Hierarchical Attention Networks for Cyberbullying Detection on the Instagram Social Network

Lu Cheng* Ruocheng Guo* Yasin Silva† Deborah Hall‡ Huan Liu*

Abstract

Cyberbullying has become one of the most pressing online risks for young people and has raised serious concerns in society. The emerging literature identifies cyberbullying as repetitive acts that occur over time rather than one-off incidents. Yet, there has been relatively little work to model the hierarchical structure of social media sessions and the temporal dynamics of cyberbullying in online social network sessions. We propose a hierarchical attention network for cyberbullying detection that takes these aspects of cyberbullying into account. The primary distinctive characteristics of our approach include: (i) a hierarchical structure that mirrors the structure of a social media session; (ii) levels of attention mechanisms applied at the word and comment level, thereby enabling the model to pay different amounts of attention to words and comments, depending on the context; and (iii) a cyberbullying detection task that also predicts the interval of time between two adjacent comments. These characteristics allow the model to exploit the commonalities and differences across these two tasks to improve the performance of cyberbullying detection. Experiments on a real-world dataset from Instagram, the social media platform on which the highest percentage of users have reported experiencing cyberbullying, reveal that the proposed architecture outperforms the state-of-the-art method.

Keywords: Cyberbullying; Hierarchical Attention Network; Social Media

1 Introduction

Instances of cyberbullying are increasing at an alarming rate. Recent statistics reported by the American Psychological Association and the White House indicate that more than 40% of young people in the US report that they have been bullied on social media platforms [10]. A number of factors have likely contributed to this rise, including the affordability of mobile devices and the growing number of social media platforms. As a result, there has been a marked increase in research–in fields such as psychology and computer science–aimed at identifying, predicting, and ultimately preventing cyberbullying.

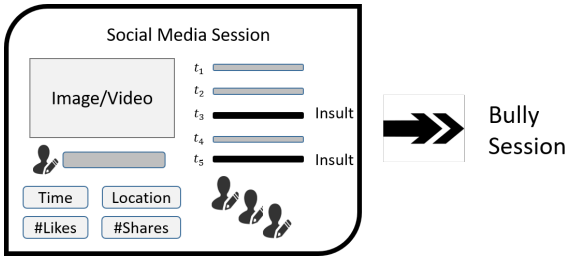


Figure 1: A social media session includes an image/video, a sequence of comments, and social media attributes. A cyberbullying session is typically composed of multiple insulting comments.

Bullying is commonly defined as a repetitive act of aggression that involves a *power imbalance* between the perpetrator and the victim [10]. Essential components of this definition include the *persistence* and *repetition* of the aggressive acts over time [10]. Existing efforts to automatically detect cyberbullying have focused chiefly on textual analysis of online messages (e.g., keywords [26, 25, 8] and sentiment analysis [9]) and have yielded prediction models with satisfactory performance. Previous work has, however, largely overlooked temporal aspects of cyberbullying behavior [31]. Given a sequence of comments, temporal analyses allow us to model the evolution of and correlations among individual comments that, together, comprise cyberbullying. Hence, analyses involving temporal characteristics enable us to identify instances of cyberbullying that take into account the full history of a social media session, including an image/video (with caption), comments (with time stamps), social information (e.g., #likes, #shares) and, importantly, the temporal relations among these components [15]. We illustrate a cyberbullying social media session in Figure 1.

A straightforward approach for incorporating temporal analyses is to extract temporal features (e.g., duration of a session, time intervals between comments) and feed them into off-the-shelf machine learning models [31]. However, simple concatenation of textual and

*School of Computer Science and Engineering, ASU.
†School of Mathematical and Natural Sciences, ASU.
‡School of Social and Behavioral Sciences, ASU.
{lcheng35,rguo12,ysilva,d.hall,huanliu}@asu.edu

temporal features may not make the best use of these features, as they are from different distributions and make unique contributions to achieve the overall goal. This approach also ignores structural properties of a social media session: a media session consists of an image/video, a caption, posted time and social attributes, and comments (each comment consists of words and time stamp). As shown in previous studies [36], modeling the knowledge of document structure improves the document representations. Furthermore, different words and comments in a social media session are differentially informative and their meaning is context dependent. For example, although both of the following comments *"You're a fucking gay!," and "Haha, I'm a gay, too."* include the word *gay*, the first one is more likely to be an instance of bullying. Words and comments that are more effective at cyberbullying detection in specific contexts should receive greater attention. To this end, a more effective framework for identifying cyberbullying instances should capture the hierarchical structure of a session, pay distinct attention to words and comments based on their context, and leverage temporal and social information in addition to textual information.

In this work, we focus on Instagram,[1] the social media platform on which the highest percentage of users report that they have experienced cyberbullying [11]. Instagram allows users to upload photos and videos and to post and comment on any photo or video that other users have made public. Bullies can post humiliating images, edit others' images and re-post them, post insulting comments, captions, or hashtags, and even create fake profiles pretending to be someone else [14]. As described below, the core contribution of this paper is the proposal of the Hierarchical Attention Networks for Cyberbullying Detection (HANCD) framework, which is designed to model the hierarchical structure, attention mechanisms [1, 36] for words and comments, temporal characteristics, and social information of a session to improve cyberbullying detection.

- **Problem.** We study the practical problem of cyberbullying detection taking into account the temporal dynamics of social media sessions. Because cyberbullying on social media takes place across a stream of comments that are typically relatively close together in time [31], we seek to jointly model cyberbullying detection and predict the time interval between adjacent comments. To this end, HANCD can exploit the commonalities and differences across the cyberbullying detection and time interval prediction tasks to improve the performance of cyberbullying detection.
- **Algorithm.** We propose a novel cyberbullying detection framework that constructs a hierarchical session representation with the aggregations of words into comments (and a caption), and then aggregates the comments, the caption, and the time and social information into a session. The model consists of two levels of attention mechanisms, one at the word level and the other at the comment level, that can capture the differential importance of words and comments in different contexts. HANCD uses context to discover when a sequence of words or comments is relevant.
- **Evaluation.** We perform empirical experiments on a real-world dataset crawled from Instagram to corroborate the efficacy of the proposed framework. Our experimental evaluation shows that HANCD outperforms previously proposed methods including the state-of-the-art approach. We also perform studies to investigate the sensitivity of the model parameters. Results reveal that our model is robust and can consequently be used for various application purposes.

2 Problem Statement

Let $\mathcal{C} = \{\int_1, \int_2, ..., \int_N\}$ be a corpus of N social media sessions. Each media session includes the caption of the posted image/video and subsequent comments, denoted together as $\{c_1, c_2, ...c_C\}$. Moreover, each comment has a time stamp and each post and owner has social media attributes (p) such as #likes, #shares, #followers. The text of the i-th comment in a session $c_i = \{w_{i1}, w_{i2}, ..., w_{iL_i}\}$ is composed of L_i words and the associated time is denoted as t_i. To train the hierarchical attention network, each media session is also associated with a binary label $y_i = \{0, 1\}$ with 1 representing a bullying session and 0 representing otherwise.

With the above notation, we now define the cyberbullying detection problem as the process of learning how to leverage *context*, *structural*, *temporal*, and *social information* to discover if a sequence of words and comments is relevant to cyberbullying.

3 Proposed Framework

The proposed framework of the Hierarchical Attention Networks for Cyberbullying Detection (HANCD) is shown in Figure 2. It consists of several components: a word sequence encoder, a word-level attention layer, a comment sequence encoder, a comment-level attention layer, contextual information, a hidden layer to embed the social media attributes and a weighted loss function, which jointly optimizes both the task of cyberbullying

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

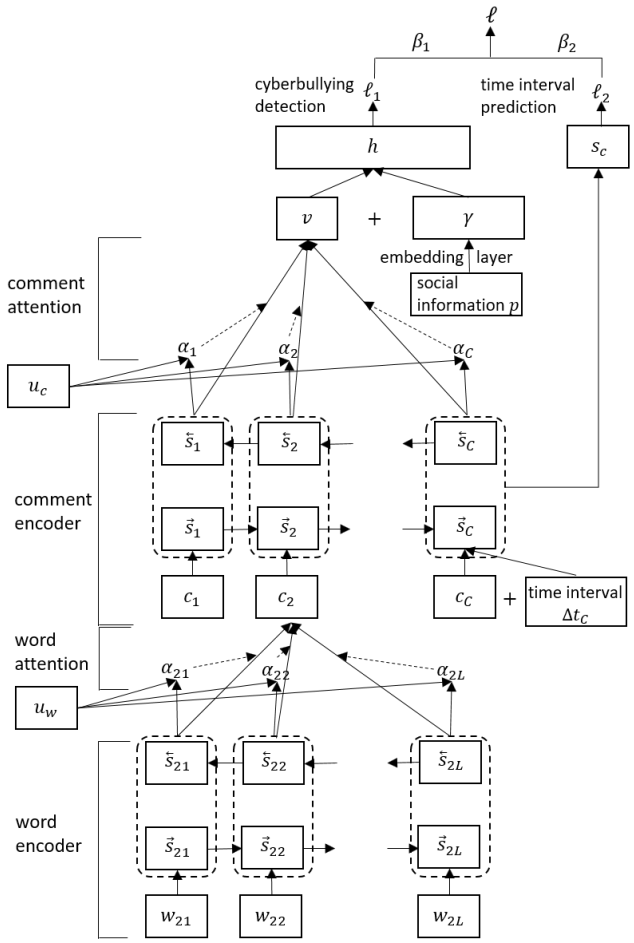


Figure 2: Hierarchical Attention Networks for Cyberbullying Detection.

detection and time interval prediction. The details of the different components are presented next.

3.1 Bidirectional GRU-RNN To model the continuous temporal phenomena, we use the bi-directional GRU [1] based RNN to encode the sequence of words and comments. GRU adds a gating mechanism to standard RNN [6] and has been found to have better performance on smaller datasets [7], which suits the case of cyberbullying detection since these datasets are hard to obtain.

There are two types of gates in the GRU framework: the update gate z_t and the reset gate r_t. Each gate only depends on the previous hidden state and the bias. Together, they control the update of the states. The new state s_t at time step t computed by GRU is a linear interpolation between the previous state s_{t-1} and the current state $\tilde{s}_t$, obtained from the information of the t-th step.

$$s_t = (1 - z_t) \odot s_{t-1} + z_t \odot \tilde{s}_t. \tag{3.1}$$

The updated state $\tilde{s}_t$ is computed with the following equation:

$$\tilde{s}_t = \tanh(W_s x_t + r_t \odot (U_s s_{t-1} + b_s)), \tag{3.2}$$

where x_t is the sequence vector at time t. The update gate z_t enables each hidden unit to maintain the information of its previous activation and the reset gate r_t controls how much and what information from the past state should be reset. Their output can be computed by Eq. 3.3-3.4.

$$z_t = \sigma(W_z x_t + U_z s_{t-1} + b_z). \tag{3.3}$$

$$r_t = \sigma(W_r x_t + U_r s_{t-1} + b_r), \tag{3.4}$$

where W_z, W_r, U_z, U_r are the related weight matrices. We then use a stack of bidirectional GRUs to encode the sequence. The bidirectional GRU can summarize information from both directions for words and comments and can, therefore, integrate the contextual information in the annotation.

3.2 Hierarchical Attention Studies have shown that an improved representation of a document can be learned by considering the structure of the document in the model architecture [36]. Similarly, in a social media session, words form comments and comments, time information, social information form a session. Hence, we first construct a representation for each word and aggregate those into a comment representation, and then construct a session representation in a similar way. We also apply attention mechanisms at both the word-level and comment-level encoders to differentiate the importance of words and comments in different contexts.

HANCD first projects all the text of a session into a vector representation h for cyberbullying detection. In the following subsection, we detail how to build the hierarchical attention network step by step.

Word Encoder and Attention

Given a comment i with L_i words w_{it}, we first embed the words to a latent space via an embedding matrix W_e,

$$w_{it} \to x_{it} : x_{it} = W_e w_{it}, \forall t \in [1, L_i], i \in [1, C]. \tag{3.5}$$

The bidirectional GRU is employed to capture the contextual information and more fine-grained annotations of words. The forward GRU $\overrightarrow{GRU}$ reads the comments c_i from w_{i1} to w_{iL_i} and the backward GRU $\overleftarrow{GRU}$ reads

from w_{iL_i} to w_{i1}. Thus, the forward/backward hidden states are computed as follows:

$$\overrightarrow{s}_{it} = \overrightarrow{GRU}(x_{it}), \quad \forall t \in [1, L_i], i \in [1, C],$$

$$\overleftarrow{s}_{it} = \overleftarrow{GRU}(x_{it}), \quad \forall t \in [L_i, 1], i \in [1, C].$$

Then, the annotation for a given word w_{it} is a concatenation of the forward hidden state $\overrightarrow{s}_{it}$ and the backward hidden state $\overleftarrow{s}_{it}$, i.e., $s_{it} = [\overrightarrow{s}_{it}, \overleftarrow{s}_{it}]$.

Words are not equally informative regarding cyberbullying detection and the same words can have different meanings within the context of different social media sessions. Instead of relying on handcrafted features, we adopt an attention mechanism [1, 36] to automatically capture the words that are more important to the meaning of the comment and aggregate the representation of weighted words to form a comment vector. Specifically, we first feed the word annotation s_{it} to a fully connected layer and get the hidden state of s_{it}:

$$h_{it} = \tanh(W_w s_{it} + b_w), \tag{3.6}$$

where W_w is the connection weight matrix of the network between two layers. To model the importance of each word, we assume that there is a word-level latent vector u_w that contains all of the contextual information in a comment [36]. We then calculate the similarity between the context vector u_w and the hidden state h_{it} as follows:

$$\alpha_{it} = \frac{\exp(h_{it}^\top u_w)}{\sum_t \exp(h_{it}^\top u_w)}. \tag{3.7}$$

Here α_{it} is a normalized importance weight for word w_{it}. Finally, the comment representation is the sum of the weighted word-level hidden states.

$$c_i = \sum_t \alpha_{it} s_{it}. \tag{3.8}$$

Comment Encoder and Attention
Given a sequence of comment vectors c_i, we can get the vector for a social media session in a similar way. Note that each comment is also associated with a time stamp. Given time information of a sequence of comments $(t_1, t_2, ..., t_C)$, we first calculate a sequence of time intervals $(\Delta t_1, \Delta t_2, ...\Delta t_C)$ with $\Delta t_i = t_i - t_{i-1}, i \in [2, C], \Delta t_1 = 0$ and then the concatenation $o_i = [c_i, \Delta t_i]$, is fed to the bidirectional GRU in the comment encoder (as shown in Figure 2):

$$\overrightarrow{s}_i = \overrightarrow{GRU}(o_i), i \in [1, C], \tag{3.9}$$

$$\overleftarrow{s}_i = \overleftarrow{GRU}(o_i), i \in [C, 1]. \tag{3.10}$$

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

$$h_i = \tanh(W_c s_i + b_c), \tag{3.11}$$

$$\alpha_i = \frac{\exp(h_i^\top u_c)}{\sum_i \exp(h_i^\top u_c)}, \tag{3.12}$$

$$v = \sum_i \alpha_i s_i, \tag{3.13}$$

where u_c is a comment-level context vector. Here, the latent vector v summarizes both textual and temporal information in a social media session. Both, the word-level and comment-level context vectors can be randomly initialized and learned in the training process [36].

3.3 Cyberbullying Detection with Time Interval Prediction To incorporate social information, we first feed the social information into a hidden layer to get the embedding γ. The final representation of a social media session is $h = [v, \gamma]$, the concatenation of v and γ. This feature then can be used to detect cyberbullying for session n:

$$p_n = \sigma(W_n h + b_n). \tag{3.14}$$

The first loss function is

$$\ell_1 = -\sum_{n=1}^{N} \log p_n. \tag{3.15}$$

As a result, HANCD is able to classify a social media session as bullying or not based on the text, time, and social information. Previous research indicates that cyberbullying on social media takes place across a stream of comments that are typically relatively close together in time-i.e., with shorter time intervals between adjacent comments [31]. As described below, we seek to simultaneously optimize cyberbullying detection and time interval prediction by using different weights to augment the efficacy of cyberbullying detection.

We first get the latent representation s_i for comment i, which summarizes all the available information of i, e.g. text, social information. Then the objective function of time interval prediction is defined as

$$\ell_2 = \sum_{n=1}^{N} \sum_{i=1}^{C} \frac{1}{2} ||A_n s_i + q_n - \Delta t_i||^2, \tag{3.16}$$

where A_n is the weight matrix and q_n is the bias term. The final weighted loss function of HANCD is

$$\ell = \beta_1 \ell_1 + \beta_2 \ell_2, \tag{3.17}$$

where β_1 and β_2 are the weights of cyberbullying detection and time interval prediction, respectively, in the overall function.

4 Experimental Evaluation

In this section, we evaluate the effectiveness of HANCD using the Instagram dataset[2] collected by Hosseinmardi et al. [15]. This dataset includes 2218 social media sessions, among which 1540 are labeled as *Normal* and 678 are labeled as *Bullying*. Each social media session has at least 15 comments and more than 40% of the comments by users other than the session owner have at least one negative word [15]. The average number of comments in a social media session is 71. We use 80% of the data for training and the remaining for testing, unless otherwise stated.

4.1 Baselines We compare HANCD with several baseline methods, including classification models - Naive Bayesian, Logistic Regression, Random Forest [3], XGBoost [4] and KNN trained on different sets of textual features. These features are count vectors, word-level TF-IDF vectors, N-Gram-level TF-IDF vectors, character-level TF-IDF vectors, word embeddings, and psychological features from Linguistic Inquiry Word Count (LIWC) [27]. Details of these features are provided below.

Count Vector It is a matrix of the Instagram dataset in which every row represents a social media session from the corpus, every column represents a term from the corpus, and every cell represents the frequency count of a particular term in a particular social media session.

TF-IDF Vectors A TF-IDF score represents the relative importance of a term in the social media session and the entire corpus. It can be generated at different levels of input tokens (words, characters, n-grams).

- Word-Level TF-IDF (Word TFIDF): Matrix representing TF-IDF scores of every word in different sessions.
- N-gram-level TF-IDF (N-gram TFIDF): N-grams are the combination of N words together. This matrix represents TF-IDF scores of N-grams.
- Character-Level TF-IDF (Char TFIDF): Matrix representing TF-IDF scores of character-level n-grams in the corpus.

LIWC We also perform psychometric analysis to obtain psychological features through LIWC, which counts words that belong to specific categories of feelings, personality traits, and psychological motives. Previous research shows that psychometric analysis can improve the performance of cyberbullying detection models [26].

Word Embedding This is a form of representing words and social media sessions using a dense vector representation. We use pre-trained word embedding[3] in the experiments.

In addition, we compare our model with several end-to-end deep learning models, including LSTM [12], CNN [19], and HAN [36], as well as some existing cyberbullying detection models, i.e., Xu et al. [35] and Soni and Singh [31]. We briefly introduce these two models here.

- Xu et al. This SVM classifier[4] is trained with several NLP features including unigrams, unigrams+bigrams, and POS-colored N-grams.
- Soni and Singh. This is the first computational method to model the temporal dynamics of commenting behavior as point processes. It defines several temporal features to distinguish characteristics between cyberbullying and regular social media sessions. We implemented this state-of-the-art cyberbullying detection model following the original design using several machine-learning models and report the best results.

4.2 Results Because the Instagram dataset is imbalanced, we report the F1 and AUC scores here for fair comparison. The results are shown in Table 1-2. We highlight the following observations:

- The proposed model (HANCD) gives the best F1 and AUC scores among all the models. Specifically, HANCD outperforms the best baseline model Soni & Singh by 5.8% and 5.1% w.r.t the F1 score and the AUC score respectively. Whereas the Soni and Singh model considers temporal, textual, and social features (e.g. #followers), it neither incorporates knowledge of the structure of social media sessions nor differentiates the importance of words and comments in different contexts. The Xu et al. model is based on textual features and relies on the existence of comments containing special keywords like "bully*" [35]. The results underscore the advantages of modeling the hierarchical structure of social media sessions and the two-level attention mechanism. HANCD also obtained better F1 and

[2] https://sites.google.com/site/cucybersafety/home/cyberbullying-detection-project/dataset

[3] https://nlp.stanford.edu/projects/glove/

[4] http://research.cs.wisc.edu/bullying/data.html

Table 1: Performance comparisons of different models (F1 score).

Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.476	0.521	0.501	0.479	0.559	0.236
Naive Bayesian	0.614	0.469	0.607	0.534	0.482	0.355
Logistic Regression	0.700	0.642	0.608	0.677	0.700	0.163
Random Forest	0.585	0.618	0.585	0.617	0.650	0.190
XGBoost	0.715	0.726	0.699	0.674	0.700	0.337
Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	
0.613	0.613	0.708	0.502	0.740	**0.783**	

Table 2: Performance comparisons of different models (AUC score).

Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.770	0.697	0.624	0.708	0.686	0.499
Naive Bayesian	0.706	0.815	0.797	0.786	0.622	0.525
Logistic Regression	0.812	0.825	0.827	0.830	0.776	0.629
Random Forest	0.788	0.804	0.788	0.781	0.743	0.544
XGBoost	0.838	0.828	0.831	0.840	0.772	0.621
Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	
0.791	0.781	0.805	0.513	0.810	**0.851**	

AUC scores than HAN, indicating the effectiveness of jointly optimizing cyberbullying detection and time interval prediction.

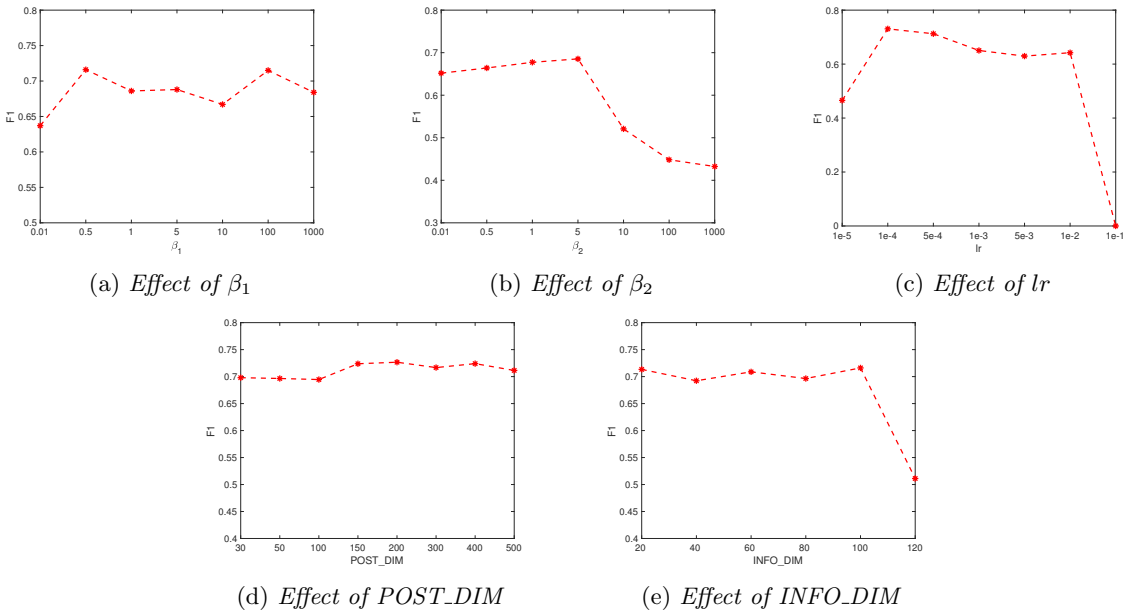
- Among the classification models, XGBoost, in most cases, has the best performance. As the winner of the Kaggle competitive data science platform, XGBoost (eXtreme Gradient Boosting) is designed for speed and performance and has recently dominated classification and regression predictive modeling problems.

- Among the different textual features, Count Vectors, TF-IDF vectors, and LIWC are considerably more effective than pre-trained word embeddings. The decreased effectiveness stems from the fact that it is difficult to train word embedding with social media data which is usually noisy, informal, and short. Social data can contain mistakes, misspelled words, established abbreviations such as *wtf* and *omfg*, and users' own abbreviations. Furthermore, a social media session is represented by the sum/average of all of its word embeddings. This can make the keywords that are useful for cyberbullying detection indistinguishable.

- Among the deep learning models, HAN outperforms LSTM and CNN on the identification of cyberbullying sessions. This suggests the importance of modeling the hierarchical structure of social media sessions and varying the level of attention to words and comments based on the context, especially when the dataset is small. Otherwise, deep learning models like LSTM and CNN can easily overfit the data.

4.3 Parameter Analysis The implementation of HANCD has five key parameters - β_1, β_2, lr, POST_DIM, and INFO_DIM, where β_1 is the weight of cyberbullying detection, β_2 is the weight of the time interval prediction task, lr is the learning rate, and POST_DIM and INFO_DIM are the embedding dimensions of words and social information, respectively. To investigate the sensitivity and effect of these parameters, we vary one parameter at a time and evaluate how it affects the overall cyberbullying detection performance. We vary the values of different parameters among different ranges due to their various numerical scales. We summarize the parameter study results (using the F1 score) in Figure 3.

As shown in Figure 3(a)-(b), HANCD is more sensitive to the weight of time interval prediction β_2 than the weight of cyberbullying detection. Specifically, as β_1 increases, HANCD puts more effort into cyberbullying detection, leading to a trend of slightly improved F1 score. HANCD is robust to changes in β_2 in the range [0.01,5]



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

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

Figure 3: Parameter sensitivity study (with 80% dataset for training)

where F1 improves as β_2 increases. However, the performance drops when $\beta_2 > 5$. The analysis of these two parameters indicates that time interval prediction indeed helps to improve the performance of cyberbullying detection. When HANCD overemphasizes the time interval prediction task, however, it diminishes the model performance. We have a similar curve for the learning rate in Figure 3(c). HANCD is robust to changes in lr in a large range but does not work well when lr is very large. As shown in Figure 3(d)-(e), HANCD is robust to changes in POST_DIM and INFO_DIM in the range [20,100]. In general, HANCD is not sensitive to the model parameters in a large range and, consequently, can be tuned for various application purposes.

5 Related Work

In this section, we review previous work related to computational models for cyberbullying detection and deep learning for text classification.

Cyberbullying Detection: Existing work in cyberbullying detection has focused on identifying characteristics of cyberbullying behavior using features from text [26, 34, 39, 28, 9, 35], social networks [16, 32], and other media sources such as images and videos [14, 22, 29, 5]. For example, Xu et al. [35] introduced several off-the-shelf tools such as Bag-of-Words models and LSA- and LDA-based representation learning to study bullying traces in social media platforms (i.e., Twitter) [35]. In doing so, they aimed to solve two major NLP tasks: text categorization, which distinguishes bullying traces from other social media posts, and role labeling, i.e., studying how a user's role evolves over time. Bellmore et al. used a dictionary including all words in a Twitter corpus to construct a frequency vector for each tweet and developed a text classifier to answer core questions about cyberbullying (*"Who, What, Why, Where, and When"*) [2]. In [13], Hosseinmardi et al. analyzed negative behaviors in the semi-anonymous question-answer pairs of the Ask.fm social network. Given the difficulty in obtaining a friendship-based social graph, the researchers constructed an interaction graph using information from the "likes" that comments received. The authors found that words connected to "cutting," "depress," "stressful," "sad," and "suicide" were prominent. Previous work, such as [13, 15], has also investigated cyberbullying on Instagram. Hosseinmardi and colleagues, for instance, applied LIWC to identify the primary categories of words used in cyberbullying social media sessions and identified specific image content (e.g., drugs) that were strongly related to cyberbullying instances [13].

Although cyberbullying represents, by definition, a harmful behavior that is repeated over time, few studies have explored cyberbullying from a temporal perspective. Soni and Singh [31] proposed a computational method for modeling the temporal dynamics of commenting behavior as point processes and defined several temporal features for distinguishing cyberbullying from regular social media sessions. Crucially, however, they did not consider the hierarchical structure of a social media session, the different importance of different

words, and how time interval prediction and text classification can jointly guide the learning process of cyberbullying detection.

Deep Learning for Text Classification Deep learning models that automatically extract context-sensitive features from raw text [23] have recently been successful for text classification. These models include the convolutional neural network (CNN) [19], the recurrent neural network (RNN) [24], the combination of CNN and RNN (RCNN) [37, 20], the CNN with attention mechanism [1, 36], the Bow-CNN model [17, 18], and the model for extreme multi-label text classification (XMTC). For example, Kim [19] applied the CNN model, originally used in computer vision [21], to text classification. CNN was found to better capture text semantics because it can identify discriminative phrases in a text using a max-pooling layer. Zhang et al. [38] applied a character-level CNN for text classification and achieved competitive performance. Socher et al. [30] used recursive neural networks for text classification and found this method to be efficient for constructing sentence representations. Due to limitations of CNN, Lai et al. [20] proposed RCNN, which can learn more precise text representations by taking advantage of both RNN and CNN. Tang et al. [33] used a hierarchical approach for sentiment classification. They first used CNN or LSTM to get latent representations of sentences and then used a bidirectional gated recurrent neural network to obtain vectors for documents. Because different words and sentences do not contribute equally to capture the meanings of documents, Yang et al. [36] proposed a hierarchical attention network (HAN) for document classification. To learn the weights of each word/sentence automatically, the authors added attention mechanisms, first proposed by [1], to both words and sentences in the bidirectional GRU. Their approach was found to outperform previous methods by a substantial margin.

In contrast to documents, social media sessions contain shorter, noisier and more informal tokens. On the other hand, they contain richer content in addition to text, such as timestamps and images. These properties enable HANCD to leverage the *multi-modal* information in social media sessions and jointly optimize cyberbullying detection and time interval prediction to augment the learning effectiveness of these two tasks.

6 Conclusions

In this paper, we proposed the Hierarchical Attention Networks for Cyberbullying Detection (HANCD) framework, which progressively builds a social media session by first aggregating words into comment vectors and then into session vectors. The proposed framework uses *context* to discover the relative importance of specific words and comments, rather than simply filter for words, devoid of context. To model the critical temporal information in a social media session, we jointly optimize the cyberbullying detection and time interval prediction tasks. By manipulating the weights of these two tasks, HANCD can capture their commonalities and differences to improve the performance of cyberbullying detection.

Because comments are posted at discrete points in time, future work can be directed to time series analysis, which models a sequence of discrete temporal data in order to extract meaningful statistics and identify important trends. Another vital direction for future research may be time series forecasting, which could be used to predict future cyberbullying instances from previously observed cases. Efforts to more accurately detect cyberbullying on social media remain a critical step toward building safer, more inclusive social interaction spaces.

Acknowledgements

This material is based upon work supported by the National Science Foundation (NSF) grant 1719722.

References

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

[2] Amy Bellmore, Angela J Calvin, Jun-Ming Xu, and Xiaojin Zhu. The five ws of bullying on twitter: who, what, why, where, and when. *Computers in human behavior*, 44:305–314, 2015.

[3] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *KDD*, pages 785–794. ACM, 2016.

[5] Lu Cheng, Jundong Li, Yasin N Silva, Deborah L Hall, and Huan Liu. Xbully: Cyberbullying detection within a multi-modal context. In *WSDM*, 2019.

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

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

[8] Maral Dadvar, de FMG Jong, Roeland Ordelman, and Dolf Trieschnigg. Improved cyberbullying detection using gender information. In *DIR*. University of Ghent, 2012.

[9] Harsh Dani, Jundong Li, and Huan Liu. Sentiment informed cyberbullying detection in social media. In *ECML PKDD*, pages 52–67. Springer, 2017.

[10] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *TiiS*, 2(3):18, 2012.

[11] L. Hackett. The annual bullying survey 2017. 2017.

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

[13] Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Amir Ghasemianlangroodi. Analyzing negative user behavior in a semi-anonymous social network. *CoRR abs*, 1404, 2014.

[14] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Analyzing labeled cyberbullying incidents on the instagram social network. In *SocInfo*, pages 49–66. Springer, 2015.

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

[16] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. Cyber bullying detection using social and textual analysis. In *SAM*, pages 3–6. ACM, 2014.

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

[18] Rie Johnson and Tong Zhang. Semi-supervised convolutional neural networks for text categorization via region embedding. In *NIPS*, pages 919–927, 2015.

[19] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[20] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *AAAI*, volume 333, pages 2267–2273, 2015.

[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[22] Homa Hosseinmardi Shaosong Li, Zhili Yang, Qin Lv, Rahat Ibn Rafiq Richard Han, and Shivakant Mishra. A comparison of common users across instagram and ask. fm to better understand cyberbullying. In *BDCloud*, pages 355–362. IEEE, 2014.

[23] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. In *SIGIR*, pages 115–124. ACM, 2017.

[24] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *ISCA*, 2010.

[25] Vinita Nahar, Xue Li, and Chaoyi Pang. An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering*, 3(5):238, 2013.

[26] Parma Nand, Rivindu Perera, and Abhijeet Kasture. " how bullying is this message?": A psychometric thermometer for bullying. In *COLING*, pages 695–706, 2016.

[27] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[28] Jing Qian, Mai ElSherief, Elizabeth M Belding, and William Yang Wang. Leveraging intra-user and inter-user representation learning for automated hate speech detection. *arXiv preprint arXiv:1804.03124*, 2018.

[29] Rahat Ibn Rafiq, Homa Hosseinmardi, Sabrina Arredondo Mattson, Richard Han, Qin Lv, and Shivakant Mishra. Analysis and detection of labeled cyberbullying instances in vine, a video-based social network. *SNAM*, 6(1):88, 2016.

[30] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. Parsing natural scenes and natural language with recursive neural networks. In *ICML*, pages 129–136, 2011.

[31] Devin Soni and Vivek Singh. Time reveals all wounds: Modeling temporal dynamics of cyberbullying sessions. In *ICWSM*, 2018.

[32] A Squicciarini, S Rajtmajer, Y Liu, and Christopher Griffin. Identification and characterization of cyberbullying dynamics in an online social network. In *ASONAM*, pages 280–285. ACM, 2015.

[33] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL (Volume 1: Long Papers)*, volume 1, pages 1555–1565, 2014.

[34] Phoey Lee Teh, Chi-Bin Cheng, and Weng Mun Chee. Identifying and categorising profane words in hate speech. In *ICCDA*, pages 65–69. ACM, 2018.

[35] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *NAACL HLT*, pages 656–666. ACL, 2012.

[36] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *NAACL HLT*, pages 1480–1489, 2016.

[37] Rui Zhang, Honglak Lee, and Dragomir Radev. Dependency sensitive convolutional neural networks for modeling sentences and documents. *arXiv preprint arXiv:1611.02361*, 2016.

[38] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, pages 649–657, 2015.

[39] Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. Content-driven detection of cyberbullying on the instagram social network. In *IJCAI*, pages 3952–3958, 2016.