Improving Cyberbullying Detection with User Interaction

Suyu Ge
Department of Electronic
Engineering, Tsinghua University
gesy17@mails.tsinghua.edu.cn

Lu Cheng
School of Computer Science and
Engineering, Arizona State University
lcheng35@asu.edu

Huan Liu
School of Computer Science and
Engineering, Arizona State University
huanliu@asu.edu

ABSTRACT

Cyberbullying, identified as intended and repeated online bullying behavior, has become increasingly prevalent in the past few decades. Despite the significant progress made thus far, the focus of most existing work on cyberbullying detection lies in the independent content analysis of different comments within a social media session. We argue that such leading notions of analysis suffer from three key limitations: they overlook the temporal correlations among different comments; they only consider the content within a single comment rather than the topic coherence across comments; they remain generic and exploit limited interactions between social media users. In this work, we observe that user comments in the same session may be inherently related, e.g., discussing similar topics, and their interaction may evolve over time. We also show that modeling such topic coherence and temporal interaction are critical to capture the repetitive characteristics of bullying behavior, thus leading to better predicting performance. To achieve the goal, we first construct a unified temporal graph for each social media session. Drawing on recent advances in graph neural network, we then propose a principled graph-based approach for modeling the temporal dynamics and topic coherence throughout user interactions. We empirically evaluate the effectiveness of our approach with the tasks of session-level bullying detection and comment-level case study. Our code is released to public. [1]

CCS CONCEPTS

• **Computing methodologies → Natural language processing**;
• **Social and professional topics → User characteristics**.

KEYWORDS

Cyberbullying Detection, User Interaction, Temporal Dynamics, Topic Coherence, Graph Attention Network
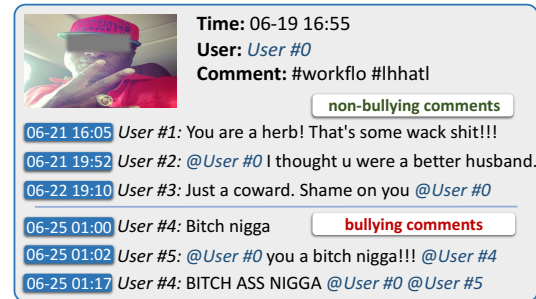
---

[1]https://github.com/gesy17/TGBully

Figure 1: An example of cyberbullying session consisting of the original post, user comments and posted time. The three comments in the upper part are non-bullying comments while those in the lower part are regarded as bullying comments.

1 INTRODUCTION

Cyberbullying is identified as a form of bullying or harassment through electronic means, which is typically characterized by posting rumors, threats, or pejorative labels on social media [43]. Due in part to the low cost and easy access to social networking sites, Cyberbullying has become increasingly prevalent in recent decades. According to the 2019 Youth Risk Behavior Surveillance System (Centers for Disease Control and Prevention), an estimated 15.7% of high school students were electronically bullied in the 12 months prior to the survey [17]. In response to the rapidly growing number of cyberbullying cases, numerous efforts (e.g., [13, 47, 50]) have been put into computational detection and mitigation of cyberbullying behaviors to tackle the socio-technical problem.

Despite the remarkable progress made towards computational cyberbullying detection, we argue that it is still essential to understand how users in a social media session interact with each other in order to mitigate the long-term impact of cyberbullying. Whereas offline bullying (or "traditional" bullying) relates to face-to-face interactions, cyberbullying typically takes place throughout a sequence of (intensive) user interactions in social media sites[9], as illustrated in Fig. 1. Modeling user interaction is important to cyberbullying detection for several reasons. First, one essential aspect that differs cyberbullying from offline bullying is its *persistence & repetition* [14, 53] characteristic. Decades of research in psychology [16] and social science [40] have shown that cyberbullying is carried out repeatedly against victims by one or more individuals. Studies of user interactions allow us to characterize the repetition by both content and temporal analysis. For example, recent findings [21, 41] suggest that (1) Most 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. Second, user interaction is a strong signal for detecting cyberbullying instances. Recent analysis of cyberbullying activity from an interdisciplinary research [21] manifests that the intensity of user activities within a bullying session is inherently different from that within a non-bullying session. Third, what makes cyberbullying detection particularly challenging compared to offline bullying is that it is not uncommon to see that users play multiple roles such as bully, victim, and witness in the same session [39, 48]. Therefore, modeling user interaction provides an alternative for the implicit monitor of users' roles over time.

However, modeling user interaction in a social media session confronts multi-faceted challenges: (1) *Sparsity*. Whereas a social media session is often involved with multiple interactions among numerous users, the interaction with a specific user can be very low. Further, most of the comments are directed towards the owner of the social media session, resulting in a "hub" where the owner degree is much larger than the degrees of users who post comments [29, 37]. Therefore, the user-user interaction graph constructed from reply relation can be sparse and less informative. (2) *Repetition*. Characterizing the repetition of cyberbullying behavior is particularly challenging because we need to satisfy at least two constraints simultaneously: similar in content and close in time [41]. Therefore, the desired models should be able to jointly capture content similarities and temporal correlations. (3) *User Characteristics*. Empirical findings within psychology identify that users' personality traits are important indicators of computer-mediated behaviors [1, 31]. However, it is rather difficult to discover such user characteristics solely based on the content in a session due to the extremely limited comments from an individual user.

To tackle these challenges, we propose **TGBully**, a **T**emporal-**G**raph-based cyber**bully**ing detection framework. TGBully is composed of three modules, i.e., a *semantic context modeling* module, a *temporal graph interaction learning* module and a *session classification* module. In the first module, we contrast the standard content analysis of a lump of comments with a hierarchical construction of a sequence of comments, each of which comprises a sequence of words. To incorporate characteristics of users who post these comments, we further extract users' historical comments, from which we infer their language behavior. The *temporal graph interaction learning* module implicitly constructs the user-user interaction graph by encoding the *topic coherence* and *temporal dynamics* of user comments in a graph attention network (GAT). In particular, we consider content similarities and time intervals between comments, and the user interactions are learned through the continuous update of the edge weights in the GAT. By doing so, TGBully models user interaction through the propagation of interactive signals in the constructed temporal graph. The *session classification* module aggregates the information representation learned from user interactions and performs the final session-level classification, i.e., determining whether a session is bullying or not. To summarize, the main contributions of this work are:

- We make novel observations of dynamic user interactions reflected by the inherently related and continuously evolved comments in a social media session. We show the importance of modeling user interactions in improving the performance of cyberbullying detection.
- We propose a principled and effective approach that models user interaction from two key aspects: temporal dynamics and topic coherence. This is achieved by a temporal graph constructed for each session and a GAT for learning the interactions in the graph.
- We evaluate TGBully on both session-level detection tasks and comment-level case study. Empirical results on two benchmark datasets validate the effectiveness of our approach and shed light on its practical implications for policy-makers.

2 RELATED WORK

Our work is closely related to cyberbullying detection and graph attention networks. Here, we briefly review the related literature.

2.1 Cyberbullying Detection

Previous work on cyberbullying detection typically builds their approaches upon natural language processing (NLP) methods [15, 35, 38]. Their goal is to identify aggressive and harassing mentions via text analysis. Hosseinmardi et al. [24] compared Linguistic Inquiry and Word Count (LIWC) values of different categories of words, they found that words connected to negative feelings such as "depress", "stressful", and "suicide" are prominent identifiers of cyberbullying. Dani et al. [13] added sentiment consistency features to a unigram model with TF-IDF values as content features.

More recent work highlighted the importance of temporal information for cyberbullying detection [7, 8, 41]. For example, Soni and Sign [41] characterized each comment on Instagram by an exponential time factor. They empirically evaluated a set of time-related features derived from the exponential factor and concluded several essential temporal traits of cyberbullying. Cheng et al. [7, 8] unified cyberbullying detection and time interval prediction in a multi-task learning framework. Built upon the hierarchical attention network, their approach aggregated contextualized comment representation for session classification meanwhile predicted time intervals from comment representations. In addition to text, numerous methods incorporated social information into cyberbullying detection [4, 28, 53]. By correlating cyberbullying behavior with online sociability, they primarily measured the power imbalance between users by the social network and interaction network. Ziems et al. [53] utilized the neighborhood overlap, verified status, number of followers, and recent tweets to indicate users' visibility among peers. Chatzakou et al. [4] extracted and compared numerous text, user, and network-based attributes. They found that compared with aggressors, bullies are less popular and participate in fewer online communities.

Prior research has put a large number of efforts into modeling user-generated contents but overlooked the interactions of users themselves within a session. Findings from various fields have evidenced the critical role user interactions play in cyberbullying detection. We thereby complement earlier work with the consideration of user interactions by jointly modeling the temporal dynamics and topic coherence in an established temporal graph.

2.2 Graph Attention Network

Recent years have witnessed the increasing popularity of graph neural networks (GNNs), which have become powerful tools to model real-world data in various domains [22, 32, 33]. Among the widely recognized GNNs, graph attention networks (GATs) [45] use self-attentional layers with automatically learned weights to attend over neighborhoods' features. Due to their prominent capability of learning salient interaction, a surge of attempts have been made to apply GATs to various text mining tasks. For instance, Hu et al. [26] proposed a Heterogeneous Graph Attention networks (HGAT) for semi-supervised short text classification. At its core, HGAT embeds heterogeneous information (topics, short texts, and entities) via a dual-level attention mechanism. Huang et al. [27] designed a novel target-dependent graph attention network (TD-GAT) to explicitly utilize the dependency relation among words for aspect-level sentiment classification.

Our work is more related to another line of research exploiting GAT for fact verification [34, 51, 52]. In both cases, characterizing interactions between facts (or users in our task) are central to the modeling process. For instance, Zhong et al. [51] constructed an XLNet-based word graph. Then they formed argument-level representations using a graph convolution network to propagate over the word graph. At last, GAT was incorporated to aggregate useful evidence for the verification task. Similarly, Zhou et al. [52] also employed GAT to perform fact reasoning where each graph node was constructed by concatenating a claim and the corresponding evidence. In these approaches, edge weights are automatically optimized during the propagation process. However, these approaches are not applicable to our task due to the unique challenges of the session-based cyberbullying detection [9]. Therefore, to make the most of GAT, we propose to explicitly introduce the two important bully-indicative features – temporal dynamics and topic coherence – into the edge weight calculation.

3 PROBLEM FORMULATION

We use the following information extracted from a social media session for cyberbullying detection:

- **Comment:** It is a sequence of comments posted by different users in a session. There are two types of comments in general: an initial comment posted by the owner and the subsequent comments. They comprise of the textual information of a session. We represent words in a comment as embeddings and analyze their contextual relation using neural networks.
- **Time:** It is the time stamp of each comment. Informed by previous work [7], we represent time by calculating the time intervals (in minutes) between comments and the initial comment.
- **User:** It denotes the username of each posted comment. Since a user may mention a previous username, we embed each username as a word and encode it into comment representation. We also add the username embedding before the first token of a comment to identify its poster.

- **History:** The historical comments of each user. We collect historical comments from previous sessions and concatenate them into a unified paragraph. [2]

Suppose that we have a corpus of sessions C and each session consists of n comments, denoted as $\mathcal{S} = \{c_1, c_2, ..., c_n\}$, where c_i is the ith comment in the session. Along with the sessions is the label set Y indicating whether a session is a bullying ($y = 1$) or non-bullying ($y = 0$) session. A user $u \in \mathcal{U}$ posts comment $c_i = \{w_1^c, w_2^c, ..., w_p^c\}$ with p words at time t_i. Besides, the user u also has a set of historical comments, and we concatenate them into one paragraph with q words as $h_u = \{w_1^h, w_2^h, ..., w_q^h\}, \forall h_u \in \mathcal{H}$. We then define the session-based cyberbullying detection task as:

Definition. *Given the corpus of sessions C, the associated label set Y, the group of users $\mathcal{U}$, and the historical comments $\mathcal{H}$, the goal of this work is to classify a social media session into the bullying/non-bullying category using the multi-modal information extracted from Comment, Time, User and History.*

4 TGBULLY: THE PROPOSED APPROACH

The goal of this work is to study user interaction in a session to improve the session-level cyberbullying detection. We herein propose **TGBully**, a principled framework consisting of three modules: (1) a *semantic context modeling* module that encodes each comment by considering both its textual content and user's language behavior reflected from her/his historical comments, (2) a *temporal graph interaction learning* module that constructs a temporal graph and models the dynamic user interaction with a bully-featured GAT. The proposed GAT jointly captures *topic coherence* and *temporal dynamics* in the modeling process; and (3) a *session classification* module that attentively aggregates information from user interaction into a session representation, based on which it then classifies the session into a bullying/non-bullying session. The overall framework of TGBully is illustrated in Fig. 2.

4.1 Semantic Context Modeling

Towards the semantic understanding of texts in a session, the *semantic context modeling* module utilizes a hierarchical pipeline [7, 49] to form two types of contextualized representations. One is comment representation, which uncovers user intent and emotion through their utterances in the current session. The other is history representation, which encodes user characteristics, e.g., personalities and language patterns reflected from their historical comments.

4.1.1 Comment Encoding. The comment encoding mainly consists of four layers. The first layer is the word embedding, which transforms a p-word comment $c = \{w_1^c, w_2^c, ..., w_p^c\}$ to a sequence of high dimensional vectors $\{\mathbf{w}_1^c, \mathbf{w}_2^c, ..., \mathbf{w}_p^c\}$ via an embedding matrix. The second is a word-level gated recurrent units (GRU) [10] layer. It aims to encode long-term contextual dependency among words. It has been shown that GRU achieves better performance than typical LSTM [23] on small datasets [20], therefore, we use a stack of bidirectional GRUs to model word sequences from both forward and backward directions. The contextualized representation

---

[2]We omit their relative temporal order due to the lack of such information.
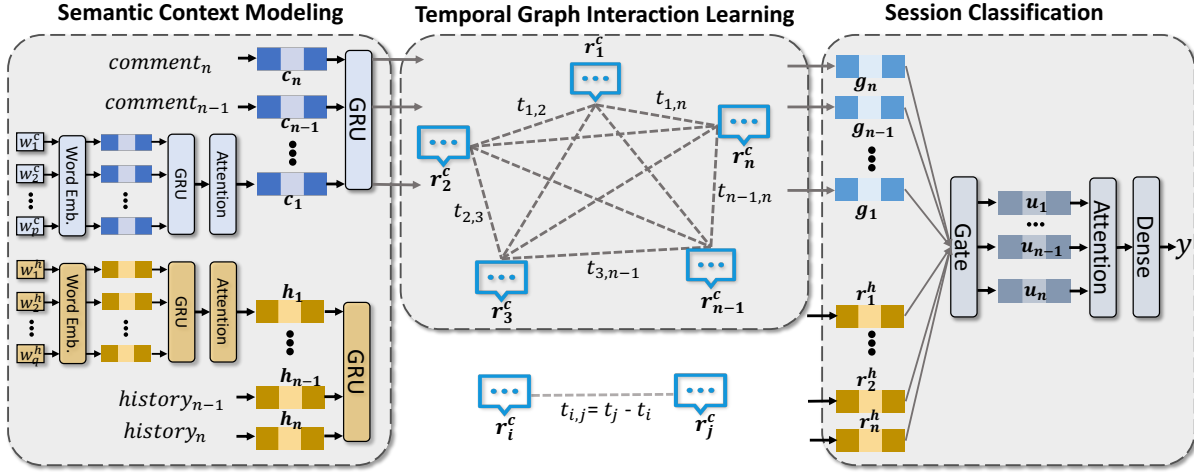
Figure 2: Description of the overall framework. TGBully consists of three modules: The *semantic context modeling* module (left) hierarchically constructs comment sequences as c (blue) and encodes users' language behavior via the representation of historical comments h (yellow). The *temporal graph interaction learning* module (middle) implicitly encodes the *topic coherence* and *temporal dynamics* in the temporal interaction graph where nodes represent user comments and edges represent time intervals between two comments. The *session classification* module (right) first aggregates the information learned from user interactions into user representation u (grey). It then performs the session-level classification by aggregating all user representations within the session into the session representation.

$\mathbf{r}_i^w$ of the ith word $\mathbf{w}_i^c$ is calculated as:

$$\mathbf{r}_i^w = [\overrightarrow{GRU}(\mathbf{w}_i^c), \overleftarrow{GRU}(\mathbf{w}_i^c)]. \tag{1}$$

However, the words in a comment are not equally significant to cyberbullying. To automatically select informative words from a comment, we herein employ word-level attention as the third layer. For the jth comment, the aggregated comment representation $\mathbf{c}_j$ is learned as:

$$\alpha_i^w = \frac{\exp(\tanh(\mathbf{v}_w^T \times \mathbf{r}_i^w + b_w))}{\Sigma_{k=1}^p \exp(\tanh(\mathbf{v}_w^T \times \mathbf{r}_k^w + b_w))}, \tag{2}$$

$$\mathbf{c}_j = \sum_{i=1}^p \alpha_i^w \mathbf{r}_i^w, \tag{3}$$

where $\mathbf{v}_w$ and b_w are the parameters of the attention network. α_i^w is the attention weight of the ith word. The final layer of this sub-module consists of comment-level bidirectional GRUs that model comments in a temporal order. Then the bi-GRU output representation of the jth comment is formulated as:

$$\mathbf{r}_j^c = [\overrightarrow{GRU}(\mathbf{c}_j), \overleftarrow{GRU}(\mathbf{c}_j)]. \tag{4}$$

4.1.2 History Encoding. As the number of comments posted by a user is rather limited, the goal of history encoding is to better capture users' language behavior and personality traits mirrored by their historical comments. Survey studies find that bullying behavior has strong connections with both hostility and neuroticism [3, 11]. Neurotic users are more emotional and sometimes even more irrational, thereby, they are more likely to engage in cyberbullying activities [1]. Further, it has been advocated that the use of historical information in social media is effective for personality

prediction [19]. We first associate the learned user characteristics with corresponding comments in the current session. We then use the same semantic context modeling approach described above for user history modeling. In particular, given n users' history in a session as inputs, where each history comment comprises q words, the output contextualized history sequence is $\{\mathbf{r}_1^h, \mathbf{r}_2^h, ..., \mathbf{r}_n^h\}$.

4.2 Interaction Learning in Temporal Graph

The *Temporal graph interaction learning* module seeks to learn user interaction and capture the repetitive characteristics of cyberbullying behavior. To accomplish this goal, the module first constructs a temporal interaction graph for each social media session. Since GAT can selectively aggregate principle neighbor information via its attentional layers, we use it to implicitly encodes the *topic coherence* and *temporal dynamics* between different users in the graph.

We begin by illustrating how the *temporal graph interaction learning* module can effectively address the two major challenges for modeling user interactions in a session. To recall, as most comments are directed towards the owner of the session, the first challenge is that social graph based on reply and friendship relation can be sparse; the second challenge relates to the unique characteristic of cyberbullying – repetition: content similarity and time proximity. We argue that these challenges can be effectively tackled by jointly modeling the *topic coherence* and the *temporal dynamics* in a session. The two aspects improve our understandings of the discussed topics and temporal correlation across time.

We first construct the temporal graph where the nodes are user comments and the edges are time intervals between comments. Each comment in our graph is inherently connected, consequently, they will not be bounded by sparse and uneven reply relationships between their users. We then propose to encode the topic

coherence and temporal dynamics in the temporal graph jointly in a unified framework. The affinity level of two neighbor user comments, namely, their content similarity and time proximity, is automatically determined and propagated through the edges of the temporal graph. At its core, the *temporal graph interaction learning* module leverages GATs to aggregate neighbor information (i.e., comments) according to their affinity levels.

4.2.1 Temporal Graph. The user interaction is implicitly learned by capturing their comment interaction in the temporal graph. Specifically, we first transform the session into a fully connected graph, where we establish connectivity between two arbitrary comments. In the temporal graph, comments are viewed as graph nodes and time intervals between them are viewed as edges. Informed by recent advances in GNNs [26, 27], we incorporate a graph attention neural network [46] to automatically aggregate information from neighbor nodes to the central node. Different significance levels of neighbor nodes are weighted by their edge weights. The GAT network encodes the temporal-level and topic-level relatedness between the central node and neighbors by introducing them into the calculation of edge weights. Then it propagates interactive signals to the central nodes by aggregating neighbors nodes according to their edge weights. Additionally, we consider self-propagation to reserve the original content information of each node. Since our temporal graph is already fully-connected, instead of stacking multiple graph layers, we only perform one-hop graph aggregation[3]. For the jth comment in the temporal graph, we denote its node vector as $\mathbf{g}_j$, which is calculated as:

$$\mathbf{g}_j = \sum_{\mathbf{r}_k^c \in \{N_j \cup \mathbf{r}_j^c\}} \pi(\mathbf{r}_k^c, \mathbf{r}_j^c, t_{k,j}) \mathbf{W}_c \mathbf{r}_k^c, \qquad (5)$$

where N_j is the set of jth comment's neighbor nodes, $\mathbf{W}_c$ is the transformation matrix of the graph network. $t_{k,j} = t_j - t_k$ is the time interval between two comments. $\pi(\mathbf{r}_k^c, \mathbf{r}_j^c, t_{k,j})$ denotes the weight of edge connecting node $\mathbf{r}_k^c$ and $\mathbf{r}_j^c$. It is a decaying factor that controls the relative contribution each neighbor makes to the central node. We detail the implementation of $\pi(\mathbf{r}_k^c, \mathbf{r}_j^c, t_{k,j})$ later.

4.2.2 Topic coherence. Semantic relatedness between comments in a session can reveal users' reactions and attitudes towards the commenting behavior of other users. For instance, a group of users may repeat the offensive terms used in a previous post to humiliate the victim. Such language repetition reflects and reinforces their aggressive motivations and potential power imbalance between bullies and victims. Crucially, capturing such topic-level repetition is useful for cyberbullying detection.

Our proposed approach is based on an important observation: In the embedding space, two comments usually locate in close proximity if they share similar content or they are coherent in the topical discussion. To measure such proximity, we compute the embedding similarity π_{time} of comments by taking dot products of their transformed node vectors in the graph. Formally,

$$\pi_{topic}(\mathbf{r}_k^c, \mathbf{r}_j^c) = (\mathbf{r}_k^c \mathbf{W}_o)^T \mathbf{r}_j^c, \qquad (6)$$

where $\mathbf{r}_k^c$ and $\mathbf{r}_j^c$ are two nodes representing two arbitrary user comments. $\mathbf{W}_o$ is the graph transformation matrix.

---

[3]Empirical experimental results also prioritize the one-hop graph aggregation.

4.2.3 Temporal dynamics. In addition to the topic coherence, temporal proximity is also an important feature for repetition. Instead of being immediate one-off responses, previous research discovered that cyberbullying sessions receive a more steady stream of comments that are consecutive in time [41]. It is also common to observe bursts of comments in which several bully users gang up on a victim in a bullying session [2]. Both the steady stream of replies and a sudden burst of comments can be recognized by capturing intervals between each chronological pair of comments. Thus, we further integrate the time-lapse factor into the constructed temporal graph. Specifically, we calculate the time interval of comments and project it into the graph embedding space as below:

$$\pi_{time}(t_{k,j}) = W_t t_{k,j} = W_t(t_j - t_k), \qquad (7)$$

where W_t is the time transformation coefficient and $t_{k,j}$ denotes the time interval.

4.2.4 Joint modeling. The ultimate goal of this work is to improve cyberbullying detection with user interaction modeling. Here, we detail how to learn the topic coherence and temporal dynamics of user interaction in a unified framework such that the cyberbullying detection performance can be optimized. One empirical observation is that the affinity level of two comments should be enhanced when they: (1) discuss consistent and coherent topics; and (2) appear closely in the time scope. Therefore, we integrate the topical factor π_{topic} and the temporal factor π_{time} into the learning of the edge weight $\pi(\mathbf{r}_k^c, \mathbf{r}_j^c, t_{k,j})$:

$$\pi(\mathbf{r}_k^c, \mathbf{r}_j^c, t_{k,j}) = \tanh(\pi_{topic}(\mathbf{r}_k^c, \mathbf{r}_j^c) + \pi_{time}(t_{k,j})). \qquad (8)$$

Note that for simplicity, we only utilize dot products of two nodes and add the results to the transformed time intervals. Further explorations of the edge weight calculation, such as alternative forms of attention mechanism and different ways of combining the two components, are left as future work. By modeling user interactions via the bully-featured GAT – GAT that encodes topic coherence and temporal dynamic jointly, we mimic the information flow in the social network to capture the evolution of the discussed topics and the repetitive characteristics of cyberbullying.

4.3 Session Classification

The *session classification* module aims to classify a social media session into a bullying or non-bullying session. In particular, it first aggregates user's history representation $\mathbf{r}^h$ and graph representation $\mathbf{g}$ into a unified user representation $\mathbf{u}$. Then it incorporates a user-level attention network to select informative users to form the final session representation $\mathbf{s}$, which is used for the final classification task.

While examining users' historical comments, an important notion is that some users show similar language behavior over time. For example, a user who constantly bullies others comments with offensive remarks whenever s/he is involved in an online discussion. By contrast, a user who occasionally bullies others may make offensive comments only once. Under such circumstance, her/his historical comments may not be informative enough to recognize the user's bullying tendency in the current session. Considering the relative importance of historical information for each user, we here use a gating mechanism to automatically weigh the relative

Table 1: Statistics of the Instagram and Vine dataset.

Datasets	# Sessions	# Bully	# Non-bully	# Comments
Instagram	2,218	678	1,540	155,260
Vine	970	304	666	78,250

importance of comments in the current session and user history. For the jth comment in session s, the merged user representation $\mathbf{u}_j$ is formulated as below:

$$\mathbf{u}_j = \beta_j \odot \mathbf{r}_j^h + (1 - \beta_j) \odot \mathbf{g}_j, \tag{9}$$

$$\beta_j = \sigma(\mathbf{W}_h \mathbf{r}_j^h + \mathbf{W}_g \mathbf{g}_j + b_c), \tag{10}$$

where $\mathbf{W}_h$, $\mathbf{W}_g$ and b_c are trainable parameters, σ is the sigmoid activation function, and β_j is the controlling weight. Since different user comments are not equally relevant to cyberbullying detection, we aggregate them using a user-level attention mechanism. The final session representation $\mathbf{s}$ is learned as:

$$\alpha_j^s = \frac{\exp(\tanh(\mathbf{v}_s^T \times \mathbf{u}_j + b_s))}{\Sigma_{k=1}^n \exp(\tanh(\mathbf{v}_s^T \times \mathbf{u}_k + b_s))}, \tag{11}$$

$$\mathbf{s} = \sum_{j=1}^n \alpha_j^s \mathbf{u}_j, \tag{12}$$

where $\mathbf{v}_s$ and b_s are parameters of the attention network, and α_j^s is the attention weight of the jth user. For session-level cyberbullying detection, we feed $\mathbf{s}$ into a single-layer dense network and predict its label y, where $y = 1$ indicates a bullying case.

5 EXPERIMENTS

In this section, we begin by an introduction of the benchmark datasets for session-based cyberbullying detection. We then detail the experimental setup including the baselines, evaluation metrics , and implementations. The focus of this section lies in answering the following questions:

- **RQ1:** How does TGBully fare against existing cyberbullying detection models w.r.t. predictive performance?
- **RQ2:** How does each module of TGBully contribute to the overall model performance?
- **RQ3:** As real-world cyberbullying datasets are often imbalanced, how does the data imputation techniques (e.g., oversampling) influence the session-level cyberbullying detection?
- **RQ4:** Apart from the session-level prediction, can TGBully effectively recognize cyberbullying instances at the comment-level?

5.1 Data

We conduct experiments on two benchmark datasets collected from Instagram [25] and Vine [36]. Both datasets were crawled using a snowball sampling method and manually annotated via the crowd-sourcing platform CrowdFlower. Definitions of cyberbullying were provided for five annotators at the beginning of the labeling process. A session is considered a cyberbullying session if no less than three people labeled it as bully. Sessions were removed from the dataset

if they contain less than 15 comments. Statistics of the two datasets are illustrated in Table 1.

Instagram: Instagram[4] is a media-based social network where users post, like, and comment on images. It has been reported as one of the top five networks with the highest percentage of users reporting experience of cyberbullying [42]. For each media-session, this dataset provides the labeled image content, the textual content of comments and the arrival time of comments. In addition, the user of each comment is identified along with social network properties (user IDs of their likes, followers, and followings) and historical information (comments and likes in earlier sessions). In total, this dataset contains 2,218 media sessions with an average number of 72 comments in each session. For each comment, the average number of words is 12.

Vine: Vine[5] is a mobile application that allows users to upload, like, and comment on six-second looping videos. For each video session, along with the video content, this dataset also records content, time, and associated user IDs of each comment. However, different from the Instagram dataset, this dataset doesn't provide users' historical comments. Instead, it provides user profile information such as location, profile description, and social network properties (user IDs of their followers and followings). In implementation, we substitute user historical comments in the history encoding module with their profile description. On average, there are 81 comments per session and 8 words per comment.

5.2 Experimental Setup

For each dataset, we randomly sample 80% sessions for training, 10% for validation, and 10% for testing. We assume each session is independent and do not consider cross-session correlation. Session length is set as 140 and sentence length is 30. We use the pretrained 400-dim Twitter embedding [18] as the textual input. The sentence-level Bi-GRU has 2×32 hidden states and the session-level Bi-GRU has 2×64. The dropout rate is 0.2. We repeat each experiment 5 times and report their average performances and standard deviations.

Evaluation Metrics. We use three widely recognized evaluation metrics in cyberbullying detection tasks, i.e., recall, micro F1, and AUC. We pay special attention to recall scores as the frequency of bullying sessions in the real-world is typically much lower compared to that of non-bullying sessions. Further, we report the micro F1 score for the interest of cyberbullying instances.

Baselines. We compare TGBully with three categories of baselines: *feature-based models* (LR, SVM and XGBoost), *neural text classification models* (CNN, LSTM, SelfAtt and HAN), and *cyberbullying detection models* (SICD, Soni & Sign and HANCD). Specifically,

- **LR.** Logistic Regression is a commonly used binary classifier. It estimates and optimizes parameters of a logistic model.
- **SVM** [12]. Support Vector Machine (SVM) is a non-probabilistic binary classifier. It maximizes the functional margin between positive and negative classes.
- **XGBoost** [6]. It is a highly scalable and efficient tree boosting system. The decision tree-based algorithm uses several regularization and parallelization techniques to improve overall performances.

---

[4]https://www.instagram.com/
[5]https://vine.co/

Table 2: Evaluation results of different methods on the Instagram dataset.

	50%			100%		
	Recall	F1	AUC	Recall	F1	AUC
LR	52.74±4.48	64.05±2.93	73.89±1.94	59.06±3.79	69.12±2.86	76.97±1.75
SVM	55.93±4.80	66.31±3.36	75.28±2.20	59.32±3.85	69.03±2.94	76.94±1.77
XGBoost	69.05±5.71	72.25±1.59	79.63±1.16	71.55±3.64	75.25±2.09	81.68±1.57
CNN	67.41±4.53	66.67±0.95	83.39±1.15	70.51±5.84	67.12±1.53	84.00±1.37
LSTM	64.18±2.24	66.63±1.37	82.22±2.45	70.01±8.37	67.42±1.60	83.87±1.59
SelfAtt	66.73±6.24	72.17±5.40	90.30±1.92	74.05±4.88	74.45±3.00	89.32±1.37
HAN	71.69±4.24	75.01±1.74	90.07±1.40	78.21±7.02	76.99±1.99	91.16±1.02
SICD	62.53±4.53	69.71±2.38	79.92±1.27	76.87±3.79	70.55±1.29	86.24±1.16
Soni & Sign	65.67±5.32	70.68±1.87	84.19±0.82	72.06±4.01	72.59±1.43	88.06±1.24
HANCD	**78.39±3.69**	76.40±2.17	91.63±2.13	76.91±4.56	77.55±2.14	91.32±1.29
TGBully	77.64±4.70	**78.12±2.04**	**91.81±0.98**	**82.57±3.10**	**80.97±2.03**	**92.91±1.30**

Table 3: Evaluation results of different methods on the Vine dataset.

	50%			100%		
	Recall	F1	AUC	Recall	F1	AUC
LR	47.42±5.32	55.92±3.21	63.72±2.25	50.00±4.06	59.65±3.31	65.85±2.30
SVM	47.69±4.98	57.22±3.17	64.31±2.42	48.48±3.28	60.38±2.76	66.81±1.78
XGBoost	63.11±4.95	61.32±2.89	69.82±1.88	64.54±3.77	64.00±3.12	70.75±1.42
CNN	60.50±6.90	53.97±2.43	67.07±0.70	61.61±2.27	56.56±1.40	72.65±2.01
LSTM	60.12±8.23	57.76±1.78	73.64±4.17	62.76±9.77	58.91±2.61	73.04±5.05
SelfAtt	53.00±7.29	59.46±4.33	81.77±1.60	59.83±6.56	64.49±3.18	81.39±1.74
HAN	71.11±6.28	66.54±2.11	82.13±1.77	70.04±2.93	68.10±2.26	82.32±1.18
SICD	61.72±5.24	58.43±5.75	76.74±1.64	66.82±3.92	62.76±2.38	78.56±2.52
Soni & Sign	64.50±5.98	60.47±2.43	80.27±1.35	64.48±5.24	63.49±1.97	80.33±3.27
HANCD	68.39±3.94	67.10±3.08	**83.62±2.08**	**71.05±4.66**	68.01±2.86	82.76±3.13
TGBully	**75.17±6.52**	**68.64±3.78**	83.54±3.35	70.18±6.65	**69.35±2.04**	**83.05±2.67**

- **CNN** [30]. The original implementation uses convolutional neural network (CNN) to classify single sentences. We apply it at comment level and average comment representations to form session representations.
- **LSTM** [23]. It designs long short-term memory units for context modeling. Similarly, we use it to model comments and classify sessions with averaged comment representations.
- **SelfAtt** [44]. It incorporates multi-head self-attention mechanism to model context dependency in sentences. In addition to word-level attention, we apply a similar comment-level attention to model relation between different comments.
- **HAN** [49]. A hierarchical bi-directional GRU framework is used to model sessions at both word level and comment level.
- **SICD** [13]. A sentiment enhanced feature-based cyberbullying model. User-post relationship and sentiment similarity between posts are used to guide the prediction.
- **Soni & Sign** [41]. A cyberbullying model that exploits temporal information and activity levels in sessions. Textual, social and temporal features are simultaneously utilized for classification.

- **HANCD** [7]. A multi-task cyberbullying model based on HAN. It adds a time interval prediction subtask to exploit temporal dynamics within sessions.

For LR, SVM, and XGBoost, we use word-level tf-idf of user comments as textual features and number of likes and follows as social features. We use the same word embedding for TGBully and all *neural text classification models*. Hyperparameters are tuned to optimize performance. We set the parameters as reported in the original papers for all *cyberbullying detection models*. One of the challenges in cyberbullying detection is the data scarcity problem. Therefore, to investigate how much these models rely on the labeled data, we also construct another training set by randomly sampling half of the original training data. We report results for both cases with half and full training data.

5.3 Performance Evaluation

To answer the question in **RQ1**, we compare the performances of all baseline methods with TGBully's on the Instagram and Vine datasets. The results are reported in Table 2 and 3.
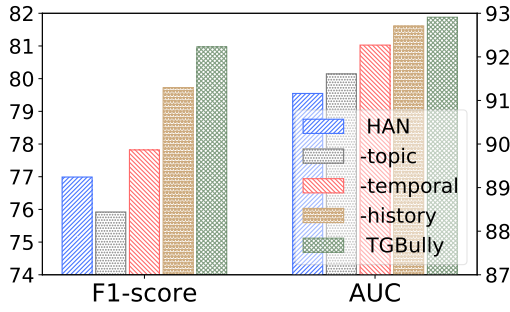
, Suyu Ge, Lu Cheng, and Huan Liu

Figure 3: TGBully and its variants on the Instagram dataset.

Comparing TGBully with other cyberbullying baselines that also consider temporal analysis, we observe that: (1) TGBully consistently achieves the best performances in most cases with different datasets and training ratios. This is in part because our approach effectively models user interactions in a graph attention neural network with a focus on temporal dynamics and topic coherence of a session. The joint modeling of the two bully-indicative aspects further leads to better recognition of the repetitive characteristics of cyberbullying behavior. The improvement also comes from the modeling of user personality traits and language behavior from users' historical comments while learning user interaction. (2) Large variations are observed in the results of all methods, especially when there is less training data. This reveals the necessity of using larger datasets in order to further improve the performance of cyberbullying detection. This result is also evidenced by recent findings in [9], which states that TGBully improves the performance by a smaller margin over baselines with the Vine dataset as Vine has fewer samples and we do not have access to the historical comments of its users.

By comparing different methods' performances, we also conclude that: (1) Neural models achieve substantially better performances than feature-based methods, especially w.r.t. Recall and AUC. Rather than representing session information with raw features, neural networks incorporate pretrained word embeddings that have rich semantic meanings and are fine-tuned during the training process. Another important reason is that neural models can better capture context dependency between words, leading to high-quality session representations, even with highly imbalanced datasets. (2) Another observation is that explicitly modeling the hierarchical structure can further improve the performance. Hierarchical neural models (SelfAtt, HAN, HANCD, and TGBully) outperform standard neural models (CNN and LSTM) partly because they better capture the interaction between different comments and consider their relative importance in a session. (3) Introducing additional temporal and sentiment information enhances both *feature-based models* and *text classification models*. For instance, SICD and Soni & Sign introduce sentiment and temporal modeling respectively, and both methods outperform LR and SVM methods by a large margin. Similarly, HANCD and TGBully improve over HAN model by modeling the temporal dynamics in a session.

5.4 Ablation Study

This set of experiments seeks to examine the impact of each module of TGBully on the overall performance (**RQ2**). Here, we use the Instagram dataset as a working example because similar findings can be observed with Vine. Results are reported in Fig. 3. We use the leave-one-out scheme to evaluate the impact of each of the three aspects – topic coherence, temporal dynamics, and historical comments – on the performance of TGBully. We also include HAN for reference because TGBully will be converted to HAN when the modeling of the three aspects is removed simultaneously.

From Fig. 3, we observe that all three aspects are important to session-level cyberbullying detection. In particular, the performance of TGBully exacerbates the most when topic coherence is removed from the learning process. Therefore, it is crucial to model semantic interaction between comments. By explicitly connecting the edge weights to the modeling of topic coherence, TGBully enforces the correlations between semantically related comments meanwhile reduces the noise induced by unrelated comments. Modeling temporal dynamics also brings substantial performance improvement. Results show that transforming time intervals between comments into edge weights in the temporal graph can help capture the evolving characteristics of sessions. Moreover, the consideration of historical comments also enhances the performance of TGBully, indicating the effectiveness of modeling users' personalities and language behavior for cyberbullying detection. By combining all three aspects together, TGBully can achieve the best performance overall.

5.5 Influence of Data Imbalance

Cyberbullying datasets are typically imbalanced. A common solution to extremely skewed data is data imputation, e.g., data oversampling. To further examine how the common data imputation techniques impact the model performance (**RQ3**), we artificially create a balanced dataset by oversampling data from the minority class. Specifically, We use SMOTE [5] to synthesize new bullying examples in the feature space and illustrate the performance comparisons of various methods in Table 4. We observe that oversampling brings converse effects to different methods. It improves performances of LR, SVM, XGBoost, and LSTM, and decreases that of HAN, HANCD, and TGBully. Two reasons may account for this. First, methods such as LR, SVM, XGBoost, and LSTM only exploit semantic information of comments. Since they capture information from a single modality, their classifiers may benefit from a more balanced data distribution in the feature space. Second, other approaches (HAN, HANCD, and TGBully) exploit multi-model information, e.g., comment hierarchy, temporal dynamics , and user history. Nevertheless, synthetic data assumes independence between different modalities and ignores their correlation. Thus, they may add extra noise to models and harm their performances.

5.6 Case Study

To answer **RQ4**, we investigate the capability of our model to detect bullying instances at the comment-level. Here, we use the comment-level attention weights that indicate the relative importance of each user comment for classification. The underlying assumption is that comments with higher attention weights are more likely to be bullying instances. To validate this, we ask three graduate

(a) A bullying example.

(b) A non-bullying example.

Figure 4: Visualization of various types of weights within a session. Shades of purple and blue denote user-level and word-level attention weights respectively. Green mesh grids denote graph weights between users (i.e., their comments), where grid at the ith row and j^{th} column represents weight w_{ij}. Darker color represents higher weights.

Table 4: Performance comparison of different methods with oversampled dataset.

	Oversampled			Original		
	Recall	F1	AUC	Recall	F1	AUC
LR	**69.50**	**71.27**	**78.81**	59.06	69.12	76.97
SVM	**60.99**	**69.08**	**77.06**	59.32	69.03	76.94
XGBoost	**73.05**	**75.74**	**81.90**	71.55	75.25	81.68
LSTM	**70.44**	**68.75**	**85.25**	70.01	67.42	83.87
HAN	77.93	76.75	**91.47**	**78.21**	**76.99**	91.16
HANCD	**77.36**	77.30	90.66	76.91	**77.55**	**91.32**
TGBully	79.02	78.67	91.81	**82.57**	**80.97**	**92.91**

students (two from computer science and one from psychology) to manually annotate the user comments in a total number of 100 sessions. The final label of each comment is determined by a majority vote. Among the 100 sessions with comment-level labels, we randomly select a cyberbullying session and a non-bullying session for illustration. We aim to visually testify the effectiveness of our method and answer **RQ4**. The results can be seen in Fig. 4. In addition to the user-level attention weights, we also present graph weights and word-level attention weights in TGBully.

From Fig. 4, we make the following observations: (1) TGBully can effectively recognize bullying users (comments) based on the learned attention weights. As illustrated in Fig. 4(a), as the 2nd, 4th and 5th comments contain humiliating words, they are predicted with higher weights indicating a high probability of being bullying comments. We found that the results are lined with ground-truth labels. In contrast, comments in Fig. 4(b) are non-bullying and their

attention weights are more evenly distributed. (2) TGBully encodes temporal dynamics and topic coherence between comments via the GAT. Comparing Fig. 4(a) and Fig. 4(b), we find that graph weights of bullying sessions are relatively dense and concentrate on specific nodes. For instance, the 2nd, 4th and 5th comments convey similar semantic meaning. Accordingly, their interactions are highlighted by the graph weights. Meanwhile, apart from being coherent in topic, the 4th and 5th comments are also consecutive in time. As a result, their correlations are more obvious than that between the 2nd and 4th comments. By contrast, weights of the non-bullying session largely concentrate on diagonals, indicating that they tend to focus on self interactions of comments and reflect limited cross-comment interactions. Despite the overlapped words and temporal proximity between comments, our method doesn't capture significant cross-comment interaction. This may happen because the GAT model only aims to capture specific bully-related interaction. (3) TGBully effectively selects offensive words. For instance, words such as "terrible", "ni**a", and "bit*h" in Fig. 4(b) have larger attention weights. We also notice that TGBully recognizes some salient words in a non-bullying session, which indicates it is discriminative to bullying words.

6 CONCLUSION AND FUTURE WORK

In this work, we make a novel observation that in a social media session, user interaction is embodied by the relations between their posted comments, and such interaction dynamically evolves with time. We further show that the topic coherence and temporal dynamics in the modeling of user interactions jointly contribute to capturing the repetitive characteristic of cyberbullying behavior. The core contribution of this work is the proposed TGBully, a

temporal graph-based cyberbullying detection framework with a focus on user interaction. TGBully consists of three modules. The first *semantic context modeling* module hierarchically constructs user comments from sequences of words and infers users' language behavior and personality from their historical comments. The second *temporal graph interaction learning* module first constructs the user interaction graph. It then encodes the topic coherence and temporal proximity through a bully-featured GAT. The third *session classification* module aggregates the information learned from user interaction in a session and performs the final session-level classification. We empirically evaluate the effectiveness of our approach with the tasks of session-level bullying detection and comment-level case study.

Our work opens several key avenues for future work. One such direction is to combine the explicit reply/social network with the introduced implicit temporal graph to together model the user interaction in a social media session. Additionally, we can incorporate other important aspects of cyberbullying into the temporal interaction graph. For instance, we can consider different roles of users (e.g., bullies and victims) and study how the role of a user evolves over time in the graph modeling process. Further investigations are needed to apply TGBully to detecting comment-level cyberbullying instances at scale.

7 ETHICS STATEMENT

The common understanding of the importance of data privacy and security is shared by the authors. To secure user privacy, we replaced usernames and ids (include their mentioning in other users' comments) with numerical indices before the annotating and prepossessing process. The insulting or offensive terms and the figures used in this paper are for illustration only, they do not represent the ethical attitude of the authors.

ACKNOWLEDGMENTS

REFERENCES

[1] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. 2016. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior* 63 (2016), 433–443.

[2] Ali Al Mazari. 2013. Cyber-bullying taxonomies: Definition, forms, consequences and mitigation strategies. In *2013 5th International Conference on Computer Science and Information Technology*. IEEE, 126–133.

[3] Osman Tolga Arıcak. 2009. Psychiatric symptomatology as a predictor of cyberbullying among university students. *Eurasian Journal of Educational Research (EJER)* 34 (2009).

[4] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*. 13–22.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *JAIR* 16 (2002), 321–357.

[6] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *SIGKDD 2016*. 785–794.

[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 *SIAM 2019*. SIAM, 235–243.

[8] Lu Cheng, Kai Shu, Siqi Wu, Yasin N Silva, Deborah L Hall, and Huan Liu. 2020. Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model. In *CIKM 2020*.

[9] Lu Cheng, Yasin Silva, Deborah Hall, and Huan Liu. 2020. Session-based Cyberbullying Detection: Problems and Challenges. *IEEE Internet Computing, Special Issue on Cyber-Social Health: Promoting Good and Countering Harm on Social Media* (2020).

[10] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *EMNLP 2014*. 1724–1734.

[11] Lucie Corcoran, Irene Connolly, and Mona O'Moore. 2012. Cyberbullying in Irish schools: An investigation of personality and self-concept. *The Irish Journal of Psychology* 33, 4 (2012), 153–165.

[12] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

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

[14] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *TIIS* 2, 3 (2012), 1–30.

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

[16] Julian J Dooley, Jacek Pyżalski, and Donna Cross. 2009. Cyberbullying versus face-to-face bullying: A theoretical and conceptual review. *Zeitschrift für Psychologie/Journal of Psychology* 217, 4 (2009), 182–188.

[17] Centers for Disease Control and Prevention. 2019. Youth Risk Behavior Survey Data. https://www.cdc.gov/healthyyouth/data/yrbs/

[18] Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia Lab @ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations. In *Proceedings of the Workshop on Noisy User-generated Text*. 146–153. https://doi.org/10.18653/v1/W15-4322

[19] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 149–156.

[20] Nicole Gruber and Alfred Jockisch. 2020. Are GRU Cells More Specific and LSTM Cells More Sensitive in Motive Classification of Text. *J. Front. Artif. Intell* 3 (2020), 1–6.

[21] 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 *The WebConf 2020 Companion*.

[22] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NIPS 2017*. 1024–1034.

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

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

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

[26] Linmei Hu, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In *EMNLP-IJCNLP 2019*. 4823–4832.

[27] Binxuan Huang and Kathleen M Carley. 2019. Syntax-Aware Aspect Level Sentiment Classification with Graph Attention Networks. In *EMNLP-IJCNLP 2019*. 5472–5480.

[28] 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.

[29] Bernardo Huberman, Daniel M Romero, and Fang Wu. 2009. Social networks that matter: Twitter under the microscope. *First Monday* (2009).

[30] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP 2014*. 1746–1751.

[31] Robin M Kowalski, Gary W Giumetti, Amber N Schroeder, and Micah R Lattanner. 2014. Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological bulletin* 140, 4 (2014), 1073.

[32] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. In *ICLR 2016*.

[33] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2016. Semi-supervised classification with graph convolutional networks. In *ICLR 2017*.

[34] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *ACL 2020*. 7342–7351.

[35] Parma Nand, Rivindu Perera, and Abhijeet Kasture. 2016. "How Bullying is this Message?": A Psychometric Thermometer for Bullying. In *COLING 2016*. 695–706.

[36] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in Vine. In *ASONAM 2015*. IEEE, 617–622.

[37] Chiara Ravazzi, Roberto Tempo, and Fabrizio Dabbene. 2017. Learning influence structure in sparse social networks. *IEEE Transactions on Control of Network Systems* 5, 4 (2017), 1976–1986.

[38] Walisa Romsaiyud, Kodchakorn na Nakornphanom, Pimpaka Prasertsilp, Piyaporn Nurarak, and Pirom Konglerd. 2017. Automated cyberbullying detection using clustering appearance patterns. In *2017 9th International Conference on Knowledge and Smart Technology*. IEEE, 242–247.

[39] Christina Salmivalli. 1999. Participant role approach to school bullying: Implications for interventions. *Journal of adolescence* 22, 4 (1999), 453–459.

[40] Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry* 49, 4 (2008), 376–385.

[41] Devin Soni and Vivek Singh. 2018. Time reveals all wounds: Modeling temporal dynamics of cyberbullying sessions. In *ICWSM 2018*. AAAI press, 684–687.

[42] Ditch the Label Anti Bullying Charity. 2013. Ditch the Label Anti Bullying Charity: The annual cyberbullying survey 2013. https://www.ditchthelabel.org/wp-content/uploads/2016/07/cyberbullying2013.pdf. Accessed: 2020-09-18.

[43] Inc. US Legal. 2019. Cyber Bullying Law and Legal Definition. https://definitions.uslegal.com/c/cyber-bullying/

[44] 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 *NIPS 2017*. 5998–6008.

[45] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. In *NIPS 2018*.

[46] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR 2018*.

[47] Hao Wu, Jiajun Bu, Chun Chen, Can Wang, Guang Qiu, Lijun Zhang, and Jianfeng Shen. 2010. Modeling dynamic multi-topic discussions in online forums. In *AAAI 2010*. Citeseer.

[48] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from Bullying Traces in Social Media. In *NAACL 2012*. Association for Computational Linguistics, 656–666.

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

[50] Rui Zhao and Kezhi Mao. 2016. Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing* 8, 3 (2016), 328–339.

[51] Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *ACL 2020*. 6170–6180.

[52] Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification. In *ACL 2019*. 892–901.

[53] Caleb Ziems, Ymir Vigfusson, and Fred Morstatter. 2020. Aggressive, Repetitive, Intentional, Visible, and Imbalanced: Refining Representations for Cyberbullying Classification. In *ICWSM 20*, Vol. 14. 808–819.