Modeling Temporal Patterns of Cyberbullying Detection with Hierarchical Attention Networks

LU CHENG, RUOCHENG GUO, YASIN N. SILVA, DEBORAH HALL, and HUAN LIU,
Arizona State University, USA

Cyberbullying is rapidly becoming one of the most serious online risks for adolescents. This has motivated work on machine learning methods to automate the process of cyberbullying detection, which have so far mostly viewed cyberbullying as one-off incidents that occur at a single point in time. Comparatively less is known about how cyberbullying behavior occurs and evolves over time. This oversight highlights a crucial open challenge for cyberbullying-related research, given that cyberbullying is typically defined as intentional acts of aggression via electronic communication that occur *repeatedly* and *persistently*. In this article, we center our discussion on the challenge of modeling temporal patterns of cyberbullying behavior. Specifically, we investigate how temporal information within a social media session, which has an inherently hierarchical structure (e.g., words form a comment and comments form a session), can be leveraged to facilitate cyberbullying detection. Recent findings from interdisciplinary research suggest that the temporal characteristics of bullying sessions differ from those of non-bullying sessions and that the temporal information from users' comments can improve cyberbullying detection. The proposed framework consists of three distinctive features: (1) a hierarchical structure that reflects how a social media session is formed in a bottom-up manner; (2) attention mechanisms applied at the word- and comment-level to differentiate the contributions of words and comments to the representation of a social media session; and (3) the incorporation of temporal features in modeling cyberbullying behavior at the comment-level. Quantitative and qualitative evaluations are conducted on a real-world dataset collected from Instagram, the social networking site with the highest percentage of users reporting cyberbullying experiences. Results from empirical evaluations show the significance of the proposed methods, which are tailored to capture temporal patterns of cyberbullying detection.

CCS Concepts: • **Security and privacy → Social aspects of security and privacy**; • **Applied computing** → *Sociology*;

Additional Key Words and Phrases: Cyberbullying, temporal analysis, hierarchical attention network, social media

ACM Reference format:
Lu Cheng, Ruocheng Guo, Yasin N. Silva, Deborah Hall, and Huan Liu. 2021. Modeling Temporal Patterns of Cyberbullying Detection with Hierarchical Attention Networks. *ACM/IMS Trans. Data Sci.* 2, 2, Article 8 (March 2021), 23 pages.
https://doi.org/10.1145/3441141

This material is based upon work supported by the National Science Foundation (NSF) Grants 1719722 and 2036127.
Authors' address: Lu Cheng, R. Guo, and H. Liu, School of Computing Science and Engineering, Ira A. Fulton School of Engineering, Arizona State University, Tempe, AZ, USA; emails: {lcheng35, rguo12, huanliu}@asu.edu; Y. N. Silva, School of Mathematical and Natural Sciences, New College of Interdisciplinary Arts and Sciences, Arizona State University, Glendale, AZ, USA; email: ysilva@asu.edu; D. Hall, School of Social and Behavioral Sciences, New College of Interdisciplinary Arts and Sciences, Arizona State University, Glendale, AZ, USA; email: d.hall@asu.edu.

2577-3224/2021/03-ART8 $15.00
https://doi.org/10.1145/3441141

1 INTRODUCTION

Cyberbullying is frequently defined as intentional acts of aggression carried out by a group or individual using electronic communication, *repeatedly over time*, against victims who cannot easily defend themselves [7, 44, 45]. A distinct aspect of this definition is the *persistence* and *repetition* of the aggressive acts. Notwithstanding the promising results, most of the established work (e.g., References [8, 9, 50]) has overlooked this key aspect of cyberbullying [22]. Comparatively fewer efforts (e.g., References [4, 53]) have been directed towards exploring the repetitive feature of cyberbullying behavior. Recent burst analysis of cyberbullying activity from interdisciplinary research bridging computer science and psychology [19] reveal noteworthy differences between the temporal characteristics of bullying versus non-bullying sessions on Instagram.[1] This suggests that cyberbullying detection in social media sessions—which typically consist of an initial post, a sequence of time-stamped comments, images/videos, and other social content such as the number of likes and shares—may benefit from models that take temporal properties of the cyberbullying behavior into account. Figure 1 displays an Instagram session where cyberbullying behavior repetitively occurs in multiple comments.

To model temporal patterns, a straightforward approach is to extract temporal features (e.g., timestamp describing when a comment is posted) and directly concatenate them with other features such as text (e.g., Bag of Words). However, given the multi-modal nature of social media data, features gleaned from social media sessions often follow different distributions and may not be compatible with each other and, in the worst case, may be independent [9]. This highlights a primary challenge of using temporal analyses in cyberbullying detection: How to effectively integrate temporal information with other features to improve model performance?

Social media sessions have an inherently *hierarchical structure*, e.g., words form a comment, comments and social content form a session, as illustrated in Figure 1. A number of studies in document classification (e.g., References [12, 52]) have shown that document representation can be improved by considering its hierarchical structure. We draw on these findings to model the hierarchy of a social media session, which can enhance the representation of a social media session in crucial ways. First, comments within a social media session can be short, and the lack of contextual information can present challenges for detecting cyberbullying sessions when comments are used independently [55]. The hierarchical structure can enrich individual comments with semantically related texts. Second, words in a comment and comments in a session are not equally relevant to cyberbullying detection. For example, whereas *"You're a f**king loser!"* and *"Yeah, I'm a loser"* both include the word *loser*, the former is more likely to indicate an instance of bullying. When constructing the hierarchical structure, we can naturally distinguish the importance of words and comments at different levels of a social media session. Moreover, studying these structural properties enables us to exploit information at different granularity levels such as textual information (e.g., tokens) at the word level, temporal information (e.g., timestamps) at the comment-level, and social content (e.g., number of likes) at the session level. In this work, we thus focus on how modeling the hierarchical structure of social media sessions renders more effective use of temporal information and improves the performance of cyberbullying detection.

In our previous work [7], we studied whether the hierarchical structure of a social media session is applicable to cyberbullying detection. We viewed temporal information as a supervised signal and utilized a multi-task learning framework (referred to as Hierarchical Attention Networks for Cyberbullying Detection, i.e., HANCD) to jointly detect cyberbullying instances and predict time. In this article, we expand the scope of Reference [7] and seek a unified way to incorporate the

[1] https://www.instagram.com/.

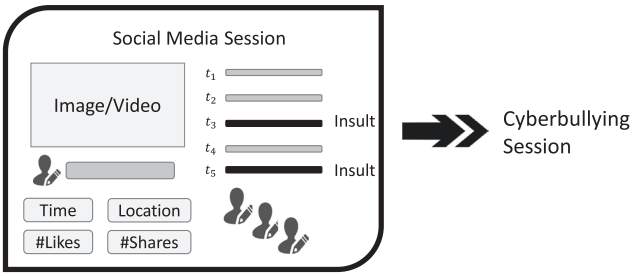


Fig. 1. A social media session may consist of an image/video, social content, and a sequence of timestamped comments that include a sequence of words. Cyberbullying is reflected in the repetitive comments labeled as "Insult," which could be posted by multiple users. The goal of cyberbullying detection is to predict whether a given social media session represents cyberbullying or not.

temporal information. We investigate how the temporal orderings of user-generated data can be utilized to compensate for the repetition of cyberbullying behavior. The new contributions include:

- Informed by the burst analysis results in Reference [19] and discussions of temporal features in cyberbullying detection from Reference [45], we build new temporal modules upon our findings from Reference [7] illuminating the hierarchical construction of a social media session for cyberbullying detection. We deploy the extracted temporal features and propose a new approach, HANT to model temporal dynamics in a consolidated manner. HANT explicitly encodes the temporal orderings of user-generated comments in a social media session.
- We perform thorough analyses and conduct extensive experiments[2] to examine the performance of HANCD and HANT. This includes (1) evaluation metrics and significance test for classification with imbalanced datasets, (2) examinations of the learned latent session representations, (3) case studies with visualized attention levels, and (4) sensitivity analyses of hyper-parameters. Additionally, to better understand the relationship between data distribution and different model architectures, we employ a widely recognized data augmentation strategy and discuss the effects of using balanced datasets on model performance.
- We significantly improve the previous paper by incorporating (1) a detailed motivation of the proposed techniques, (2) a formal problem definition and a comprehensive review of related work, (3) a discussion of how our approach is supported by recent exploratory analyses using a new cyberbullying dataset with comment-level cyberbullying labels [19], and (4) a detailed description of the presented algorithms and discussion of experimental results based on a more rigorous experimental design.

The remainder of this article is organized as follows: We review the related work on cyberbullying detection and deep learning for text classification in Section 2, and we formally define the problem of session-based cyberbullying detection in Section 3. In Section 4, we first review how to use hierarchical attention networks to construct a social media session and then introduce the details of the proposed temporal encoding for modeling the temporal characteristics of cyberbullying behavior. In Section 5, we present the quantitative and qualitative evaluations of the proposed approach. We summarize our findings and discuss directions for future work in Section 6.

[2]Code is available at https://github.com/GitHubLuCheng/Modeling-Temporal-Patterns-of-Cyberbullying-Detection.

2 RELATED WORK

In this section, we survey the related literature in two broad areas, cyberbullying detection and deep learning models for text classification. We first investigate four categories of features widely used in cyberbullying detection: content-, sentiment-, user-, and network-based features. We then review established cyberbullying detection approaches that consider repetition in the process of model development. Because our model is based on deep neural networks that have achieved state-of-the-art results on a suite of standard natural language processing (NLP) tasks, we also examine common deep learning models aimed at text classification.

2.1 Cyberbullying Detection

To date, numerous machine learning algorithms have been proposed to identify cyberbullying instances using various features, such as content, sentiment, user, and social network information. Content-based features are common in the related literature [41], e.g., cyberbullying keywords [50], profanity [17, 31], Bag of Words (BoW) [50], n-gram [54], Term Frequency Inverse Document Frequency (TFIDF) [17, 37, 54], and Linguistic Inquiry Word Count (LIWC) [6, 8]. For example, studies such as Reference [17] created profanity lexicons using word lists collected by the researchers. Xu et al. [50] introduced several NLP-based tools (e.g., BoW, Latent Semantic Analysis- and Latent Dirichlet Allocation-based representation learning) to study bullying traces extracted via the Twitter Streaming API.[3] Cheng et al. [8] proposed to personalize cyberbullying detection by considering users' unique personality traits and peer influence inferred from users' language behavior, i.e., LIWC. In addition to improving predictive accuracy, Cheng et al. [6] further sought to identify potential causes of cyberbullying among LIWC features through a causally interpretable model.

Sentiment or emotion analysis has been used to detect sentiments on social media. When applied to cyberbullying detection, sentiment-based features are often combined with content features like TFIDF to improve the performance. For instance, in [14], the authors proposed the Sentiment-Informed Cyberbullying Detection (SICD) model, which incorporates sentiment analysis into content features. Experimental results showed that capturing the sentiment consistency of bullying and non-bullying posts can improve model accuracy. By reviewing the extracted tweets, Xu et al. [50] detected seven different types of emotions in the tweets, including anger, empathy, and fear. Most models using sentiment as features rely on a lexicon of emotive words to detect the polarity (negative, positive, or neutral) of sentiments. An exception is the work by Nahar et al. [38], which leveraged Probabilistic Latent Semantic Analysis to extract sentiment features. Another common type of feature is based on users' characteristics including age, gender, sexual orientation, and race [41]. Dadvar et al. [13] studied a gender-specific corpus of MySpace[4] posts to train an SVM classifier using the TFIDF of profane words and pronouns as features. Besides the features extracted from text, the growing prevalence of online social networking systems has also provided researchers with network-based features such as the number of friends, network embeddedness, and relational centrality [46] in cyberbullying detection. For instance, in Reference [9], the authors leveraged the multi-modal information in an Instagram session such as social network features (i.e., following and followed-by relationships) to construct a heterogeneous network for all social media sessions. Huang et al. [23] extracted a set of features from the constructed ego networks to improve detection performance.

Yet, comparatively less is known about how to model key aspects of cyberbullying behavior, such as repetition, using computational approaches. Soni and Singh [45] proposed a

[3] https://developer.twitter.com/en.
[4] http://www.myspace.com/.

computational method to model the dynamic commenting behavior as point processes and extracted several temporal features for distinguishing cyberbullying from non-bullying social media sessions. More recently, Gupta et al. [19] presented key temporal characteristics of cyberbullying and trends obtained from descriptive and burst analysis of 100 social media sessions with comment-level labels. To achieve more timely and scalable detection using a repetitive process, Yao et al. [53] proposed a sequential hypothesis-testing formulation that aims to reduce the number of features while maintaining high classification accuracy. In Reference [4], the authors sought to predict the number of harassing comments a social media session will receive over a period of time. They formulated this problem as a regularized multi-task regression to study the evolution of cyberbullying behavior using historical data. Finally, our previous work [7] sought to model social media sessions in a hierarchical manner. Temporal information, such as the time intervals between two continuous comments, was extracted and used at the comment-level.

Whereas these efforts represent important initial steps towards understanding temporal aspects of cyberbullying, most of them have not explicitly examined the connections between temporal information and other features. Additionally, most of these approaches (e.g., References [4, 53]) rely on comment-level cyberbullying labels. These labels are particularly difficult to acquire, because (i) all of the contextual information (e.g., historical comments) needs to be carefully examined; and (ii) there may be a large number of comments within each social media session. A critical question that remains is how to jointly model the temporal patterns and other available information in social media sessions to optimize their contributions to cyberbullying detection without access to comment-level labels. We propose to model the hierarchical structure of a social media session to facilitate the temporal analysis at the comment-level.

2.2 Deep Learning for Text Classification

Deep learning models have been successfully applied in text classification in part because they can automatically extract context-sensitive features from raw text [33] and, therefore, largely overcome the drawbacks of conventional machine learning models that extract hand-crafted features from documents. A simple but effective type of deep learning model for text representation is the feed-forward network. These models take the BoW as input and learn a vector representation using an embedding model for each word. The text representation is the sum or average of the embeddings, which is then passed through one or more feed-forward layers, i.e., Multi-Layer Perceptrons. The final layer's representation is input into a classifier, such as logistic regression. An example of this approach is the Deep Average Network [24]. Recurrent Neural Network (RNN) [30] takes a sequence of words as input and aims to capture the dependencies among words. The most popular RNN architecture is Long Short-Term Memory (LSTM), which is designed to better capture long-term dependencies. For example, inspired by the syntactic properties of natural language that combine words into phrases, Tai et al. [47] proposed a tree-structured LSTM model to learn rich semantic representations. Because RNNs often struggle to remember long-range dependencies, Bieng et al. [16] proposed a TopicRNN model to combine the merits of RNNs and latent topic models (used to capture local and global dependencies, respectively). Whereas RNNs are trained to recognize patterns across time, Convolutional Neural Networks (CNNs) learn patterns across space [36]. One of the first CNN models for text classification—Dynamic CNN [27]—used dynamic k-max pooling to explicitly capture short- and long-range relations of words and phrases. Kim [29] later proposed a simplified CNN that used only one layer of convolution on top of word2vec [35]. They concluded that CNN can better capture text semantics, because the max-pooling layer helps identify discriminative phrases in a text. There have been efforts aimed at improving CNN-based models, such as the CNN with attention mechanism [2, 52], the Bow-CNN model [25, 26], and the combination of CNN and RNN (RCNN) [32]. Zhang et al. [56] proposed a character-level CNN that

Table 1. Primary Symbols

Notation	Definition/Description
C, N	A corpus of social media sessions, sample size of this corpus
$\int_i$, $\mathbf{c}_i$, $\mathbf{w}_{ij}$	Session i, comment i, word j in comment i
L_i, C	Number of words in comment i, number of comments in a session
D, t	Size of hidden layer, a hidden state
$\mathbf{x}_{it}$, $\mathbf{s}_{it}$, $\mathbf{h}_{it}$	Embeddings of words at different hidden layers
$\mathbf{x}_i$, $\mathbf{s}_i$, $\mathbf{h}_i$	Embeddings of comments at different hidden layers
$\mathbf{u}_w$, $\mathbf{u}_c$	Word-level context vector, comment-level context vector
α_{it}, α_i	Attention of a word, attention of a comment
Δt_i, $\mathbf{z}_i$, $\mathbf{m}$	Time interval between comment $i-1$ and i, social content of session i, session representation
β_1, β_2	Weights of cyberbullying detection and time interval prediction tasks

achieved competitive performance. In Reference [48], the authors first used CNN or LSTM to get the mappings of sentences, followed by a bidirectional gated RNN to obtain document representations. Lai et al. [32] proposed RCNN, which learns more precise text representations by taking advantage of both RNN and CNN.

Attention has become an increasingly popular term and useful tool in deep learning for NLP. It can be interpreted as a vector of importance weights that differentiate the contributions of words and sentences to the meanings of documents. For instance, Yang et al. [52] proposed a hierarchical attention network (HAN) for document classification. This model presents two distinctive properties: (1) a hierarchical structure that mirrors the structure of documents and (2) two levels of attention mechanisms applied at the word- and sentence-level. The weight of each word/sentence is learned automatically by imposing the attention mechanisms [2] to both word and sentence representations in the bidirectional GRU (Gated Recurrent Units). Their approach outperformed previous methods by a substantial margin on six text classification tasks. HAN was later extended to cross-lingual sentiment classification [57]. Shen et al. [43] introduced a directional self-attention network for RNN/CNN-free language understanding, where the attention between units in input sequence is directional and multi-dimensional. Other popular attention mechanisms include two-way attention (e.g., Attentive Pooling [42]) and co-attentive networks [28].

In contrast to documents, social media sessions include shorter, noisier, and more informal tokens. Yet, they also contain useful content in addition to text, such as temporal and social network information. These properties enable HANCD and HANT to leverage the *multi-modal* information in social media sessions to investigate key aspects of cyberbullying, such as repetition, and further improve the performance of cyberbullying detection.

3 PROBLEM STATEMENT

In this section, we first describe the primary notations used throughout this article, next introduce the proposed problem, and then briefly describe our approach built on a hierarchical attention network. At the end of this section, we highlight key challenges to efficiently solve the proposed problem.

3.1 Notations

Following commonly used notation in related work, we denote a scalar as a lowercase letter (e.g., a), a random variable as a uppercase letter (e.g., A), a vector as a boldface lowercase letter (e.g., $\mathbf{a}$), and a matrix as a boldface uppercase letter (e.g., $\mathbf{A}$). Subscripts indicate element indexes. Table 1 summarizes the primary symbols.

3.2 Session-based Cyberbullying Detection

Definition (Cyberbullying Detection within Social Media Sessions). Let C be a corpus of N social media sessions $C = \{\int_1, \int_2, \ldots, \int_N\}$, where each session includes the caption of the posted image/video, a sequence of timestamped comments $\{\mathbf{c}_1, \mathbf{c}_2, \ldots \mathbf{c}_C\}$, and social content such as the number of likes and shares. Each session owner is associated with features describing her/his number of followers and follows. The ith comment in a session is composed of L_i words $\{\mathbf{w}_{ij}\}$, i.e., $\mathbf{c}_i = \{\mathbf{w}_{i1}, \mathbf{w}_{i2}, \ldots, \mathbf{w}_{iL_i}\}$. Each session is labeled as either 1 denoting bullying session or 0 denoting non-bullying session. Let D be the dimension of the session representation. We define cyberbullying detection as learning a binary classifier $f : \mathbb{R}^D \rightarrow \{0, 1\}$ that leverages *textual* (e.g., comments), *structural* (e.g., hierarchical structure of a social media session), *temporal* (e.g., timestamp of a comment), and *social content* (e.g., number of likes) to identify if a social media session is an instance of cyberbullying [10].

In this work, we use the following information extracted from an Instagram session:

- *Text*: captions and subsequent comments $\{\mathbf{c}_i\}$ represented as BoW $\{\mathbf{w}_{ij}\}$, with the number of words limit set to 20,000 after removing stop words.
- *Time*: the timestamp t_i when a comment i was posted. It is used to calculate time intervals in HANCD and referred to as the temporal index in HANT.
- *Social Content*: a vector $\mathbf{z}_i$ describing the number of likes and shares a post has received.

3.3 Cyberbullying Detection via Hierarchical Attention Network

A social media session is inherently hierarchical and multi-modal: A comment/caption comprises a sequence of words and a session comprises a sequence of timestamped comments and key social content. To learn a high-quality session representation, rather than simply concatenating information from multiple modalities into a high-dimensional vector, our approach constructs a session in a bottom-up manner to explicitly model its hierarchical structure and incorporate the multi-modal information at different levels of the hierarchy. In addition, because the semantics of words and comments highly depend on the context—and even in the same context, the relevance of words and comments to cyberbullying is different—our approach also seeks to account for the ordering of words and comments, and measure the attention levels associated with individual words and comments. Central to this hierarchical attention network is the unique feature of our approach—it simultaneously exploits the temporal patterns of users posting comments to characterize the *repetition* of cyberbullying behavior. To this end, the proposed hierarchical attention network learns the session representations via feedback from both the labels (cyberbullying/non-bullying) of and the temporal information within social media sessions. We illustrate the hierarchical attention network for a cyberbullying social media session in Figure 2.

3.4 Challenges

- **Integration of Temporal Information.** A major challenge in the stated problem is how to effectively integrate the temporal information (i.e., the timestamps of the comments) into a cyberbullying detection framework. Directly concatenating temporal features with other features may not optimize the use of these temporal characteristics. It is crucial to model the evolving dynamics embedded within a social media session to achieve high accuracy of cyberbullying detection.
- **Scarcity of Comment-level Labels.** Modeling temporal patterns of cyberbullying behavior can benefit from labels that indicate if a comment involves cyberbullying or not [53]. To the best of our knowledge, however, there are few publicly available datasets [58] for cyberbullying detection that include comment-level labels in addition to the label for the entire

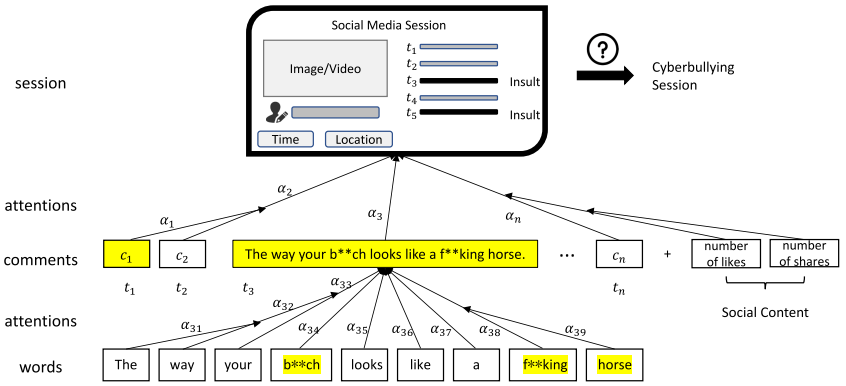


Fig. 2. Illustration of the hierarchical structure and attention mechanisms for cyberbullying detection in social media sessions. α_i denotes the attention associated to comment i; α_{ij} denotes the attention associated to word j in comment i.

social media session. The integration of more readily available comment-level information (e.g., timestamps) in detection models is, therefore, an open research challenge.

- **Social Media Data.** Compared to traditional news and documents, social media data are notorious for being a sea of noisy, short, and informal information. Social media sessions usually come in complex forms and exhibit considerable variation due to the multi-modal nature of the data [9]. These factors further complicate the process of gaining actionable knowledge from social media sessions.

4 THE PROPOSED FRAMEWORK

Both the proposed HANCD and HANT frameworks aim to capture and incorporate temporal patterns into cyberbullying detection models such that the capability of discriminating between bullying and non-bullying instances can be improved. At its core, our framework is built on the Hierarchical Attention Network (HAN) [52] that models a social media session in a bottom-up fashion. By capturing the hierarchical structure of sessions, our method can differentiate the words/comments importance and characterize the temporal patterns of users posting comments at different levels of the hierarchy, therefore, improving the representation learned for a social media session. Specifically, the components shared between HANCD and HANT include: a word and comment sequence encoder, a word- and comment-level attention layer, and a hidden layer to embed the social content. HANCD uses a time interval prediction component while HANT includes a temporal encoding component. In this section, we first describe in detail how HAN can be adapted to construct the structure of a social media session. Then, we introduce the two approaches for temporal pattern extraction that contribute to more accurate cyberbullying detection. Figure 3 illustrates the framework of HANCD and HANT.

4.1 Hierarchical Attention Network

The HAN framework proceeds from the input BoW towards the representation of an entire social media session. In particular, it consists of two main components: the bidirectional GRU-RNN [2] and the hierarchical attention structure.

4.1.1 Bidirectional GRU-RNN. We employ the GRU-based RNN to encode the sequence of words and comments. It has been shown that GRU-based RNNs work particularly well on smaller datasets [11], which is especially useful in cyberbullying detection due to the limited amount of

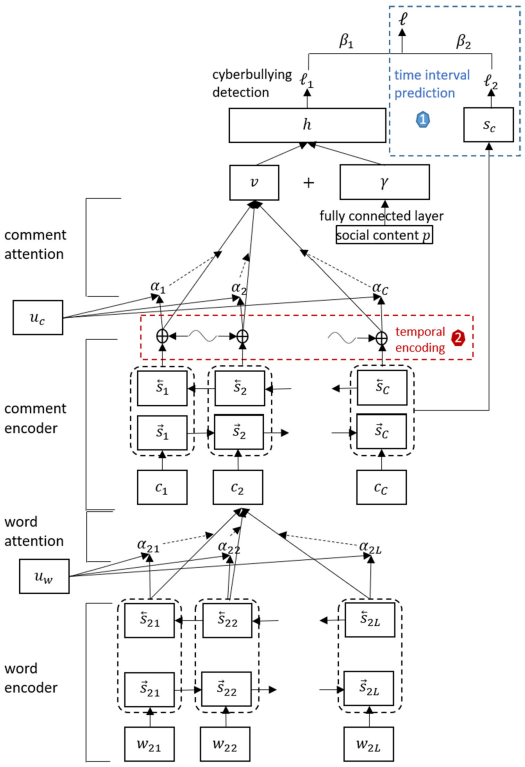


Fig. 3. Modeling temporal patterns of cyberbullying within a hierarchical attention network. HANCD (blue) jointly detects cyberbullying instances and predict time; HANT (red) seeks to capture the temporal orderings of user-generated comments.

available data. The GRU framework comprises two types of gates: the update gate and the reset gate. Each gate depends only on the previous hidden state and the bias. The new state computed by GRU is a linear interpolation between the previous state and the current state. The bidirectional GRU can summarize information of words and comments from both directions and, therefore, is able to better capture the contextual information embedded in the latent representation. In this work, we denote the forward and backward GRU as $\overrightarrow{GRU}$ and $\overleftarrow{GRU}$, respectively.

4.1.2 Hierarchical Attention Structure. Empirical results in a wide range of NLP tasks have shown that the quality of document representation can be greatly improved when attention mechanisms are properly integrated to recognize the important characters, words, and sentences in the inherently hierarchical structure of a document [15, 49, 51, 52]. In our case, a sequence of words form a comment and a sequence of comments, along with temporal information and social content form a social media session. This hierarchy incorporates the word and comment encoders, and word- and comment-level attention mechanisms. We specify, next, how to build the hierarchical attention network to model the structure of social media sessions.

Word Encoder and Attention Mechanisms

Supposing that a comment i has L_i words and $\mathbf{w}_{it}$ denotes the tth word in the ith comment, we first embed the input words in a latent space via an embedding matrix $\mathbf{W}_e$,

$$\mathbf{x}_{it} = \mathbf{W}_e \mathbf{w}_{it}, \forall t \in [1, L_i], i \in [1, C], \tag{1}$$

where C denotes the number of comments. The bidirectional GRU is employed to capture each word's contextual information, i.e., the information embedded in the neighboring words. The forward GRU ($\overrightarrow{GRU}$) reads the ith comment from the embedding of its first word $\mathbf{x}_{i1}$ to its last word $\mathbf{x}_{iL_i}$ and the backward GRU ($\overleftarrow{GRU}$) reads reversely from $\mathbf{x}_{iL_i}$ to $\mathbf{x}_{i1}$. As such, the forward and backward hidden states are computed as follows:

$$\overrightarrow{\mathbf{s}}_{it} = \overrightarrow{GRU}(\mathbf{x}_{it}), \quad \forall t \in \{1, 2, \ldots, L_i\}, i \in \{1, 2, \ldots, C\},$$
$$\overleftarrow{\mathbf{s}}_{it} = \overleftarrow{GRU}(\mathbf{x}_{it}), \quad \forall t \in \{L_i, L_i - 1, \ldots, 1\}, i \in \{1, 2, \ldots, C\}.$$

The embedding of a given word $\mathbf{w}_{it}$ is then the concatenation of the forward hidden state $\overrightarrow{\mathbf{s}}_{it}$ and the backward hidden state $\overleftarrow{\mathbf{s}}_{it}$, i.e., $\mathbf{s}_{it} = [\overrightarrow{\mathbf{s}}_{it}, \overleftarrow{\mathbf{s}}_{it}]$.

As expected, words are not equally informative regarding detecting cyberbullying instances. Here, we adopt additive attention [2, 52] to automatically highlight the words and comments that are more important for learning discriminative representations of social media sessions in cyberbullying detection. Specifically, we first feed the word embedding $\mathbf{s}_{it}$ to a fully connected layer and get its hidden state:

$$\mathbf{h}_{it} = \tanh(\mathbf{W}_s \mathbf{s}_{it} + \mathbf{b}_s), \tag{2}$$

where $\mathbf{W}_s$ is the weight matrix of the fully connected layer and $\mathbf{b}_s$ is the bias term. We then assume that there is a word-level latent vector $\mathbf{u}_w$ that contains all the contextual information in the comment [52]. We calculate the similarity between the context vector $\mathbf{u}_w$ and $\mathbf{h}_{it}$ as follows:

$$\alpha_{it} = \frac{\exp(\mathbf{h}_{it}^\top \mathbf{u}_w)}{\sum_t \exp(\mathbf{h}_{it}^\top \mathbf{u}_w)}. \tag{3}$$

Here, α_{it} is a normalized importance weight for word $\mathbf{w}_{it}$. Finally, the comment representation is the sum of the weighted word embeddings:

$$\mathbf{c}_i = \sum_t \alpha_{it} \mathbf{s}_{it}. \tag{4}$$

Comment Encoder and Attention Mechanisms

Given a sequence of comment representations $\{\mathbf{c}_1, \ldots, \mathbf{c}_i, \ldots \mathbf{c}_C\}$ generated by the word encoder, we can compute the representation of a social media session in a similar way to the aforementioned procedure. Given the timestamps of a sequence of comments $(t_1, t_2, \ldots, t_C)$, we first calculate a sequence of time intervals $(\Delta t_1, \Delta t_2, \ldots \Delta t_C)$ with $\Delta t_i = t_i - t_{i-1}, i \in [2, C], \Delta t_1 = 0$. $\mathbf{c}_i$ is fed into the bidirectional GRU of the comment encoder (as shown in Figure 3):

$$\overrightarrow{\mathbf{s}}_i = \overrightarrow{GRU}(\mathbf{c}_i), \quad \overleftarrow{\mathbf{s}}_i = \overleftarrow{GRU}(\mathbf{c}_i), \quad i \in [1, C]. \tag{5}$$

Similarly, we concatenate the forward and backward hidden states $\overrightarrow{\mathbf{s}}_i$, $\overleftarrow{\mathbf{s}}_i$ to get the embedding of comment i, i.e., $\mathbf{s}_i = [\overrightarrow{\mathbf{s}}_i, \overleftarrow{\mathbf{s}}_i]$, which emphasizes the comment i as well as summarizes the contextual information from the neighboring comments of i. The representation of a social media session $\mathbf{v}$ can then be obtained with the attention mechanism illustrated in the following equations:

$$\mathbf{h}_i = \tanh(\mathbf{W}_c s_i + \mathbf{b}_c), \quad \alpha_i = \frac{\exp(\mathbf{h}_i^\top \mathbf{u}_c)}{\sum_i \exp(\mathbf{h}_i^\top \mathbf{u}_c)}, \quad \mathbf{v} = \sum_i \alpha_i \mathbf{s}_i, \tag{6}$$

where $\mathbf{h}_i$ is the hidden state parameterized by $\mathbf{W}_c$ and $\mathbf{b}_c$, and $\mathbf{u}_c$ is a comment-level context vector. Both word- and comment-level context vectors can be randomly initialized and learned in the training process [52]. We also project social content $\mathbf{z}_i$ into a latent space via a fully connected layer and get the dense vector γ:

$$\gamma = \tanh(\mathbf{W}_z \mathbf{z}_i + \mathbf{b}_z). \tag{7}$$

The final representation of a social media session is $\mathbf{m} = [\mathbf{v}, \gamma]$.

4.2 Modeling Temporal Patterns in Cyberbullying Detection

In this subsection, we introduce two approaches for modeling the temporal patterns of cyberbullying behavior. The first method, referred to as HANCD, was first introduced in our previous work [7]. It is based on multi-task learning where we jointly detect cyberbullying and predict the time intervals between adjacent comments. The second method, referred to as HANT, is motivated by the positional encoding mechanism in the Transformers [20], an advanced deep learning model used primarily in the field of NLP. In HANT, we explicitly encode the temporal ordering of comments with trigonometric functions.

Time Interval Prediction. Results from early studies showed that cyberbullying on social media takes place across a stream of comments that are typically relatively close in time, i.e., time intervals between adjacent comments are relatively short in a bullying sessions [45]. More recent findings reported in Reference [19] further shed light on the significance of using time intervals to facilitate cyberbullying detection. In particular, the authors investigated the temporal properties of bullying and non-bullying sessions using comment-level cyberbullying labels and concluded that: (1) In both bullying and non-bullying sessions, most of the bullying comments occurred in the first hours after an initial post and the bullying comment counts in bullying sessions are significantly larger than those of non-bullying sessions; and (2) a relatively short time interval between consecutive cyberbullying comments is observed in bullying sessions. Given these distinct patterns regarding time intervals between bullying and non-bullying sessions, HANCD seeks to predict time intervals to glean and utilize these temporal patterns to augment the efficacy of cyberbullying detection. The joint learning of multiple tasks can result in high-quality representations [1].

Conventionally, cyberbullying detection is viewed as a binary classification task. We thereby define the first objective function of HANCD as the log loss parameterized by $\mathbf{W}_n$ and $\mathbf{b}_n$:

$$p_n = \sigma(\mathbf{W}_n\mathbf{m} + \mathbf{b}_n), \quad \ell_1 = -\frac{1}{N}\sum_{n=1}^{N} y_n \log p_n + (1 - y_n)\log(1 - p_n), \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function. Other loss functions for binary classification such as the Hinge loss are left to be explored in future work. For time interval prediction, we first extract the comment representation $\mathbf{s}_i$ from the hierarchical attention network. Then, the second objective function is defined as minimizing the mean squared error between the predicted and true time intervals:

$$\ell_2 = \frac{1}{NC}\sum_{n=1}^{N}\sum_{i=1}^{C} |||\mathbf{A}_n\mathbf{s}_i + \mathbf{q}_n - \Delta t_i|||^2, \tag{9}$$

where $\mathbf{A}_n$ and $\mathbf{q}_n$ are the weight matrix and bias, respectively. The final objective function of HANCD is the weighted sum of the cyberbullying detection loss and the time interval prediction loss:

$$\ell = \beta_1\ell_1 + \beta_2\ell_2, \tag{10}$$

where β_1 and β_2 are the hyper-parameters balancing the cyberbullying detection task and time interval prediction task, respectively, in the overall function.

Temporal Encoding. The second method—HANT—focusing on the temporal ordering of the comments in addition to their textual content aims to capture the semantics of a conversation in a social media session. In addition to the insights from the positional encoding in Transformers, this method is also guided by important findings from the burst analysis of the 100 social media sessions with comment-level labels in Reference [19]. Specifically, this study concluded that (1) a sequence of intense cyberbullying comments (in cyberbullying sessions) occurred during the first few hours after the initial posts, and the intensity is even higher within the first hour; (2) the comment count does not decrease monotonically but shows spikes (bursts of activity) over time.

As typically characterized by *repetitive acts*, cyberbullying behavior can implicitly present this *periodicity*. In this method, we propose temporal encoding, a simple but effective technique to incorporate temporal ordering that reflects the periodicity of cyberbullying behavior. Compared to HANCD, HANT provides a more unified way to integrate the temporal information for cyberbullying detection: Whereas HANCD models the intensity of interactions using time intervals between adjacent comments, HANT uses temporal encoding to explicitly model the structural information of the stream of comments based on their posted time.

Let $t_j \in \mathbb{N}$ be the timestamp of comment j in a social media session, $\mathbf{p}_{t_j} \in \mathbb{R}^{TIME\_DIM}$ be its corresponding encoding, and $TIME\_DIM$ be the encoding dimension. We define the encoding function $g : \mathbb{N} \rightarrow \mathbb{R}^{TIME\_DIM}$ as follows:

$$\mathbf{p}_{t_j}^{(i)} = g(t_j)^{(i)} := \begin{cases} \sin(\omega_k \cdot t_j), & \text{if} \quad i = 2k \\ \cos(\omega_k \cdot t_j), & \text{if} \quad i = 2k+1, \end{cases} \tag{11}$$

where $\omega_k = \frac{1}{100^{2k/d}}$ is the angular frequency, i.e., the rate of change of the function argument in units of radians per second. Therefore, with different values of $k \in \mathbb{Z}^*$, $g(\cdot)$ forms a geometric progression from 2π to $100 \cdot 2\pi$ on the wavelengths. This encoding method maps each timestamp t_j into a latent vector that considers the periodicity of bullying comments:

$$\mathbf{p}_{t_j} = \begin{bmatrix} \sin(\omega_1 \cdot t_j) \\ \cos(\omega_1 \cdot t_j) \\ \cdots \\ \sin(\omega_{TIME\_DIM/2} \cdot t_j) \\ \cos(\omega_{TIME\_DIM/2} \cdot t_j) \end{bmatrix}_{TIME\_DIM \times 1}. \tag{12}$$

Next, we feed the temporal encoding to a fully connected layer and get the hidden state of $\mathbf{p}_{t_j}$:

$$\tilde{\mathbf{p}}_{t_j} = \tanh(\mathbf{W}_p \mathbf{p}_{t_j} + \mathbf{b}_p), \tag{13}$$

where $\mathbf{W}_p$ and $\mathbf{b}_p$ are the weights and bias term, respectively. The final comment representation $\tilde{\mathbf{c}}_j = [\mathbf{c}_j, \tilde{\mathbf{p}}_{t_j}]$ is then used in Equation (5) to encode the session representations. The final objective function is the log loss similar to Equation (8).

5 EVALUATION

We conduct both quantitative and qualitative analyses on a real-world *Instagram* dataset from Reference [22] to evaluate the performance of HANCD (based on time interval prediction) and HANT (based on temporal encoding). We seek to examine the following aspects of the proposed models:

- How does the proposed framework fare against conventional models for cyberbullying detection and the state-of-the-art model that integrates the use of temporal patterns in cyberbullying?
- How does the performance of HANT differ from that of HANCD?
- How does data oversampling influence the performance of the proposed framework and conventional models?
- How robust are the proposed methods w.r.t. different model parameters?

5.1 Dataset

As one of the most widely used social networking sites, Instagram allows users to upload photos and videos and comment on posts that other users have made public. Cyberbullying on Instagram can thus take the form of posting insulting comments, captions, or hashtags, posting humiliating images/videos of others, and editing and re-posting images/videos originally posted by others

[22]. The *Instagram* dataset from Reference [22] was collected using a snowball sampling method. For each public user, the collected data includes the media object (i.e., image) that the user had posted and the text of the 150 most recently posted comments (or fewer, depending on the total number of comments for an image), the list of users who follow or are followed by the user, and the list of users who have commented/liked the media objects shared by the user. The average number of comments per session in this dataset is 71. Data encoding in terms of whether the session constituted cyberbullying was performed on CrowdFlower[5] and each session was labeled by five different contributors. The final decision comes from the label with most votes. This dataset includes 2,218 social media sessions, among which 1,540 are labeled as *Non-bullying* and 678 are labeled as *Bullying*. We use 80% of the data for training and the remaining for testing.

5.2 Baseline Methods

The baseline models consist of common text classification models—KNN, Naïve Bayesian, Random Forest, Logistic Regression, and XGBoost [5]—trained on different sets of textual features. These features include BoW, word-, N-Gram- and character-level TF-IDF vectors, and pre-trained word embeddings, as well as psychological features obtained from Linguistic Inquiry Word Count (LIWC) [40]:

Count Vector. It is a matrix where every cell represents the frequency count of a particular term in a particular social media session.

TF-IDF Vector. TF-IDF score is a numerical statistic that is intended to reflect how important a term is to a social media session in a collection. It can be generated at different levels of input tokens: Word-Level TF-IDF (**Word TF-IDF**), N-gram-level TF-IDF (**N-gram TF-IDF**) and Character-Level TF-IDF (**Char TF-IDF**).

LIWC. LIWC represents the output of psychometric analyses of text. It counts words belonging to certain categories of personality traits, feelings, and psychological motives. Previous findings have concluded that using features from psychometric analysis can improve the performance of cyberbullying detection models [8, 39].

Word Embedding. In this approach, we use the pre-trained Google News corpus (3B running words) word vector model.[6] The representation of a social media session is simply the average of all of the word embeddings.

Moreover, we compare the proposed framework with deep learning models for text classification including LSTM, CNN, and HAN, as well as existing cyberbullying detection models that do not require comment-level labels; this includes Xu et al. [50] and Soni and Singh [45]:

- *LSTM [21].* LSTM is a special kind of RNN capable of learning long-term dependencies between words. It is one of the most popular architectures used in NLP tasks.
- *CNN [29].* CNN is a class of deep, feed-forward artificial neural networks. It has been applied to various NLP tasks showing promising results. CNN is able to detect patterns of multiple sizes by varying the size of the kernels and concatenating their outputs.
- *HAN [52].* HAN is used in document classification and has two features: (1) a hierarchical structure and (2) two levels of attention mechanisms applied at the word- and sentence-level.
- *Xu et al. [50].* This pre-trained model extracts textual features including unigrams, unigram+bigrams, and POS-tagged N-grams to train a Support Vector Machine on labeled Twitter data.

[5] http://www.figure-eight.com/.
[6] https://code.google.com/archive/p/word2vec/.

Table 2. Performance Comparisons of Different Models (Precision Score)

Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.818±0.07	0.632±0.05	0.705±0.03	0.466±0.03	0.683±0.04	0.529±0.03
Naïve Bayesian	0.625±0.05	**0.895±0.06**	0.779±0.03	0.717±0.06	0.489±0.05	0.439±0.02
Logistic Regression	0.713±0.02	0.821±0.05	<u>0.873±0.04</u>	0.797±0.05	0.766±0.03	0.807±0.02
Random Forest	0.753±0.02	0.780±0.03	0.770±0.05	0.789±0.03	0.842±0.03	0.802±0.03
XGBoost	0.799±0.03	0.802±0.04	0.806±0.03	0.825±0.04	0.839±0.03	0.788±0.04
Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	HANT
0.668±0.04	0.682±0.04	0.759±0.03	0.087±0.02	0.794±0.03	0.763±0.07	0.723±0.05

Table 3. Performance Comparisons of Different Models (Recall Score)

Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.363±0.04	0.541±0.03	0.454±0.05	0.784±0.04	0.589±0.02	0.831±0.03
Naïve Bayesian	0.761±0.06	0.266±0.06	0.547±0.05	0.348±0.05	0.501±0.03	<u>0.879±0.04</u>
Logistic Regression	0.670±0.05	0.566±0.06	0.517±0.05	0.579±0.05	0.601±0.02	0.584±0.03
Random Forest	0.559±0.05	0.506±0.05	0.507±0.06	0.458±0.06	0.523±0.02	0.495±0.03
XGBoost	0.671±0.04	0.657±0.04	0.679±0.05	0.654±0.03	0.583±0.01	0.661±0.04
Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	HANT
0.656±0.03	0.643±0.05	0.715±0.06	**1.000±0.02**	0.691±0.03	0.802±0.06	0.801±0.02

- *Soni & Singh [45].* This is the state-of-the-art computational model that considers the temporal dynamics of cyberbullying behavior without the need for comment-level cyberbullying labels. It models users' commenting behavior as point processes and extracts several temporal features to distinguish between bullying and non-bullying social media sessions.

We implemented *Soni & Singh* using several machine learning models and report the results of the best model—XGBoost. Because real-world cyberbullying datasets are typically imbalanced, i.e., each class does not make up an equal proportion of the dataset, the tradeoff between recall and precision may be affected. We therefore report Precision, Recall, F1, and AUC scores for a holistic understanding of the models' performance. All presented results are averaged over 10 runs.

5.3 Quantitative Results

5.3.1 Original Dataset. We first evaluate the models using the original imbalanced *Instagram* dataset. We report the mean and standard deviation in Tables 2–5 with the best and second-best approaches highlighted in bold and underscored font, respectively. We can observe the following:

- Standard text classification models such as KNN often present skewed Precision and Recall scores, e.g., a low Precision and a high Recall or vice versa. An extreme case can be seen in the results of Xu et al., which presents a rather low Precision but perfect Recall (1.000). Neural Network–based models such as LSTM, HAN, HANCD, and HANT achieve more balanced Precision and Recall. This result implies that deep learning models can better balance the performance w.r.t. each class in cyberbullying detection, where the real-world datasets are typically imbalanced. Among the various common text classification models, XGBoost

Table 4. Performance Comparisons of Different Models (F1 Score)

Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.502±0.04	0.581±0.03	0.548±0.05	0.584±0.02	0.670±0.02	0.646±0.03
Naïve Bayesian	0.685±0.03	0.405±0.07	0.641±0.03	0.465±0.05	0.494±0.03	0.586±0.02
Logistic Regression	0.689±0.02	0.668±0.04	0.648±0.04	0.668±0.03	0.673± 0.02	0.677±0.02
Random Forest	0.641±0.03	0.612±0.04	0.608±0.04	0.577±0.04	0.645±0.01	0.610±0.02
XGBoost	0.728±0.02	0.721±0.02	0.679±0.03	0.728±0.02	0.688±0.02	0.717±0.05
Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	HANT
0.661±0.02	0.663±0.02	0.734±0.02	0.502±0.03	0.739±0.03	**0.778±0.01**	0.763±0.02

Table 5. Performance Comparisons of Different Models (AUC Score)

Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.664±0.02	0.702±0.02	0.686±0.03	0.698±0.02	0.755±0.01	0.754±0.03
Naïve Bayesian	0.781±0.02	0.626±0.03	0.740±0.02	0.644± 0.02	0.627±0.02	0.696±0.03
Logistic Regression	0.777±0.02	0.757±0.03	0.742±0.02	0.757±0.02	0.757±0.01	0.762±0.01
Random Forest	0.740±0.02	0.722±0.03	0.720±0.03	0.703±0.03	0.738±0.01	0.721±0.01
XGBoost	0.799±0.02	0.793±0.02	0.764±0.02	0.797±0.01	0.765±0.01	0.791±0.02
Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	HANT
0.746±0.02	0.756±0.02	0.808±0.02	0.513±0.02	0.809±0.02	**0.839±0.01**	0.836±0.01

consistently outperforms other models regardless of the input features. There is no clear evidence showing the superiority of any of the included text features over others.

- HANCD and HANT present the best and the second-best overall performance regarding F1 and AUC scores compared to the baseline methods. Specifically, HANCD improves over the best baseline model (i.e., Soni & Singh) by 5.3% and 3.8% w.r.t. F1 and AUC scores, respectively. HANT outperforms Soni & Singh by 3.2% and 3.3% w.r.t. F1 and AUC scores, respectively. Two-tailed t-tests[7] further confirm the significant improvement of HANCD and HANT over all baseline models. Specifically, we performed a series of t-tests in which we compared the mean for each evaluation metric for a given model (across the 10 runs) with the mean for each of the other models. Each comparison yielded a t-value and an associated significance level (i.e., p-value), where a p-value below .05 indicates a significant improvement of one model over the other. (In other words, a p-value of $<.05$ supports the rejection of the null hypothesis that the two models being compared have identical means for a given evaluation metric.) These results thus corroborate the effectiveness of modeling temporal patterns of cyberbullying detection using a hierarchical attention network.
- Among neural network–based models, HAN, HANCD, and HANT outperform LSTM and CNN in the identification of cyberbullying sessions w.r.t. all of the evaluation metrics. This finding suggests that it is critical to model the hierarchical structure of social media sessions and the attention levels associated with words and comments in session-based cyberbullying detection. When comparing the two temporal modeling methods, HANCD consistently

[7]Implemented using the Python package scipy.stats.ttest_ind.

Table 6. Performance Comparisons of Different Models Using Balanced Data (Precision Score)

Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.397±0.05	0.348±0.03	0.356±0.03	0.342±0.03	0.519±0.03	0.401±0.02
Naïve Bayesian	0.429±0.04	0.589±0.03	0.600±0.04	0.548±0.04	0.494±0.04	0.495±0.02
Logistic Regression	0.606±0.06	0.703±0.03	0.701±0.03	0.660±0.03	0.580±0.03	0.624±0.02
Random Forest	0.716±0.02	0.751±0.03	0.731±0.04	0.780±0.02	0.780±0.05	0.774±0.04
XGBoost	0.749±0.02	0.772±0.01	0.731±0.04	**0.785±0.02**	0.724±0.05	0.733 ± 0.02

Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	HANT
0.623±0.04	0.556±0.04	0.655±0.09	0.087±0.02	0.685±0.03	0.747±0.06	0.658±0.04

Table 7. Performance Comparisons of Different Models Using Balanced Data (Recall Score)

Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.480±0.02	0.946±0.02	0.927±0.04	0.980±0.01	0.668±0.04	0.926±0.02
Naïve Bayesian	0.785±0.03	0.828±0.05	0.770±0.05	0.796±0.03	0.498±0.03	0.830±0.02
Logistic Regression	0.760±0.03	0.752±0.05	0.717±0.05	0.761±0.05	0.729.0.04	0.809±0.03
Random Forest	0.779±0.03	0.642±0.03	0.611±0.03	0.552±0.03	0.617±0.05	0.603±0.05
XGBoost	0.734±0.03	0.728±0.03	0.642±0.03	0.679±0.04	0.620±0.04	0.734±0.03

Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	HANT
0.669±0.04	0.702±0.06	0.823±0.06	**1.000±0.02**	0.757±0.03	0.831±0.07	0.870±0.0.05

outperforms HANT regarding the average Recall, F1, and AUC scores, however, the results of a t-test shows no significant differences between the results of the two models. We conjecture that the intensity of user interactions might be more informative than the temporal ordering of comments for session-based cyberbullying detection. More evidence is needed to make conclusive claims.

5.3.2 Oversampling Imbalanced Data. A problem with cyberbullying detection using real-world data is that there are too few bullying instances for a model to effectively learn the decision boundary. To better understand how the data imbalance may influence the performance of HANCD, HANT, and existing models, we employed a widely recognized data augmentation[8] technique referred to as the Synthetic Minority Oversampling Technique (SMOTE) [3]. Specifically, we oversampled the bullying instances in the training data so the number of bullying and non-bullying instances are equal. Data for testing remained imbalanced. We re-executed the previous experiments and present the results in Tables 6–9.

We start by observing that most of the standard text classification models (e.g., Logistic Regression and Random Forest) present a large improvement in terms of Recall, rendering a more balanced Precision-Recall tradeoff and better overall performance w.r.t. F1 and AUC. However, the imbalanced Precision-Recall issue of KNN appears to be exacerbated when it is trained on the synthetically balanced data. This suggests that a simple data augmentation strategy may not be appropriate for models that are sensitive to noisy data, e.g., KNN. Another observation is that

[8]Undersampling the majority class will largely reduce the number of available samples for training, which is not appropriate for deep learning models.

Table 8. Performance Comparisons of Different Models Using Balanced Data (F1 Score)

Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.433±0.03	0.508±0.04	0.514±0.04	0.507±0.03	0.583±0.02	0.560±0.03
Naïve Bayesian	0.554±0.04	0.688±0.03	0.673±0.03	0.648±0.03	0.495±0.03	0.620±0.02
Logistic Regression	0.672±0.04	0.726±0.03	0.708±0.03	0.706±0.03	0.645±0.02	0.705±0.02
Random Forest	0.746±0.02	0.692±0.02	0.665±0.03	0.646±0.03	0.688±0.05	0.677±0.05
XGBoost	0.741±0.02	0.749±0.01	0.683±0.02	0.728±0.03	0.666±0.03	0.733±0.02

Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	HANT
0.643±0.02	0.627±0.01	0.723±0.03	0.502±0.03	0.719±0.02	**0.782±0.01**	0.747±0.02

Table 9. Performance Comparisons of Different Models Using Balanced Data (AUC Score)

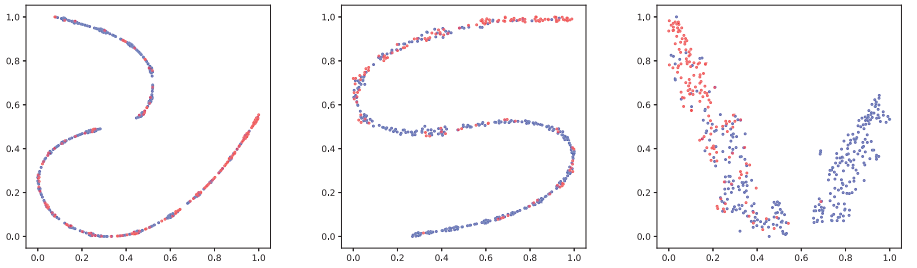
Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.573±0.02	0.570±0.01	0.583±0.02	0.562±0.01	0.682±0.02	0.655±0.01
Naïve Bayesian	0.655±0.02	0.783±0.02	0.768±0.02	0.759±0.01	0.632±0.02	0.726±0.02
Logistic Regression	0.767±0.02	0.804±0.02	0.789±0.02	0.792±0.03	0.743±0.02	0.796±0.02
Random Forest	0.819±0.01	0.773±0.02	0.755±0.01	0.741±0.02	0.769±0.03	0.762±0.03
XGBoost	0.811±0.02	0.815±0.02	0.768±0.01	0.797±0.02	0.756±0.02	0.807±0.02

Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	HANT
0.740±0.01	0.733±0.02	0.808±0.02	0.513±0.02	0.808±0.02	**0.845±0.02**	0.825±0.02

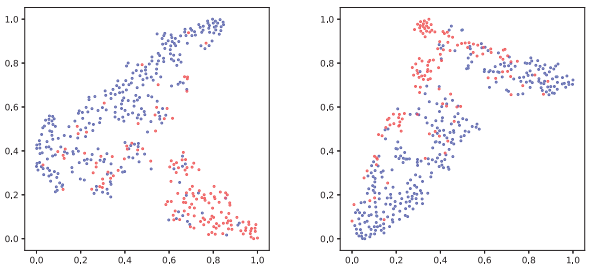
most of the models (e.g., LSTM and HAN) that highly depend on the structural information show slightly decreased F1 and AUC scores. We surmise that SMOTE, which generates synthetic data by simple linear interpolation, can disrupt the structural dependencies of text and nonlinear correlations among different modalities within a session, e.g., text and temporal information. Of particular interest is that the performance of HANCD is slightly improved by the data augmentation technique. This is partly explained by the fact that HANCD separates the cyberbullying detection task from the temporal dynamic fitting task. The advantage of incorporating additional synthetic data for training thus outweighs the associated disadvantage when applied to HANCD. By contrast, HANT unifies the temporal ordering and sequence text modeling. Therefore, independently oversampling data from correlated modalities can significantly reduce their inherent relationships. With balanced data, HANCD still achieves the best overall performance and HANT achieves the second-best AUC.

5.4 Qualitative Analysis

We further examine the latent representations of social media sessions learned by various deep learning models via 2D t-SNE visualizations [34] with perplexity values set to 30. Results for CNN, LSTM, HAN, HANT, and HANCD are presented in Figure 4. On one hand, we observe that for CNN and LSTM, the representations of bullying and non-bullying sessions are mostly overlapping. Models that explicitly consider the hierarchical structure of a social media session, on the other hand, can learn more discriminative representations. HAN, HANT, and HANCD exhibit two separate clusters, but the left cluster of HAN in Figure 4(c) contains a more mixed set of bullying and non-bullying instances than the clusters in HANT and HANCD. The 2D t-SNE visualizations



(a) *t-SNE visualization of LSTM* (b) *t-SNE visualization of CNN* (c) *t-SNE visualization of HAN*



(d) *t-SNE visualization of HANCD* (e) *t-SNE visualization of HANT*

Fig. 4. 2D t-SNE visualizations of the learned representations (Perplexity score = 30).

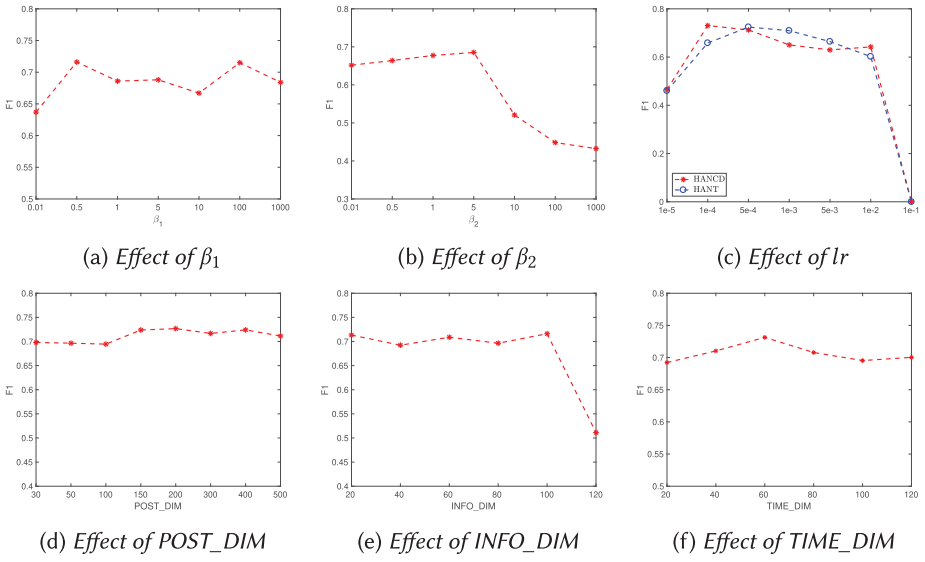
of bullying and non-bullying instances in HANT and HANCD are more spread over the latent space compared to other methods.

To summarize, from the quantitative results and the t-SNE visualizations of the learned representations, we conclude that both HANCD and HANT can learn more disciminative session representations and accomplish more accurate detection of cyberbullying instances compared to common text classification models, deep learning models, and prior cyberbullying detection models in a session-based cyberbullying detection task. When employing data oversampling techniques, one needs to pay special attention to the design of the deployed models and inherent relationships among different information sources these models rely on.

5.5 Parameter Analysis

HANCD and HANT together include six key parameters: β_1, β_2, lr, POST_DIM, INFO_DIM, and TIME_DIM. β_1 and β_2 balance between the task of cyberbullying detection and the task of time interval prediction in HANCD. lr is the learning rate, and POST_DIM and INFO_DIM are the embedding dimensions of words and social content, respectively. TIME_DIM is a unique parameter of HANT denoting the embedding dimension of temporal encoding. To investigate the sensitivity of these parameters, we vary one parameter at a time and evaluate how it affects the cyberbullying detection performance w.r.t. F1 score. Due to different numerical scales, we vary different parameters among various ranges. We run HANCD to analyze the sensitivity of β_1, β_2, POST_DIM, and INFO_DIM, and HANT to study TIME_DIM. As lr is one of the most important parameters in HANCD, we examine the sensitivity of lr in both HANCD and HANT. We summarize the parameter study results in Figure 5.

As shown in Figures 5(a)–(b), HANCD is more sensitive to β_2, the weight of time interval prediction, than β_1, the weight of cyberbullying detection. Specifically, as β_1 becomes larger, HANCD



(a) *Effect of* β_1 (b) *Effect of* β_2 (c) *Effect of lr*

(d) *Effect of POST_DIM* (e) *Effect of INFO_DIM* (f) *Effect of TIME_DIM*

Fig. 5. Parameter sensitivity study (with 80% dataset for training). *lr* is examined in both HANCD and HANT.

this fuckin bitch .
that 's fucking disgusting its fanfic about zayn harry and lux its nasty .
she is sick bitch ... i m disgusted .
that was most fucked up fanfic i have ever read in my whole entire life .... wow just wow .
what hell is wrong with her .
why would you right that why would you think of that .

who has more muscle dam you guys just start something that is n't meant to happen he 's being him so let him be .
hey comment back and tag my name .
he 's my friend 's cousin .
how old r u ur white piece of trash .
plz say 12 cocky sumbitch ai n't ya .

Fig. 6. Two Instagram sessions identified as bullying.

pays more attention to better performing cyberbullying detection, leading to a trend of slightly improved F1 score. HANCD is robust to β_2 in a certain range ([0.01, 5]) and the F1 score improves as β_2 becomes larger. However, the performance drops significantly when $\beta_2 > 5$. Based on the analysis of these two parameters, we can conclude that time interval prediction can improve the performance of cyberbullying detection when the parameter is set properly. Results in Figures 5(c)–(e) reveal that HANCD and HANT are robust to *lr*, POST_DIM, and INFO_DIM in a large range, whereas an extremely large *lr* and INFO_DIM can reduce the cyberbullying detection performance significantly. As shown in Figure 5(f), HANT is robust to the changes of TIME_DIM. In general, HANCD and HANT are not sensitive to most of the model parameters in a large range and, consequently, can be tuned for various application purposes.

5.6 Case Study

To validate the capability of selecting informative comments and words in a session, we visualize the attention levels of two Instagram sessions that were identified as bullying and two identified as non-bullying. The results can be seen in Figures 6–7. Each figure contains two examples (or sub-figures). Every line in each sub-figure is a comment. Shades of blue denote the relative importance of comments, and shades of red denote the relative importance of words (in both cases, darker shades represent higher attention levels). Because all of the selected sessions have multiple comments, only a portion of the content is shown here. Figure 6 shows that the hierarchical

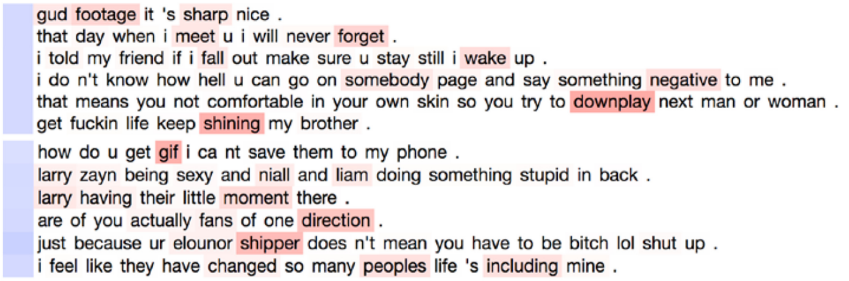


Fig. 7. Two Instagram sessions identified as non-bullying.

attention network can select the words that are strongly associated with bullying, such as *trash, sumb*tch, f*ckin, b*tch, disgusted*, and *hell*. In the second example in Figure 7, we observe that hierarchical attention networks can also deal with complex cross-comment contexts: although the session might appear to be a bullying session based on the second comment from the bottom, the hierarchical attention network predicts the session as non-bullying, because it also considers the contextual information.

6 CONCLUSION

This article studies one of the key characteristics of cyberbullying behavior—*repetition*—that has been largely overlooked in prior research. Due to the limited accessibility of social media datasets with comment-level cyberbullying labels, it is especially important to leverage the auxiliary temporal information to understand the evolving behavior of users posting cyberbullying comments. Our contribution focuses on using the temporal information of social media sessions to capture the repetitive nature of cyberbullying. That is, we provide new insights on HANCD [7], which explores the commonalities and differences between cyberbullying detection and time interval prediction, and further propose a unified approach—HANT—that explicitly models the temporal ordering of the sequence of comments. A defining aspect of these two approaches is that they build on a hierarchical attention network that enables us to construct a social media session in a bottom-up manner. We then incorporate temporal information at the comment-level to ultimately refine the session representations. Extensive experiments show the significance of the time-informed hierarchical attention network for cyberbullying detection.

The present work motivates several key avenues in the field. Whereas the work presented in this article focuses on the case of Instagram sessions, a complementary line of research could study and integrate the session structures of other popular social media platforms (e.g., multiple comment levels in Facebook and multiple ways of retweeting a tweet in Twitter) and run additional experiments using data collected from these platforms. Future work can also be directed towards using time series forecasting to predict future cyberbullying instances based on previously observed cases. Early detection is especially crucial to help prevent the occurrence of cyberbullying behavior and mitigate its negative impact on victims. In addition, other mechanisms for analyzing conversations that happen across multiple comments or even multiple sessions could be used to identify implicit cyberbullying behaviors. Due to the complexity of data labeling, one may also consider using temporal information and social network information to develop unsupervised cyberbullying detection models. Causal learning is central to understanding cyberbullying behaviors, given its potential to improve both the generalizability and interpretability of cyberbullying detection models [6, 18]. Ultimately, efforts to more accurately detect and interpret cyberbullying remain a critical step toward building safer and more inclusive social interaction spaces.

REFERENCES

[1] Wasi Uddin Ahmad, Xueying Bai, Zhechao Huang, Chao Jiang, Nanyun Peng, and Kai-Wei Chang. 2018. Multi-task learning for universal sentence embeddings: A thorough evaluation using transfer and auxiliary tasks. *arXiv preprint arXiv:1804.07911* (2018).

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[3] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16 (2002), 321–357.

[4] Charalampos Chelmis and Mengfan Yao. 2019. Minority report: Cyberbullying prediction on Instagram. In *Proceedings of the 10th ACM Conference on Web Science*. 37–45.

[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794.

[6] Lu Cheng, Ruocheng Guo, and Huan Liu. 2019. Robust cyberbullying detection with causal interpretation. In *Proceedings of the World Wide Web Conference*. 169–175.

[7] Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, and Huan Liu. 2019. Hierarchical attention networks for cyberbullying detection on the Instagram social network. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 235–243.

[8] Lu Cheng, Jundong Li, Yasin Silva, Deborah Hall, and Huan Liu. 2019. PI-bully: Personalized cyberbullying detection with peer influence. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*.

[9] Lu Cheng, Jundong Li, Yasin N. Silva, Deborah L. Hall, and Huan Liu. 2019. XBully: Cyberbullying detection within a multi-modal context. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. 339–347.

[10] Lu Cheng, Yasin Silva, Deborah Hall, and Huan Liu. 2020. Session-based cyberbullying detection: Problems and challenges. *IEEE Internet Comput., Spec. Iss. Cyber-soc. Health: Promot. Good Counter. Harm Soc. Media* (2020).

[11] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[12] Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2016. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781* (2016).

[13] Maral Dadvar, F. M. G. de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the 12th Dutch-Belgian Information Retrieval Workshop (DIR'12)*. University of Ghent.

[14] Harsh Dani, Jundong Li, and Huan Liu. 2017. Sentiment informed cyberbullying detection in social media. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 52–67.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[16] Adji B. Dieng, Chong Wang, Jianfeng Gao, and John Paisley. 2016. TopicRNN: A recurrent neural network with long-range semantic dependency. *arXiv preprint arXiv:1611.01702* (2016).

[17] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *Proceedings of the Social Mobile Web Conference*.

[18] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: Problems and methods. *ACM Comput. Surv.* 53, 4 (2020), 1–37.

[19] Aabhaas Gupta, Wenxi Yang, Divya Sivakumar, Yasin N. Silva, Deborah L. Hall, and Maria Camila Nardini Barioni. 2020. Temporal properties of cyberbullying on Instagram. In *Proceedings of the World Wide Web Conference*.

[20] L. Hackett. 2017. The annual bullying survey 2017. DitchThe Label. Retrieved from https://www.ditchthelabel.org/research-papers/the-annual-bullyingsurvey-2017.

[21] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780.

[22] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the Instagram social network. *arXiv preprint arXiv:1503.03909* (2015).

[23] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-aware Multimedia*. 3–6.

[24] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Meeting of the Association for Computational Linguistics*. 1681–1691.

[25] Rie Johnson and Tong Zhang. 2014. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058* (2014).

[26] Rie Johnson and Tong Zhang. 2015. Semi-supervised convolutional neural networks for text categorization via region embedding. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 919–927.

[27] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188* (2014).
[28] Seonhoon Kim, Inho Kang, and Nojun Kwak. 2019. Semantic sentence matching with densely-connected recurrent and co-attentive information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6586–6593.
[29] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).
[30] Stefan Kombrink, Tomáš Mikolov, Martin Karafiát, and Lukáš Burget. 2011. Recurrent neural network based language modeling in meeting recognition. In *Proceedings of the 12th Conference of the International Speech Communication Association*.
[31] April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. 2013. Detecting cyberbullying: Query terms and techniques. In *Proceedings of the 4th ACM Conference on Web Science*. 195–204.
[32] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 333. 2267–2273.
[33] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 115–124.
[34] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, Nov. (2008), 2579–2605.
[35] Tomáš Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 3111–3119.
[36] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep learning based text classification: A comprehensive review. *arXiv preprint arXiv:2004.03705* (2020).
[37] Vinita Nahar, Xue Li, and Chaoyi Pang. 2013. An effective approach for cyberbullying detection. *Commun. Inf. Sci. Manag. Eng.* 3, 5 (2013), 238.
[38] Vinita Nahar, Sayan Unankard, Xue Li, and Chaoyi Pang. 2012. Sentiment analysis for effective detection of cyber bullying. In *Proceedings of the Asia-Pacific Web Conference*. Springer, 767–774.
[39] Parma Nand, Rivindu Perera, and Abhijeet Kasture. 2016. "How bullying is this message?": A psychometric thermometer for bullying. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*. 695–706.
[40] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawr. Erlb. Assoc.* 71, 2001 (2001).
[41] Semiu Salawu, Yulan He, and Joanna Lumsden. 2017. Approaches to automated detection of cyberbullying: A survey. *IEEE Trans. Automat. Control* 11, 1 (2017), 3–24.
[42] Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive pooling networks. *arXiv preprint arXiv:1602.03609* (2016).
[43] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. DiSAN: Directional self-attention network for RNN/CNN-free language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
[44] Peter K. Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *J. Child Psychol. Psychiat.* 49, 4 (2008), 376–385.
[45] Devin Soni and Vivek Singh. 2018. Time reveals all wounds: Modeling temporal characteristics of cyberbullying. In *Proceedings of the 12thInternational AAAI Conference on Web and Social Media*.
[46] Anna Squicciarini, Sarah Rajtmajer, Y. Liu, and Christopher Griffin. 2015. Identification and characterization of cyberbullying dynamics in an online social network. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 280–285.
[47] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075* (2015).
[48] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Meeting of the Association for Computational Linguistics*, Vol. 1. 1555–1565.
[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 5998–6008.
[50] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 656–666.

[51] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 5754–5764.
[52] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.
[53] Mengfan Yao, Charalampos Chelmis, and Daphney Stavroula Zois. 2019. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *Proceedings of the World Wide Web Conference*. 3427–3433.
[54] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D. Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proc. Content Anal. WEB* 2 (2009), 1–7.
[55] Justin Zhan and Binay Dahal. 2017. Using deep learning for short text understanding. *J. Big Data* 4, 1 (2017), 34.
[56] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 649–657.
[57] Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Attention-based LSTM network for cross-lingual sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 247–256.
[58] Caleb Ziems, Ymir Vigfusson, and Fred Morstatter. 2020. Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification. *arXiv preprint arXiv:2004.01820* (2020).

Received June 2020; revised September 2020; accepted December 2020