRAINED CYBERBULLYING MODEL (LSTM)

This implementation section typically provides explanations of LSTM model structure, including the data tokenization, all hyper-parameter settings, dataset splitting, and evaluation. All models were created and trained using Keras, a high-level neural network API compatible with the open-source machine learning TensorFlow framework. The best parameters and outcomes were obtained through multiple rounds of experimentation.

Data Split: Data was split into training, validation, and test sets using the train_test_split function from the sklearn.model_selection module. This function is commonly used to split a dataset into two or more parts for training and testing purposes. First, dataset is divided into training and a temporary set, 40% of the data will be used for temporary, while 60% will be used for training. In order to allow the model to learn over different data instances and to uncover its reliability and consistency of results over repeated executions, the random number generator "random_state" is set to 42. The temporary data is split into validation and test sets indicates that half of the temporary data will be used for validation, and the other half for testing.

Tokenization: first of all, three inputs which are reply_ids, tweet texts, and bystander_role_labels, and one output label named class_labels stores in xlsx file. Before feeding them to the network, it's necessary to convert each input into a series of words tokenization, a crucial role in text processing serving as a fundamental step. It involves the creation of tokens through the division of textual content into words, phrases, or other significant components. Essentially, it is a form of text segmentation [36]. The text is tokenized using Tokenizer functions from TensorFlow's Keras API. This guaranteed that all sequences have the same length to be prepared for training step implementation.

Word Embedding Layer: To handle the tokenized input, an embedding layer was added to the network and textual data was provided as an input. embedding layer embedded high dimensional text data in low dimensional vector space for generating dense vector representation of data. The maximum number of vocabularies is 10,000 unique words that are embedded in a low dimensional vector with 128 dimensions. In order to encode the textual input data, word embedding has demonstrated a highly effective technique in converting discrete tokens into continuous vectors for text classification tasks. These vectors are learned during the training process, capturing semantic relationships between words. The text is padded using pad_sequences functions from TensorFlow's Keras API.

Recurrent LSTM Layer: The labels are mapped and encoded to numerical values, making them compatible with RNN-LSTM architecture. The next layer is an LSTM with 64 units(neurons). LSTMs learn sequences of words through a process called backpropagation. LSTM is a multilayers steak that works together to transform their output in a non-linear fashion based on their input. Each layer further refines the previous layer's outputs leading to enhanced selectivity and invariance in the representation. The mapping is acquired from the input data by updating the weights of internal state of the neuron through backpropagation process. Weight adjustments are made based on the error gradient at the output. The backpropagation algorithm computes this gradient and distributes it back as weight updates to the earlier layers [37]. The goal is to minimize the error in the output, and this iterative process of weight adjustment is carried out over numerous repetitions according to the epoch parameter.

Concatenation Layer: Combining the three LSTM layers using a concatenation layer, to build a multi-input classifier. On top of this layer, we add the output layer (dense layer) with a neuron dedicated to each class we aim to predict. To normalize the output values between 0 and 1, we apply a softmax activation function. At this point, the output from each neuron signifies the probability of the sample belonging to its corresponding class.

Training the LSTM Layers: We train all layers simultaneously, treating them as a unified classification system with multiple inputs. For our implementation, we utilize Keras for the deep learning models. In terms of training, we employ categorical cross-entropy as the loss function and Adam as the optimization algorithm. The choice of the epoch hyperparameter depends on various factors, including the complexity of the problem, the dataset's size, and the learning rate, which can be determined through multiple experiments.

VI. RESULTS AND DISCUSSION

In this section, we will explain and discuss the results of the evaluation experiments. Two groups of experiments were performed to evaluate the performance of the proposed chained fine-grained cyberbullying detection model: the first group aims to evaluate the bystander roles identification model, and the second group aims to evaluate the performance of the fine-grained cyberbullying detection model. In the last subsection, we included a comparison experiment with the proposed model to explore the impact of involving information related to bystanders and tweet threads on the fine-grained detection of cyberbullying.

We utilize three commonly recognized evaluation metrics in multi-class cyberbullying detection tasks, i.e., precision (Pc), recall (Rc), and F1-measure (F) [18], [20]. We pay special attention to evaluate the performance in terms of weighted metrics due to the class imbalance nature of our datasets, and to ensure that the majority class doesn't heavily influence the metric results.

They are calculated by the following equations:

$$P_c = \frac{TP}{TP + FP} \tag{12}$$

$$\text{Weighted } P_c = \sum_{i=1}^{N} W_i \times P_i \tag{13}$$

$$R_c = \frac{TP}{TP + FN} \tag{14}$$

$$\text{Weighted } R_c = \sum_{i=1}^{N} W_i \times R_c \tag{15}$$

$$F = 2 \times \frac{P_c \times R_c}{P_c + R_c} \tag{16}$$

$$\text{Weighted } f = \sum_{i=1}^{N} W_i \times F_i \tag{17}$$

$$W_i = \frac{\text{No. of samples in class } i}{\text{Total number of samples}} \tag{18}$$

In addition, the reliability of obtained results are evaluated using confidence intervals (CI) on test data. A finely tuned CIs show how trustworthy and precise of the obtained results.

A. THE PERFORMANCE OF THE BYSTANDER ROLE MODEL

1) FINE-TUNING BERT

In this section, we tested the performance of fine-tuned BERT with the parameters explained in Section V. To evaluate the model detection ability on an unseen dataset and on different training input, we conducted three experiments. We noticed an enhancement in the BERT model's performance when incorporating inputs with additional information identifying the bystander entity, the results are provided in Table 5 for all the examined approaches. Specifically, including the reply ID and retaining the user screenname in the replies (tweet text) proved to be more effective than other experiments. The reason for this is that these types of inputs help the model grasp data representation, including the relation between the main tweet and various replies from different bystanders involved in the same thread.

2) FREEZING BERT

Since the pre-trained BERT model has 12 layers already been fine-tuned during the pre-training process on a large corpus of text, we experienced incorporating the BERT model into our architecture without re-training the lower layers. The weights of the BERT layers are typically frozen to retain the weights and benefit from encoded rich knowledge. We fine-tune only the additional layers that we have added (the intermediate layer and output layer). Due to the challenging nature of our classification task, freezing the lower layers depending more on their pre-trained representations while the additional upper layers focus on learning task-specific features leads to a decrease in the performance as shown in Table 5.

3) RoBERTa

To assess BERT's effectiveness in comparison to alternative pre-trained models, we incorporate RoBERTa into our study since recent research has demonstrated that RoBERTa often outperforms BERT in classification tasks [38]. It is developed by Facebook AI and stand of 'A Robustly Optimized BERT Pre- Training Approach'. RoBERTa is an enhanced version of BERT with improvements in pre-training methodologies, training data, training duration, and other hyper-parameters. These enhancements can result in variations in performance across different tasks. We implemented a similar version of BERT in terms of parameter numbers, which is 'roberta-base'. However, Table 5 showed evident that fine-tuning BERT significantly outperformed other alternative approaches, achieving the highest performance with a weighted Recall and weighted F1-Score of 0.83.

B. PERFORMANCE OF THE CHAINED FINE-GRAINED CYBERBULLYING DETECTION MODEL

The output of bystander role detection model is fed into the fine-grained cyberbullying detection model; to obtain optimal parameter, we perform multi experiments.

Parameter Setting: The experiment was initially executed on 20 epochs to complete passes through the entire dataset and 32 batch size to moderate usage of computing resources and maintaining a reasonable training speed balance with training efficiency. The model is compiled using the Adam optimizer and a loss function (sparse_categorical_crossentropy) tailored for multi-class classification. The learning rate of Adam optimizer equals to 0.001, which is the default in Keras. Refer to Table 6 for a deeper look into the parameter.

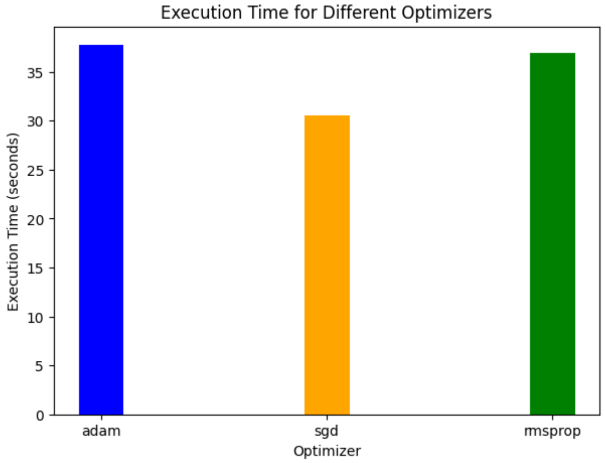


FIGURE 6. **The performance of different Optimization models(1).**

1) FIND OPTIMAL EPOCH

We trained the models using different numbers of epochs: 20, 30, 50, and 100. As seen in Table 7, it was observed that the model trained with 20 epochs yielded the most promising results, as it achieved high precision, recall, and F1 scores while minimizing losses. Following this, we further trained the models under the following configurations: 20 epochs starting from scratch, implementing early stopping, and setting a patience level of 5 epochs. We used early stopping; training is halted if the validation error ceases to improve, and the weights from the best epoch are retained. It is form of regularization that helps to prevent overfitting through terminating the training iteration when the validation error (loss) starts to increase or stops improving in a certain number of iterations.

In conjunction with early stopping, we experienced adding dropout layers to the LSTM layers with rate =.5. This addition resulted in a slight enhancement. However, the final experiment detailed in Table 7 demonstrated that replacing an LSTM with a BiLSTM, utilizing the same configurations, didn't yield any notable improvement.

2) OPTIMIZATION MODELS

For the optimizer, we performed the second experiment testing different optimizers, the comparison is on execution time and computation costs. Fig. 6 & 7 show the results of using the Adam, SGD, and RMSprop optimizers. It can be seen from Fig. 6 that the lowest execution time (30.58) was achieved using SGD optimizer; however, the lowest validation loss (0.30) is obtained by Adam optimizer as shown in Fig. 7. Adam is regarded as the most effective optimizer for efficiently training neural networks in less validation loss and acceptable execution time; nevertheless, RMSprop is the worst.

3) ADAM OPTIMIZER WITH DIFFERENT HIDDEN LAYER SIZES

It is worth noting the importance of considering both the dataset size and available computational resources when

TABLE 5. Results of bystander roles detection model.

Model	Training Inputs	W-Precision	W-Recall	W-F1_Score
fine-tuning BERT	• tweet text	.68	.70	.66
fine-tuning BERT	• reply_id • tweet text	.768	.77	.75
fine-tuning BERT	• tweet text • user screen-name	.80	.74	.73
fine-tuning BERT	• reply_id • tweet text • user screen-name	**.82**	**.83**	**.83**
freezing-BERT	• reply_id • tweet text • user screenname	.71	.70	.69
fine-tuning RoBERTa	• reply_id • tweet text • user screen-name	.78	.78	.76

TABLE 6. Parameter settings.

Parameter	Value or Description
input dimension size	10,000
output dimension size	128
hidden units size	64
optimizer	adam
loss function	'sparse_categorical_crossentropy'
epochs#	20
batch size	32
activation function	Default: hyperbolic tangent (tanh)
output layer activation	softmax

TABLE 7. Running different experiences to find the optimal.

#Epochs	W-Precision	W-Recall	W-F1	Test loss	Val loss
20	.85	.87	.86	.44	.43
30	.85	.87	.86	.54	.41
50	.86	.85	.84	.81	.62
100	.83	.85	.84	.69	.52
Output of implementing: dropout rate(DR)=0.5, BiLSTM, early stopping(ES).					
DR to LSTM	.86	.90	.88	.23	.18
DR to LSTM + ES	.87	.90	.89	.22	.17
DR to BiLSTM + ES	.85	.90	.87	.23	.18

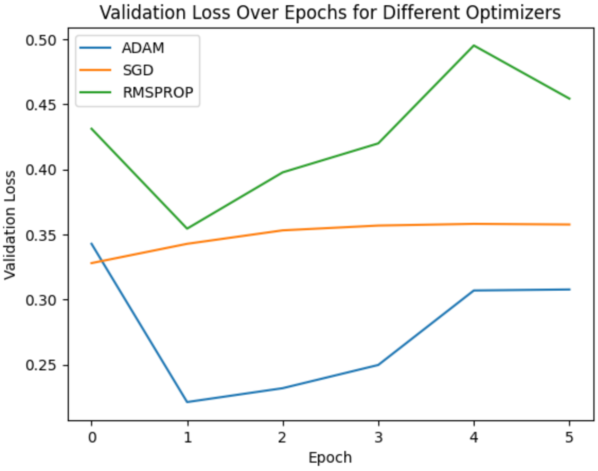


FIGURE 7. The performance of different Optimization models(2).

TABLE 8. The performance over different unit sizes.

Time (sec.)	units size	Computation cost (loss) Test Loss	Val loss	W-Precision	W-Recall	W-F1
36.40	32	.26	.21	.85	.89	.87
47.21	**64**	.23	.18	.86	.91	.89
92.83	128	.26	.20	.84	.85	.84

TABLE 9. Comparison the performance of the proposed model with and without bystander roles.

Approach	W-Precision	W-Recall	W-F1_Score	Test Loss
proposed	**.87**	**.90**	**.89**	**.22**
baseline	.72	.78	.71	.59

determining the appropriate size of units. We conducted the third experiment to test the performance of different unit sizes within a range suitable for our dataset. The performance is evaluated in terms of weighted metrics, execution time and computation cost. Table 8 illustrates the use of 32-, 64-, and 128-unit sizes respectively. Based on the findings, using 64-unit size achieves superior results in the overall evaluation metrics. However, there is a trade-off between time consumption and accuracy with the 32-unit size.

C. EVALUATION

1) IMPLEMENTING THE PROPOSED MODEL ON BASELINE DATASET WITHOUT BYSTANDER ROLES INFORMATION

The main challenge in studying cyberbullying detection with bystander role information is the lack of a dataset and a standardized evaluation framework. This absence makes fair comparisons difficult. Our research in cyberbullying detection differs from previous studies as none of them explored bystander roles. Therefore, it wouldn't be fair to directly compare our model with past results, as existing datasets don't align with our focus on bystander roles.

To assess the performance, we use the CY23 dataset as baseline dataset. It contains individual tweets with fine-grained cyberbullying label, labeled without any references to bystander roles or threads. We then compared the performance of our proposed model with CYBY23 - provided in our previous research [12]- to its performance with the baseline dataset. As shown in Table 9, we see that the model we propose, which takes into account the roles of bystanders, performs better in terms of accuracy and efficiency than the model run on the baseline dataset. In short, the results of our proposed approach, which takes into account the roles of bystander, are more significant in terms of all weighted metrics and loss. Bystander roles contribute significantly to the advancement of cyberbullying detection. To further verify the significance of the better performance achieved by our method considering bystander roles as a feature, paired confidence intervals have been considered. CI are calculated

TABLE 10. Finding reliability using CI.

Approach	CI
Proposed	**95% CI [0.78, 0.97]**
baseline	95% CI [0.66, 0.75]

at the 95% confidence level represent the lower and upper confidence limits around the mean that reported based on computing t-distribution. Table 10 shows results using CI metric for each model. The noteworthy score is that our proposed model has produced the 95.0% CI estimates that weighted F1-Score lies between 0.78 and 0.97.

2) COMPARISON WITH THE STATE-OF-THE-ART APPROACHES

As this is the first introduction of a session (thread)-based model for detecting cyberbullying with features associated with bystander roles, there is no standardized session (thread)-based dataset with the bystander roles feature and there are no studies using this feature in the field of cyberbullying detection. It is challenging to establish a standardized evaluation and conduct an experimental comparison with the performance of previous cyberbullying detection models. In other words, we could not implement our model on the available datasets because our chain model is based on a two-layer classification: Bystander roles, fine-grained cyberbullying. Therefore, we compared the performance of our model with the results reported in the literature for the most similar state-of-the-art models, as shown in Table 11.

In [19], the annotation process is session-based, where annotators are provided with the entire thread and user information to effectively label the five critical cyberbullying criteria. Additionally, five logistic regression classifiers are trained independently, each corresponding to one of the cyberbullying criteria. This process is a key factor contributing to high computation and time-consuming costs, as well as the loss of critical correlation between the dependent cyberbullying attributes. Consequently, the reported poor model detection accuracies, ranging from 17.5% to 77.9% F1-score, do not accurately represent the model's significant ability to distinguish cyberbullying from other related behaviors, such as aggression or crude joking. Previously proposed methods demonstrated poor performance in handling unbalanced data, as they are not explicitly designed to address the naturally unbalanced representation of real-world data [18]. This is evident in the reported detection accuracy of the 'serious' severity class. Although the training process involved balancing the three fine-grained cyberbullying classes, the 'serious' severity in the validation and test sets was lower than the other two classes. Consequently, their proposed Hierarchical Squashing Attention Network (HSAN) model exhibits poor performance in predicting the 'serious' severity class. It is the only study comparable to our proposed method in detecting the severity of cyberbullying instances. However, they neglect to consider the influence of bystander roles feature on the severity of cyberbullying events.

The highest detection score achieved by [20] using transformer-based models is lower than that of our proposed approach, primarily due to limitations in the distribution of bystander roles. Their dataset includes only two types of bystanders: 'harasser assistants' and 'defenders.' Moreover, 'harasser assistants' are merged with 'harassers' due to their very low frequency. This limitation arises from the platform used to collect their corpus, specifically the ASKfm platform, which consists only of question-answer pairs, limiting the availability of continuous interaction conversations necessary to capture different types of bystander roles.

Another study by Ge et al. [16] have contributed to session-based cyberbullying detection research with their proposed TGBully, a temporal graph-based cyberbullying detection framework that emphasizes user interactions. However, the detection accuracy is lower than that of our proposed approach due to certain limitations. The primary shortcoming lies in the substantial effort, both in computational and time costs, required to capture the repetitive characteristics of cyberbullying behavior. This process involves linking users' interactions with their historical comments to capture language behavior and personality. In reality, cyberbullying is unlike traditional forms of bullying, where repeated actions are often a defining criterion. Some authors argue that online abuse, whether it be a video, comment, or picture, can persist permanently in the public domain and remain accessible without being repeated [39].

In recent research on binary detection of cyberbullying, Long Short-Term Memory (LSTM) networks have been implemented [13]. Feature extraction methods were used in the study, including Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), which focus on individual words to capture their frequency. While these methods are effective in capturing word frequency, they fall short in capturing contextual or semantic information. This limits their ability to understand the exact meaning of words. Therefore, the reported recognition accuracy is lower than that achieved by our proposed approach..

In summary, the results obtained by our proposed approach are significant and valuable for advancing the research on cyberbullying detection. The discussion and comparison with earlier efforts with the same aim are limited due to the different implemented settings.

D. THEORETICAL AND PRACTICAL IMPLICATIONS OF RESEARCH

From a **theoretical** perspective, there are several social and educational studies that argue for the influence of "bystanders" on cyberbullying, as discussed in Section III. However, most existing research in the field of cyberbullying detection has primarily focused on binary classifications, often neglecting the role of bystanders and the broader conversational context. This is due to the limitation of available session (thread)-based datasets with the bystander roles feature and the challenges associated with the tedious

TABLE 11. Comparison with the state of the art models.

study	Dataset	F1-score	Precision	approach
[19]	Twitter dataset 6897 tweets 80% training 20% testing 5-cross validation	77.9	78.7	✗participants' actors (bully& victim) ✗bystander roles ✗Fine grained bullying detection ✓session based
[18]	Gossip Chinese dataset 5000 dialogues 3500 for training 1000 for testing 500 for validation	61.76	63.56	✗participants' actors (bully& victim) ✗bystander roles ✓Fine grained bullying detection ✓session based
[20]	ASKfm dataset 5375 posts 90% training 10% testing 5-cross validation	60.04	60.62	✓participants' actors (bully& victim) ✓bystander roles ✗Fine grained bullying detection ✗session based
[16]	Instagram dataset 2218 sessions 80% training 10% testing 10% validation	80.97	-	✗participants' actors (bully& victim) ✗bystander roles ✗Fine grained bullying detection ✓session based
[13]	not indicated	-	69.99	✗participants' actors (bully& victim) ✗bystander roles ✗Fine grained bullying detection ✗session based
proposed	Twitter dataset 112 threads- 640 tweets 60% training 20% testing 20% validation	89	87.9	✗participants' actors (bully& victim) ✓bystander roles ✓Fine grained bullying detection ✓session based

labeling of such datasets collected by threads (thread: initial post + associated replies), which requires a lot of time and human effort.

The proposed models would serve as a foundation for future studies examining the impact of involving bystander roles identification and has several **practical** implications for fine-grained detection models. The key advancements include: (i) Multi-label Detection: Instead of single-label detection, the proposed model supports multi-label detection. (ii)Capturing Label Dependency: The model captures label dependency through chained sequential classification, enhancing the understanding of relationships between labels. (iii)Upgraded Chain Classifier: The research presents an upgraded version of the built-in chain classifier in scikit-multilearn, that is primarily designed for traditional machine learning classifiers.

The fundamental and crucial step is sequence classification with deep learning. Specifically, the prediction of the first model is appended to the training set of the next model in the sequence, creating a continuous learning process.

Bystander roles and other features, such as cyberbullying criteria, have recently received much attention in cyberbullying detection research. However, their explicit dependence on final detection has not been captured through distinct training and classification layers. In the proposed approach, separate training and classification layers are performed, and the results of each layer are used as a feature for the next detection layer, resulting in a classification chain with updates to the training set. This improvement in fine-grained cyberbullying detection aims to advance recent research efforts by significantly capturing the dependency between features.

VII. CONCLUSION

Bystander identification is crucial for enhancing the detection of cyberbullying. This paper aims to identify and evaluate the impact of identifying bystander roles in developing fine-grained cyberbullying detection. The paper presents a reliable classifier chain that identifies bystander roles to detect the severity of cyberbullying and distinguish it from instances of aggression. Furthermore, the proposed chain model, consisting of two layers of classification, demonstrates improved performance with an f1-score of 89%. The results show that fine-tuning BERT significantly enhances performance in the task of bystander roles classification compared to another alternative pre-trained model. We experimented with various regularization and optimization techniques to enhance the LSTM; however, this resulted in a cost/accuracy trade-off, yielding only marginal improvement while slightly increasing computational costs.

VIII. RESEARCH LIMITATIONS AND FUTURE DIRECTIONS

While the proposed findings point to promising directions for future work to recognize that cyberbullying detection can be improved by incorporating the surrounding conversation through bystander roles, this study has limitations which may impact the results.

In the chained classification model, misclassifications in bystander roles in the first layer might lead to incorrect contextual information being passed to the cyberbullying detection layer, resulting in reduced accuracy and reliability of the overall model. Most earlier studies address this issue by determining the optimal order of the chain. However, our study maintains a fixed order because our research is grounded in theoretical findings indicating that bystander replies significantly impact the dynamics of cyberbullying. Therefore, this study focuses solely on fine-tuning the pre-trained BERT model to leverage knowledge from large datasets, thereby reducing the potential for cascading errors. All experimental stages of the proposed methods in this study were limited to CYBY23 dataset as there are no other datasets available taking into consideration the roles of bystanders in cyberbullying threads. The present research is implemented on Twitter dataset. However, it could be generalized in the future to session-based datasets with bystanders roles and cyberbullying classes from other social media platforms such as Facebook and YouTube to investigate whether the same pattern of cyberbullying severity is observed. This research does not extend to specifically identifying participants as bullies or victims. Instead, it focuses on identifying the roles of bystanders: 'instigators', 'defenders', those who remain 'impartial', and any others who do not fit into these categories are tagged as 'others'.

In the future, we aim to address the issue of error propagation using mitigation mechanisms such as ensemble methods, confidence thresholds, uncertainty estimation, and feedback loops. Also, we intend to collect a more extensive expert-labeled corpus by collaborating with experts in the field of cyberbullying. Qualified experts can provide accurate annotations for a large dataset to further improve the performance of the proposed model. The future goal is also to improve the model for handling imbalanced classes to account for the representation of real-world data that is inherently imbalanced. Furthermore, we would want to balance the dataset through increasing the size of the lowest classes or through using data augmentation techniques and repeating our previous experiments.

REFERENCES

[1] E. A. Vogels. (2021). *The State of Online Harassment Pew Research Center*. [Online]. Available: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

[2] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," *Psychol. Bull.*, vol. 140, no. 4, pp. 1073–1137, 2014.

[3] B. Iranzo, S. Buelga, M.-J. Cava, and J. Ortega-Barón, "Cyberbullying, psychosocial adjustment, and suicidal ideation in adolescence," *Psychosocial Intervent*, vol. 28, no. 2, pp. 75–81, 2019.

[4] S. C. Hunter, J. M. Boyle, and D. Warden, "Perceptions and correlates of peer-victimization and bullying," *Brit. J. Educ. Psychol.*, vol. 77, no. 4, pp. 797–810, 2007.

[5] R. Kowalski, S. Limber, S. Limber, and P. Agatston, *Cyberbullying: Bullying in the Digital Age*. Hoboken, NJ, USA: Wiley, 2012.

[6] J. W. Patchin and S. Hinduja, *Cyberbullying Prevention and Response: Expert Perspectives*, 1st ed., Evanston, IL, USA: Routledge, 2011, doi: 10.4324/9780203818312.

[7] S. Bastiaensens, H. Vandebosch, K. Poels, K. Van Cleemput, A. DeSmet, and I. De Bourdeaudhuij, "Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully," *Comput. Hum. Behav.*, vol. 31, pp. 259–271, Feb. 2014, doi: 10.1016/j.chb.2013.10.036.

[8] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Arch. Suicide Res.*, vol. 14, no. 3, pp. 206–221, Jul. 2010.

[9] M. Duggan. (2014). *Online Harassment Pew Research Center*. [Online]. Available: https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/

[10] N. K. Mishra and P. K. Singh, "Linear ordering problem based classifier chain using genetic algorithm for multi-label classification," *Appl. Soft Comput.*, vol. 117, Mar. 2022, Art. no. 108395.

[11] J. Read, B. Pfahringer, G. Holmes, E. Frank, D. Xin, S. Takamichi, H. Saruwatari, W. Liu, and I. W. Tsan, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, pp. 333–359, 2011.

[12] H. S. Alfurayj, N. S. Yee, and S. L. Lutfi, "Bystanders unveiled: Introducing a comprehensive cyberbullying corpus with bystander information," in *Proc. IEEE Region 10 Conf.*, Oct. 2023, pp. 1012–1017.

[13] A. D. Ma and D. K. Daniel, "Cyberbullying detection on social networks using LSTM model," in *Proc. Int. Conf. Innov. Sci. Technol. Sustain. Develop. (ICISTSD)*, Aug. 2022, pp. 293–296.

[14] T. Li, Z. Zeng, Q. Li, and S. Sun, "Integrating GIN-based multimodal feature transformation and multi-feature combination voting for irony-aware cyberbullying detection," *Inf. Process. Manage.*, vol. 61, no. 3, May 2024, Art. no. 103651, doi: 10.1016/j.ipm.2024.103651.

[15] F. Elsafoury, S. Katsigiannis, S. R. Wilson, and N. Ramzan, "Does BERT pay attention to cyberbullying?" in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1900–1904.

[16] S. Ge, L. Cheng, and H. Liu, "Improving cyberbullying detection with user interaction," in *Proc. Web Conf.*, Apr. 2021, pp. 496–506.

[17] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, and S. Mishra, "Identifying differentiating factors for cyberbullying in vine and Instagram," in *Proc. Annu. Int. Conf. Inf. Manage. Big Data*, vol. 1410, 2020, pp. 348–361.

[18] J.-L. Wu and C.-Y. Tang, "Classifying the severity of cyberbullying incidents by using a hierarchical squashing-attention network," *Appl. Sci.*, vol. 12, no. 7, p. 3502, Mar. 2022.

[19] C. Ziems, Y. Vigfusson, and F. Morstatter, "Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification," in *Proc. 14th Int. AAAI Conf. Web Social Media*, 2020, pp. 808–819.

[20] 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.

[21] R. Mahieu, "'We're not coming from mars; we know how things work in morocco!' how diasporic Moroccan youth resists political socialisation in state-led homeland tours," *J. Ethnic Migration Stud.*, vol. 45, no. 4, pp. 674–691, Mar. 2019, doi: 10.1080/1369183x.2017.1409177.

[22] H. Machackova, "Bystander reactions to cyberbullying and cyberaggression: Individual, contextual, and social factors," *Current Opinion Psychol.*, vol. 36, pp. 130–134, Dec. 2020, doi: 10.1016/j.copsyc.2020.06.003.

[23] Y. Zhao, X. Chu, and K. Rong, "Cyberbullying experience and bystander behavior in cyberbullying incidents: The serial mediating roles of perceived incident severity and empathy," *Comput. Hum. Behav.*, vol. 138, Jan. 2023, Art. no. 107484, doi: 10.1016/j.chb.2022.107484.

[24] H. S. Alfurayj and S. L. Lutfi, "Exploring Bystanders' roles in labeled cyberbullying threads on Twitter: A preliminary analysis," in *Proc. IEEE Region 10 Conf.*, Oct. 2023, pp. 1018–1023.

[25] C. Olah. (2015). *Understanding LSTM Networks*. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[26] T. Komatsu, S. Watanabe, K. Miyazaki, and T. Hayashi, "Acoustic event detection with classifier chains," in *Proc. Interspeech*, Aug. 2021, pp. 46–50.

[27] J. Li, X. Zhu, and J. Wang, "AdaBoost. C2: Boosting classifiers chains for multi-label classification," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 8580–8587.

[28] W. Liu and I. W. Tsang, "On the optimality of classifier chain for multi-label classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 712–720.

[29] D. Xin, S. Takamichi, and H. Saruwatari, "Exploring the effectiveness of self-supervised learning and classifier chains in emotion recognition of nonverbal vocalizations," in *Proc. ICML Exvo. Workshop*, 2022, pp. 1–19.

[30] M. Behzadi, I. G. Harris, and A. Derakhshan, "Rapid cyber-bullying detection method using compact BERT models," in *Proc. IEEE 15th Int. Conf. Semantic Comput. (ICSC)*, Jan. 2021, pp. 199–202.

[31] J. Briskilal and C. N. Subalalitha, "An ensemble model for classifying idioms and literal texts using BERT and RoBERTa," *Inf. Process. Manage.*, vol. 59, no. 1, Jan. 2022, Art. no. 102756, doi: 10.1016/j.ipm.2021.102756.

[32] 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.

[33] C. Graney-Ward, B. Issac, L. Ketsbaia, and S. M. Jacob, "Detection of cyberbullying through BERT and weighted ensemble of classifiers," *TechRxiv*, Jan. 2022, doi: 10.36227/techrxiv.17705009.v1.

[34] K. B. Nelatoori and H. B. Kommanti, "Multi-task learning for toxic comment classification and rationale extraction," *J. Intell. Inf. Syst.*, vol. 60, no. 2, pp. 495–519, Apr. 2023.

[35] H. Almerekhi, H. Kwak, J. Salminen, and B. J. Jansen, "Are these comments triggering? Predicting triggers of toxicity in online discussions," in *Proc. Web Conf. 2020*, Jul. 2020, pp. 3033–3040.

[36] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manage.*, vol. 50, no. 1, pp. 104–112, Jan. 2014, doi: 10.1016/j.ipm.2013.08.006.

[37] R. A. D. Oliveira and M. H. J. Bollen, "Deep learning for power quality," *Electric Power Syst. Res.*, vol. 214, Jan. 2023, Art. no. 108887, doi: 10.1016/j.epsr.2022.108887.

[38] Z. Zhao, Z. Zhang, and F. Hopfgartner, "A comparative study of using pre-trained language models for toxic comment classification," in *Companion Proc. Web Conf.*, Apr. 2021, pp. 500–507.

[39] R. Slonje, P. K. Smith, and A. Frisén, "The nature of cyberbullying, and strategies for prevention," *Comput. Hum. Behav.*, vol. 29, no. 1, pp. 26–32, Jan. 2013, doi: 10.1016/j.chb.2012.05.024.

HAIFA SALEH ALFURAYJ received the master's degree from Old Dominion University, USA. She is currently pursuing the Ph.D. degree with Universiti Sains Malaysia. She was a Lecturer with the School of Computer Sciences, Qassim University, Saudi Arabia. Her research interests include machine learning, big data analytics, sentiment analysis, and natural language processing (NLP). Her current research focuses on leveraging AI techniques to enhance cyberbullying detection by investigating causal factors, including the influence of bystanders.

SYAHEERAH LEBAI LUTFI (Member, IEEE) received the master's degree in software engineering from Universiti Malaya, and the DEA and Ph.D. degrees from the Speech Technology Group (Grupo de Technología del Habla-THAU), Universidad Politécnica de Madrid, Spain, in June 2013. She is currently a tenured academic at the School of Computer Sciences, Universiti Sains Malaysia (USM). Her research interests include human-computer interactions, with a particular focus on affective computing, behavior analytics, personality, culture, and mood and emotion analysis and modeling. She has made significant contributions to her field, with over 60 papers published in reputable journals and conferences. Since 2015, she has been serving as the EU Erasmus+ Coordinator for the collaboration between Universidad Politécnica de Madrid and USM. Additionally, she has held the position of Deputy Director of Global Engagement at the International Mobility and Collaboration Center (IMCC), USM, since 2019.

RAMESH PERUMAL (Member, IEEE) was born in Tamil Nadu, India, in January 1986. He received the Bachelor of Engineering (B.E.) degree in electronics and communication engineering from Anna University, India, in 2007, the Master of Science (M.S.) degree in embedded system design from Manipal University, in 2011, India, the Master of Engineering (M.E.) degree in communication systems from Anna University, in 2013, and the Ph.D. degree in electrical engineering with a specialization in neuroengineering from National Tsing Hua University, Taiwan, in 2021. He is currently a Software Enabling and a Optimization Engineer with Intel Microelectronics (M) Sdn., Bhd., Penang, Malaysia. His research interests include biosignal processing, neuroinformatics, predictive modeling, Kalman filter, adaptive deep brain stimulation, hardware-accelerated computing, machine learning, model optimization, computer vision, and edge computing. With a decade of professional experience in software development and edge AI applications.