XBully: Cyberbullying Detection within a Multi-Modal Context

Lu Cheng[†], Jundong Li[†], Yasin N. Silva[‡], Deborah L. Hall[⋆], Huan Liu[†]
[†]Computer Science and Engineering, Arizona State University
[‡]Mathematical and Natural Sciences, Arizona State University
[⋆]Social and Behavioral Sciences, Arizona State University
{lcheng35,jundongl,ysilva,d.hall,huanliu}@asu.edu

ABSTRACT

Over the last decade, research has revealed the high prevalence of cyberbullying among youth and raised serious concerns in society. Information on the social media platforms where cyberbullying is most prevalent (e.g., Instagram, Facebook, Twitter) is inherently multi-modal, yet most existing work on cyberbullying identification has focused solely on building generic classification models that rely exclusively on text analysis of online social media sessions (e.g., posts). Despite their empirical success, these efforts ignore the multi-modal information manifested in social media data (e.g., image, video, user profile, time, and location), and thus fail to offer a comprehensive understanding of cyberbullying. Conventionally, when information from different modalities is presented together, it often reveals complementary insights about the application domain and facilitates better learning performance. In this paper, we study the novel problem of cyberbullying detection within a multi-modal context by exploiting social media data in a collaborative way. This task, however, is challenging due to the complex combination of both cross-modal correlations among various modalities and structural dependencies between different social media sessions, and the diverse attribute information of different modalities. To address these challenges, we propose XBully, a novel cyberbullying detection framework, that first reformulates multi-modal social media data as a heterogeneous network and then aims to learn node embedding representations upon it. Extensive experimental evaluations on real-world multi-modal social media datasets show that the XBully framework is superior to the state-of-the-art cyberbullying detection models.

KEYWORDS

Cyberbullying Detection, Multi-Modality, Social Media, Network Embedding, Collaborative Learning

ACM Reference Format:
Lu Cheng, Jundong Li, Yasin N. Silva, Deborah L. Hall, Huan Liu. 2019. XBully: Cyberbullying Detection within a Multi-Modal Context. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3289600.3291037

WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5940-5/19/02…$15.00
https://doi.org/10.1145/3289600.3291037

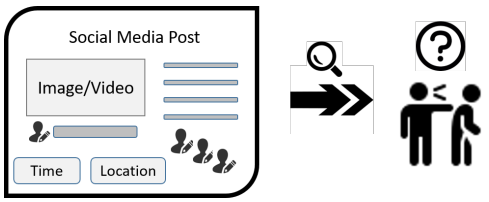


Figure 1: Illustration of cyberbullying detection within a multi-modal context: the left side figure represents a social media session (e.g., post) with rich user-generated information such as an image, video, user profile, time, location and comments. In addition, different sessions are inherently connected with each other through user-user social relations. The goal is to predict if a particular session is bullying or not by leveraging its multi-modal context information.

1 INTRODUCTION

Cyberbullying, commonly defined as the electronic transmission of insulting or embarrassing comments, photos, or videos, has become increasingly prevalent on social networks. Reports from the American Psychological Association and the White House, for example, reveal that more than 40% of teenagers in the US indicate that they have been bullied on social media platforms [7]. The growing prevalence and severity of cyberbullying on social media and the link between cyberbullying and such negative outcomes as depression, low self-esteem, and suicidal thoughts and behaviors have led to the identification of cyberbullying as a serious national health concern. It has also motivated a surge in research in psychology and computer science aimed at better understanding the nature and key characteristics of cyberbullying in social networks.

Within the computer science literature, existing efforts toward detecting cyberbullying have primarily focused on text analysis. These works attempt to build a generic binary classifier by taking high-dimensional text features as the input and make predictions accordingly. Despite their satisfactory detection performance in practice, these models inevitably ignore critical information included in the various social media modalities such as image, video, user profile, time, and location. For example, Instagram[1] allows users to post and comment on any public image to express their opinions and preferences. In light of this, bullies can post humiliating images or insulting comments, captions, or hashtags, edit and then re-post someone else's images, and even create fake profiles pretending to other individuals altogether [13]. Therefore, it is critical to exploit the rich user-generated content within a multi-modal

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

context to gain greater insight into cyberbullying behaviors and generate more accurate predictions. Fig. 1 illustrates the cyberbullying detection problem within a multi-modal context.

Despite the potential benefit, performing cyberbullying detection within a multi-modal context presents multiple challenges. First, information from different modalities might be complementary, thereby facilitating better learning performance, especially when the data is sparse. However, heterogeneous information from different modalities might not be compatible and, in the worst case, some modalities may be entirely independent. Thus, a key problem that has not been sufficiently addressed in cyberbullying detection is how to effectively encode the cross-modal correlation among different types of modalities. Second, social media data are often not independent and identically distributed (i.i.d.) but rather intrinsically correlated, either directly or indirectly, limiting the applicability of conventional text analysis approaches. For example, if two social media sessions (e.g., posts) are from the same user or are posted by a pair of friends, their content similarity is expected to be high based on the homophily principle [20]. Considering this, it is important to model structural dependencies among different social media sessions when performing cyberbullying detection. Third, although multi-modal social media data can be useful in understanding human behavior, it is difficult to directly make use of it because different modalities are frequently associated with rather diverse feature types (e.g., nominal, ordinal, interval, ratio, etc.), and in some cases, some modalities that identify particular entities (e.g., users) cannot be simply represented as feature vectors[2]. Therefore, it is crucial that the solution framework uses an expressive way to represent modalities with diverse feature types.

To address the above challenges, we propose a novel cyberbullying detection framework, XBully, that models multi-modal social media data in a collaborative way. Specifically, to capture cross-modal correlation among modalities as well as the structural dependencies among different social media sessions, we model the multi-modal social media data as a heterogeneous network by exploiting *co-existence* and *neighborhood* relations (explained later) and aim at learning the embedding representations for nodes in the network. Due to data sparsity, we identify a number of *hotspots* for each mode, which provide a succinct high-level summarization of similar modality attribute values. For nominals (modalities without attributes), we form nodes in the constructed heterogeneous network using their meta information, e.g., user IDs. After learning the embedding representation for nodes in the heterogeneous network, each social media session can be represented as a numerical vector by concatenating the node embeddings in that session. Using these vectors, various off-the-shelf machine learning models can be directly applied to provide accurate cyberbullying detection and a deeper understanding of cyberbullying behaviors. The main contributions of this work are:

- **Problem Formulation:** We formally define the problem of cyberbullying detection within a multi-modal context. The definition is a result of multiple modalities exploited in a collaborative fashion.
- **Algorithms:** We propose a novel cyberbullying detection framework (XBully) with three core components: (1) a hotspot detector that identifies centroids for each mode; (2) a module that constructs a heterogeneous network by leveraging *co-existence* and *neighborhood* relations of the detected hotspots and instances of nominals; and (3) a principled joint embedding module that effectively encodes cross-modal correlation and structural dependencies among different social media sessions to learn noise-resilient embedding representations. The resulting embeddings enable better detection and understanding of cyberbullying behaviors.
- **Evaluations:** We perform experiments on two real-world social media datasets to corroborate the efficacy of the proposed framework.

[2]We refer to modalities with attributes and without attributes as modes and nominals, respectively.

2 PROBLEM DEFINITION

In this section, we first introduce the problem of cyberbullying detection within a multi-modal context, briefly describe our approach to solve it via network representation learning, and highlight key challenges.

2.1 Multi-Modal Cyberbullying Detection

Definition 1. Cyberbullying Detection within a Multi-Modal Context *Given a corpus of social media sessions C (e.g., posts) with M modalities, cyberbullying detection within a multi-modal context aims at identifying instances of cyberbullying by leveraging multiple modalities such as textual features, spatial locations, and visual cues, as well as the relations among sessions.*

The definition of multi-modal cyberbullying detection builds on the concept of multi-modality learning in machine learning [2]. Here, we emphasize the multi-modal context of social media sessions. In our experiments, we use the following modalities extracted from an Instagram session:

- User - It is a typical type of nominals and we use the relations among users to decode the dependencies between social media sessions.
- Image - The associated meta-information of an image forms a tuple composed of the number of shares, the number of likes, and the labels describing the category of this image.
- Profile - The meta-information of a user forms a tuple with the number of followers, the number of follows, the total number of comments, and the total number of likes received.
- Time - The timestamp of posting an image. We consider the time of the day (24h range) and convert the raw time to the range of [0, 86400] by calculating its offset (in seconds) w.r.t. 12:00 am.
- Text - We perform psychometric analysis on the textual information of the session, i.e., description of the image and comments, and obtain the psychological features through LIWC [24]. We base this on previous research indicating that such psychometric analysis can provide insights about cyberbullying behaviors [22].

2.2 Cyberbullying Detection via Multi-Modal Network Representation Learning

Let C be a corpus of social media sessions. We define each session $s \in C$ as a tuple $< \mathbf{x}_{s1}, \mathbf{x}_{s2}, ..., \mathbf{x}_{sM}, y_{s1}, ..., y_{sN} >$ where M and N denote the number of modes and nominals, respectively. $\mathbf{x}_{sm}$ is the feature vector of s in mode $m \in (1, 2, 3...M)$. For example, a geo-tagged social media session may have location information $\mathbf{x}_{sm} = [34.0489, 111.0937]$. In addition, different sessions are inherently connected with each other via social relations among users. Our goal is to represent the original corpus C as a heterogeneous network G by capturing its multi-modal nature and learn high-quality embeddings for each node in this network.

In contrast to simply concatenating the raw multi-modal feature vectors of each modality, the learned node embeddings in the resultant heterogeneous network capture both *structural dependencies* among different social media sessions and *cross-modal correlations* among different modalities in a joint framework.

2.3 Challenges

- **Number of Distinct Feature Values.** Social media data usually comes in complex forms and exhibits considerable variations due to its multi-modal nature. We are often confronted with diverse feature types and the number of unique feature values each mode can take is often exceedingly large, which can cause the problem of data sparsity. Furthermore, the available training data for each node in the network is often limited, further complicating the training process.
- **Cross-Modal Correlation and Structural Dependencies.** An effective network embedding model for multi-modal social media data should preserve the node proximities in terms of both cross-modal correlation and structural dependencies among different sessions. Conventional network embedding models such as Deepwalk [25], LINE [32], and node2vec [10], which primarily focus on encoding structural information of homogeneous networks, cannot be effectively applied in our problem. Metapath2vec [9] is a recently proposed heterogeneous network embedding model that relies on a set of predefined meta-paths to find the neighborhoods around nodes. However, in our problem, the number of meta-paths is often very large, making metapath2vec inapplicable.
- **Information Noise.** While the rich multi-modal data can provide valuable and complementary insights for identifying cyberbullying behaviors, such data can also be cluttered and noisy, thus complicating the process of gaining actionable knowledge from it.

To address these challenges, we propose mode hotspots detection using kernel density estimation (KDE) [23] to reduce the number of unique feature values. Motivated by [39], after the hotspots are identified, we then construct a heterogeneous network by leveraging *co-existence* relations, to exploit the connections among different modalities in the same session, and *neighborhood* relations, to connect nodes of the same modality in different sessions. We develop a graph-based joint embedding module to capture the cross-modal correlation and structural dependencies in a joint framework. This embedding module maps all the mode hotspots and nominal nodes in a heterogeneous network into a common latent space. To alleviate the negative impacts of noise, we also identify the most informative neighbors for each node to refine the learned embeddings. The overall framework is further explained in Fig. 2.

3 THE XBULLY FRAMEWORK

This section presents the proposed XBully model in detail. Specifically, we first show how to identify succinct yet accurate summarizations of groups of similar feature values (mode hotspots). Then, we present a principled way to capture both cross-modal correlation and structural dependencies in a joint framework for embedding representation learning. We also discuss how to alleviate the negative impact of noise during the embedding training phase by allowing nodes to borrow strength from each other in a collaborative way.

3.1 Mode Hotspot Detection

Previous work has shown that high-dimensional feature representation not only suffers from data sparsity but also poses great challenges to downstream learning tasks due to the curse of dimensionality [18]. To address this issue, we propose the concept of mode hotspots based on KDE [23], which is a non-parametric method to estimate the density function from a collection of data samples. With KDE, we do not need to establish any prior knowledge about the data distribution, as it provides automatic discovery of arbitrary modes from complex data spaces [39].

Definition 2. Mode Hotspots *Given a corpus of social media sessions C, the mode hotspots for mode m ($m = 1, 2..., M$) are the set of local maximums of the kernel density function estimated from m.*

Then, given n sessions containing mode m in a d-dimensional feature space $X_m = (\mathbf{x}_{1m}, \mathbf{x}_{2m}, ..., \mathbf{x}_{nm})$, the kernel density at any point $\mathbf{x}$ with mode m is given by:

$$f(\mathbf{x}) = \frac{1}{n\delta_m^d} \sum_{i=1}^{n} K(\frac{\mathbf{x} - \mathbf{x}_{im}}{\delta_m}), \tag{1}$$

where $K(\cdot)$ is a predefined kernel function, and δ_m is the kernel bandwidth for mode m. We further leverage the advanced mean-shift algorithm used in [39] to identify the mode hopspots.

3.2 Network Representation Learning

Following [39], now we investigate how to build a heterogeneous network by exploiting the *co-existence* and *neighborhood* relations, such that both the cross-modal correlations and structural dependencies are properly captured. The *co-existence* relation is established between two nodes when they co-exist in the same social media session and the *neighborhood* relations among mode hotspots are built upon the idea of *modality continuity* [33] which relies on the idea that nearby objects are more closely related than distant objects. We first define the node kernel, based on which the *neighborhood* relations are formed:

Definition 3. Node Kernel *For two mode hotspots u_i and u_j in mode m with feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, the kernel strength between them is :*

$$w(u_i, u_j) = \begin{cases} \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\delta_m^2)}{2\pi\delta_m^2}, & if \|\mathbf{x}_i - \mathbf{x}_j\| \le \delta_m \\ 0, & \text{otherwise.} \end{cases}$$

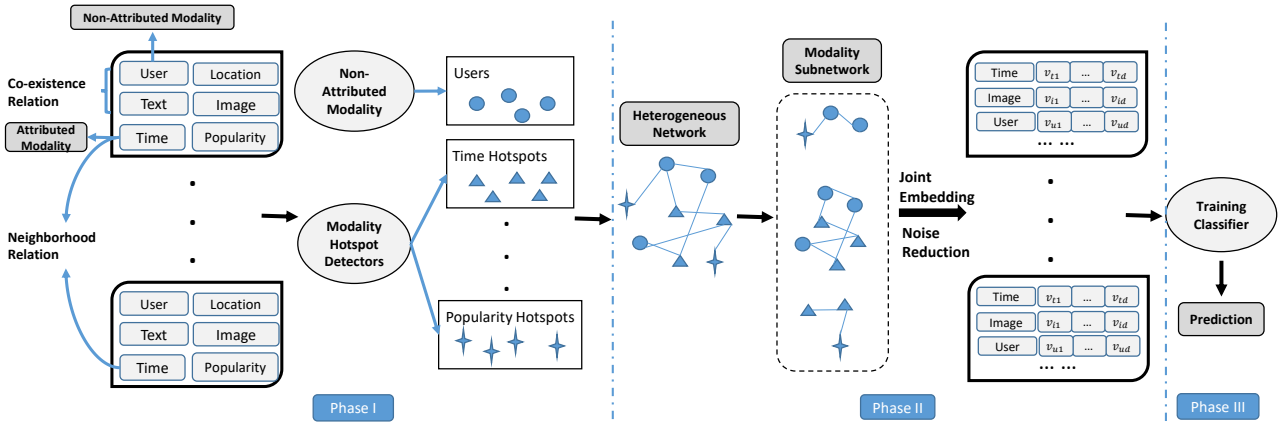


Figure 2: The proposed XBully framework. Given a corpus of social media sessions, we first attempt to discover hotspots for each mode (Phase I). Then, based on the detected hotspots and instances of nominals, we leverage the *co-existence* and *neighborhood* relations to construct a heterogeneous network that is later divided into several modality subnetworks (Phase II). Each subnetwork consists of two modalities. Nodes in these subnetworks are then mapped into the same latent space through network representation learning. Finally, we concatenate embeddings of nodes in each session and apply off-the-shelf machine learning models for cyberbullying detection (Phase III).

Therefore, the neighbors of a mode hotspot v in the heterogeneous network are the set of mode hotspots that produce a *non-zero* kernel strength value with hotspot v. In addition to that, for nominal nodes, we define the *neighborhood* relations based on their structural information by making use of the dependencies (e.g., social relations) between different sessions. For example, an Instagram session could have five different modalities - user, image, profile, time, and comments (text). From the definition of *co-existence* relations, we construct the following 10 types of edges in the heterogeneous network: *user-image*, *user-profile*, *user-time*, *user-text*, *image-profile*, *image-text*, *image-text*, *profile-time*, *profile-text* and *time-text*. Moreover, the *neighborhood* relations also generate the following 4 edge types: *image-image*, *profile-profile*, *time-time* and *text-text* using Definition 3 and *user-user* edges with nominal nodes are built using the social relations among users. With the above defined edge types, we define the weight of an edge considering the following three scenarios: (a) the normalized co-existence count (between 0 and 1); (b) kernel strength (between 0 and 1); and (c) the dependencies between nominal nodes (0 or 1). Since the resultant network has different types of nodes and edges, it would be inappropriate to directly apply a conventional network embedding algorithm such as Deepwalk [25] or node2vec [10] to learn the embeddings for each node. Instead, we build on [31] to decompose the heterogeneous networks into multiple modality subnetworks (using two modalities) and learn embeddings within each subnetwork. This approach, the learned embeddings can capture the node proximity across different types of edges. In what follows, we provide the details of the joint embedding model.

First, let us denote G_S as the set of all modality subnetworks. For any two different modalities $A, B \in (1, 2, ..., M + N)$, we can construct a modality subnetwork $G_{AB} \in G_S$. Then, the probability of node j with modality B generated from node i with modality A is defined by the following conditional probability:

$$p(j|i) = \frac{exp(v_j^T \cdot v_i)}{\sum_{k \in B} exp(v_k^T \cdot v_i)}, \tag{2}$$

where v_j denotes the embedding representation of node j with modality B and v_i is the embedding vector of node i with modality A. Next, we learn embeddings by minimizing the distance between the conditional distribution of the context nodes given the center node and the empirical distribution. The empirical distribution of node i is defined as $p'(j|i) = \frac{w_{ij}}{d_i}$, where w_{ij} is the weight of the edge $i - j$ and d_i is the out-degree of node i, i.e., $d_i = \sum_{j \in B} w_{ij}$. Therefore, we define the loss function as follows:

$$O_{AB} = \sum_{i \in A} d_i KL\big(p'(\cdot|i)||p(\cdot|i)\big), \tag{3}$$

where the $KL(\cdot)$ denotes the KL-divergence between two probability distributions. By omitting the constants, the above loss function can be reformulated as follows:

$$O_{AB} = -\sum_{i \in A, j \in B} w_{ij} \log p(v_j|v_i). \tag{4}$$

As each modality subnetwork is composed of four different types of edges, $A-A$, $A-B$, $B-A$, and $B-B$, the overall loss function of a modality subnetwork G_{AB} is computed as follows:

$$Z_{AB} = O_{AA} + O_{AB} + O_{BB} + O_{BA}. \tag{5}$$

3.3 Embedding Refinement

While multi-modal information is useful in improving the embedding quality, the above model can be problematic when the network is very sparse. In addition, when the discovered mode hotspots are very noisy, the embedding representation learning phase may be adversely affected. To address these problems, we propose a noise-resilient embedding refinement approach to *adaptively* choose the

most informative neighbors for each node. The core idea of the refinement method is to find the best locally weighted context vectors (predictors) to reconstruct the embedding of the center node. Specifically, given n embedding vectors $v_1, ..., v_j, ..., v_n \in \mathbb{R}^d$, the problem is to estimate $\hat{v}_i$ using an estimator of the form $\hat{v}_i = \sum_{j=1, i\neq j}^{n} \alpha_{ji} v_j$, s.t. $\sum_j \alpha_{ji} = 1$, where α_{ji} denotes the extent to which the embedding v_i is influenced by v_j. Our solution builds on the algorithm presented in [1] to adaptively learn the optimal neighborhood structure for each center node and automatically quantify the influence from other nodes. This can be formulated as follows:

$$\begin{aligned} R_i = |v_i - \sum_{j=1, i\neq j}^{|V_{AB}|} \alpha_{ji} v_j|, \quad i \in V_{AB}, \\ \text{s.t.} \sum_{j=1, j\neq i}^{|V_{AB}|} \alpha_{ji} = 1, \end{aligned} \tag{6}$$

where V_{AB} represents the node set in modality subnetwork G_{AB}. By integrating the above embedding refinement component, the new objective function for embedding representation learning is:

$$Z_{AB} = O_{AA} + O_{AB} + O_{BA} + O_{BB} + \lambda \sum_{i=1}^{|V_{AB}|} R_i, \tag{7}$$

where λ is a parameter that balances the contribution of the refinement component. As a result, the overall objective function of our multi-modal network embedding is:

$$O = \sum_{G_S} Z_{AB}, \qquad G_{AB} \in G_S. \tag{8}$$

To solve the final objective function, we alternate between the updates of embedding variables and the influential matrix with entry α_{ij}. To update the embedding vectors, we use the stochastic gradient descent (SGD) to optimize different modality subnetworks with negative sampling [21]. Specifically, for an edge e_{ij}, we randomly select K nodes that are not connected with node i as negative samples. The influential matrix is then updated via the algorithm proposed in [1] with the input of updated embedding vectors.

4 EXPERIMENTS

In this section, we aim to answer the following research questions: (1) Is the proposed XBully framework superior to existing models that solely rely on text information for cyberbullying detection? (2) How effective is the noise-resilient embedding refinement component for embedding representation learning? (3) Does the proposed multi-modal network embedding method help achieve better detection performance than those of conventional network embedding methods? (4) What kind of insights can XBully provide for psychology and social scientists? (5) How robust is the proposed model w.r.t the different model parameters? Each experiment was run 10 times. This section reports the averaged experimental results.

4.1 Experimental Setup

Datasets. Our experiments are performed using two real-world social media datasets[3]. Each social media session in the Instagram dataset [13] includes image descriptions, user comments, and the creation time of the session. The dataset also has user profile information and social relations among users. The second dataset [26] was collected from Vine, a mobile application website that allows users to record and edit six-second looping videos. Each Vine session is associated with video descriptions, user comments, and the creation time of the session. Basic statistics of these datasets are shown in Table 1. Please refer to [14, 26] for additional details.

Table 1: Dataset statistics.

Datasets	# Sessions	# Bullying	# Normal	# Comments
Instagram	2,218	678	1,540	155,260
Vine	970	304	666	78,250

Baseline Methods. To answer the first two research questions, we compare XBully with the commonly used feature engineering approach, two recently proposed cyberbullying models, and a variant of XBully without the noise-resilient embedding refinement.

- Raw Features (Raw): This is a concatenation of all the multi-modal features such as network feature and text feature.
- *Bully* [37]: A pretrained classifier[4] based on textual analysis.
- *SICD* [6]: The state-of-the-art cyberbullying detection model which uses sentiment information embedded in the user-generated content to guide the prediction.
- XBully variant (Variant): A variant of XBully without the noise-resilient embedding refinement.

We also compare the proposed model with three widely used network embedding models - DeepWalk [25], Node2vec [10] and GraRep [3]. To reduce the effect of model variances on performance evaluation, we tested these methods on three classification models, including *Random Forest*, *Linear SVM* and *Logistic Regression*. Multiple training datasets are generated by extracting increasing fractions (10% to 90%) of the entire datasets and the remaining parts are used as the test datasets.

Parameter Settings. The XBully framework has the following parameters: (1) the hotspot detection bandwidth h_m for each mode $m \in M$; (2) the weight of the noise-resilient component λ; and (3) the embedding dimension d. By default, we set parameters for the Instagram dataset as $h_t = 150$ (time), $h_i = 100$ (image), $h_l = 100$ (LIWC), $h_u = 500$ (user profile), $\lambda = 0.01$ and $d = 500$. For the Vine dataset, we set $h_t = 50$ (time), $h_l = 5$ (LIWC), $\lambda = 0.01$ and $d = 500$. A detailed parameter sensitivity analysis is presented later in this section.

4.2 Performance Evaluation

In this subsection we aim to answer the first three research questions. To this end, two common evaluation metrics are calculated - Macro F1 (Mac F1) and Micro F1 (Mic F1). A macro-average computes the metric independently for each class and then takes the average as the output, whereas a micro-average will aggregate the contributions of all classes to compute the average metric. In binary classification, Micro F1 is equal to Accuracy. Table 2 and Table 3 report the cyberbullying detection performance of various methods

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

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

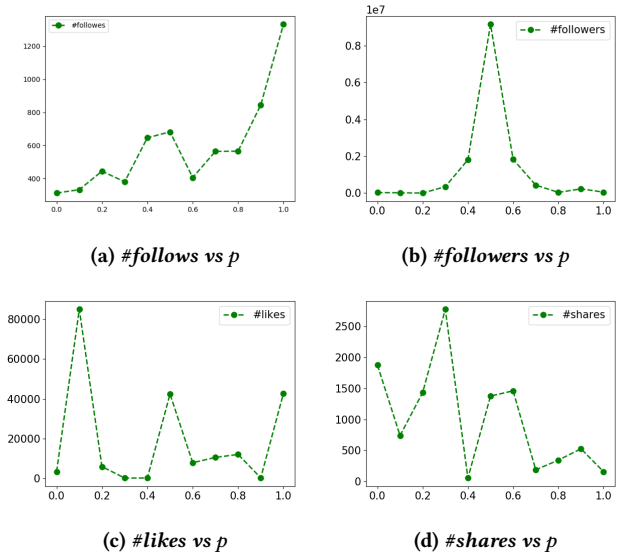
on the two datasets. We make the following observations from these results:

- In most cases and with both datasets, XBully significantly outperforms the method that works with concatenated multi-modal features (Raw) and the methods with homogeneous network embedding (Deepwalk, Node2vec and GraRep). The better performance of XBully against the other methods shows the effectiveness of performing cyberbullying detection with multi-modal network representation learning as it captures both the cross-modal correlations and structural dependencies.
- XBully is also superior to the two cyberbullying detection methods (*Bully* and *SICD*) under Macro F1 in both datasets and under Micro F1 in the Instagram dataset. Observe that, in the few experiments where *SICD* outperforms XBully under Micro F1 in the Vine dataset, XBully's Macro F1 scores are significantly higher (more than 50% higher). These results show that multi-modal information can indeed provide complementary insights to achieve better learning performance.
- The improvements of XBully over baseline methods are consistent across different classifiers. This indicates that the learned embedding representations are effective and can be easily generalized to various off-the-shelf machine learning models.
- XBully achieves better detection performance than the variant without the embedding refinement component. This result highlights the benefit of collaboratively refining the embeddings by integrating information from similar nodes during the learning process, which in turn makes the learned embedding representation more robust to noise.

4.3 Qualitative Analysis

We are also interested in findings that offer insight into common social behaviors of users who have experienced cyberbullying versus users who have not. To this end, we interpret the confidence level of each label in the datasets as an indicator of the probability that the session is cyberbullying (p).[5] To understand how XBully can provide insights for social scientists and psychologists, we incorporated the new type of node into the previous model by treating it as another modality and retrained all of the embeddings. Afterwards, we made queries w.r.t. p in the range of $(0.1, 0.2, ..., 1.0)$ and XBully returned the most similar mode hotspots based on the *cosine* similarity. The question we aim to answer is how user behavior in social media platforms varies in relation to the probability of cyberbullying. The following analysis is based on the experiments on the Instagram dataset.

As shown in Fig. 3(a), *#follows* gets larger as p increases, i.e., the number of users a session owner follows increases when the probability of cyberbullying increases. This may indicate that users who are more active in following others on social media have a higher probability of experiencing cyberbullying. In Fig. 3(b), the shape of the distribution of *#followers* at different probability levels of cyberbullying approximately follows a normal distribution. This pattern suggests that the difference between *#followers* of users who experienced cyberbullying and users who did not is small. Fig. 3(c)-(d) shed light on the relationships between the popularity of a social media session and p, as indicated by the number of likes a post receives in Fig. 3(c) and the number of times a post is shared in Fig. 3(d). In Fig. 3(c), three peaks can be seen, at $p = 0.1, 0.5$, and 1.0. In this figure, *#likes* at $p = 0.1$ is twice as large as those at $p = 0.5$ and 1.0. In Fig. 3(d), *#shares* decreases as p gets larger. One possible explanation is that most social media users are normal users who are not specifically interested in cyberbullying-related content. Although tentative, the trends in Fig. 3 elucidate a potentially novel way for interdisciplinary researchers to measure social influence within the context of social media interactions, particularly as they relate to cyberbullying risk.



(a) *#follows vs p* (b) *#followers vs p*

(c) *#likes vs p* (d) *#shares vs p*

Figure 3: Qualitative analysis

4.4 Parameter Study

The XBully framework has four parameters (h_t, h_i, h_l, h_u) for mode hotspot detection, and two parameters (d and λ) for the joint embedding module. To investigate the effect of these parameters, we run experiments on the Instagram dataset to vary one parameter at a time and evaluate how it affects the classification performance. The results presented in Fig. 4 show that XBully is not very sensitive to kernel bandwidth parameters except for h_l. The performance of XBully increases moderately when h_l becomes larger, i.e., the number of detected text hotspots is relatively small. In Fig. 4(e), we can see that when embedding dimensionality increases, the performance of XBully first improves and then remains stable. Fig. 4(f) shows that the best performance is achieved when λ is around 0.01. In general, XBully is not sensitive to most of the model parameters, and consequently can be tuned for various application purposes.

5 RELATED WORK

Cyberbullying Detection Prior work on cyberbullying detection has relied primarily on text analysis to identify cyberbullying cases

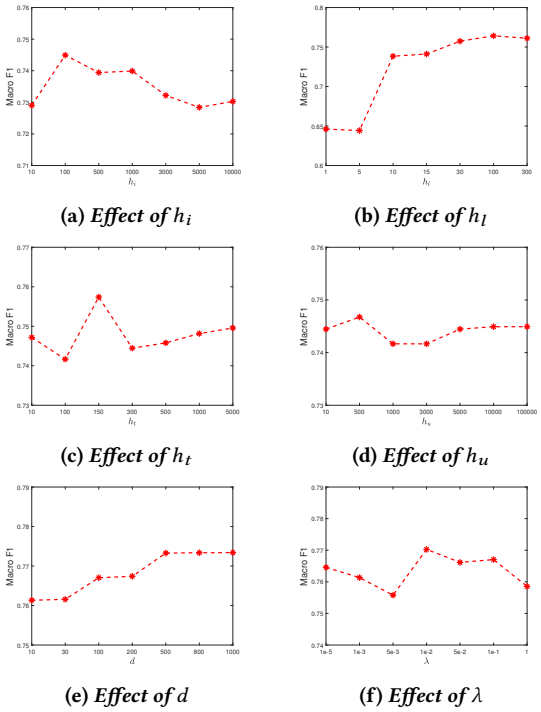
[5]For details of the confidence level, please refer to [14]

Table 2: Performance comparison of various methods on the Instagram dataset.

Percentages		10%		30%		50%		70%		90%	
Metrics		Mac F1	Mic F1	Mac F1	Mic F1	Mac F1	Mic F1	Mac F1	Mic F1	Mac F1	Mic F1
Random Forest	Raw	0.528	0.838	0.573	0.835	0.517	0.830	0.532	0.827	0.543	0.860
	DeepWalk	0.461	0.668	0.445	0.680	0.450	0.678	0.470	0.679	0.432	0.716
	Node2vec	0.519	0.714	0.550	0.712	0.584	0.717	0.562	0.716	0.599	0.770
	GraRep	0.459	0.671	0.456	0.680	0.464	0.680	0.460	0.671	0.455	0.707
	Variant	0.551	0.844	0.680	0.874	0.778	0.905	0.854	0.926	0.932	0.959
	XBully	**0.566**	**0.853**	**0.702**	**0.887**	**0.814**	**0.920**	**0.865**	**0.937**	**0.963**	**0.982**
Linear SVM	Raw	0.459	0.559	0.459	0.564	0.515	0.692	0.540	0.793	0.582	0.847
	DeepWalk	0.523	0.598	0.518	0.581	0.522	0.591	0.508	0.593	0.540	0.635
	Node2vec	**0.586**	0.663	0.577	0.635	0.612	0.665	0.582	0.643	0.622	0.680
	GraRep	0.513	0.585	0.534	0.603	0.515	0.621	0.505	0.626	0.568	0.712
	Variant	0.568	0.812	0.659	0.828	0.747	0.863	0.796	0.890	0.782	0.914
	XBully	0.570	**0.819**	**0.668**	**0.840**	**0.781**	**0.886**	**0.821**	**0.904**	**0.837**	**0.928**
Logistic Regression	Raw	0.459	0.828	0.460	0.830	0.465	0.819	0.451	0.82	0.461	0.856
	DeepWalk	0.512	0.634	0.523	0.620	0.508	0.618	0.491	0.602	0.514	0.644
	Node2vec	**0.581**	0.681	**0.584**	0.661	**0.602**	0.675	0.572	0.656	0.610	0.707
	GraRep	0.506	0.623	0.538	0.648	0.499	0.638	0.495	0.646	0.494	0.698
	Variant	0.495	0.837	0.522	0.832	0.536	0.835	0.543	0.826	0.615	0.874
	XBully	0.497	**0.841**	0.528	**0.836**	0.593	**0.849**	**0.599**	**0.848**	**0.621**	**0.878**
Cyberbully models	*Bully*	0.274	0.331	0.271	0.325	0.267	0.318	0.277	0.334	0.278	0.335
	SICD	0.447	0.646	0.443	0.604	0.383	0.537	0.438	0.512	0.358	0.559

Table 3: Performance comparison of various methods on the Vine dataset.

Percentages		10%		30%		50%		70%		90%	
Metrics		Mac F1	Mic F1	Mac F1	Mic F1	Mac F1	Mic F1	Mac F1	Mic F1	Mac F1	Mic F1
Random Forest	Raw	0.651	0.716	0.663	0.729	0.641	0.709	0.663	0.725	0.749	0.784
	DeepWalk	0.575	0.695	0.635	0.738	0.677	0.763	0.683	0.759	0.638	0.691
	Node2vec	0.576	0.704	0.626	0.733	0.655	0.746	0.679	0.753	0.638	0.691
	GraRep	0.589	0.703	0.633	0.723	0.671	0.751	0.694	0.763	0.650	0.691
	Variant	0.659	0.738	0.676	0.733	0.682	0.738	0.709	0.766	0.705	0.753
	XBully	**0.661**	**0.740**	**0.678**	**0.758**	**0.711**	**0.779**	**0.717**	**0.777**	**0.757**	**0.784**
Linear SVM	Raw	0.409	0.683	0.432	0.439	0.575	0.701	0.578	0.588	0.547	0.557
	DeepWalk	0.568	0.646	0.582	0.661	0.597	0.643	0.571	0.639	0.528	0.557
	Node2vec	0.569	0.659	0.592	0.658	0.579	0.627	0.599	0.649	0.620	0.649
	GraRep	0.590	0.664	0.610	0.669	0.634	0.689	0.644	0.715	0.629	0.649
	Variant	0.636	0.715	0.622	0.689	0.650	0.711	0.678	0.732	0.671	0.722
	XBully	**0.657**	**0.717**	**0.641**	**0.704**	**0.651**	**0.733**	**0.678**	**0.742**	**0.700**	**0.753**
Logistic Regression	Raw	0.648	0.732	0.683	0.748	0.684	0.755	0.676	0.746	0.705	0.753
	DeepWalk	0.554	0.668	0.581	0.691	0.631	0.705	0.598	0.680	0.589	0.639
	Node2vec	0.535	0.672	0.603	0.705	0.641	0.715	0.629	0.708	0.612	0.660
	GraRep	0.578	0.672	0.626	0.717	0.650	0.732	0.644	0.725	0.633	0.670
	Variant	0.618	0.724	0.663	0.741	0.696	0.757	0.705	0.770	0.658	0.732
	XBully	**0.670**	**0.737**	**0.700**	**0.756**	**0.700**	**0.759**	**0.706**	**0.790**	**0.769**	**0.804**
Cyberbully models	*Bully*	0.283	0.354	0.314	0.417	0.350	0.480	0.389	0.533	0.481	0.619
	SICD	0.417	0.715	0.436	**0.773**	0.506	0.775	0.474	**0.900**	0.473	**0.897**



(a) *Effect of* h_i (b) *Effect of* h_l

(c) *Effect of* h_t (d) *Effect of* h_u

(e) *Effect of* d (f) *Effect of* λ

Figure 4: Parameter sensitivity study (with 50% dataset for training)

in online social networks like YouTube, Formspring, MySpace, Instagram, and Twitter [6, 7, 13, 37]. For example, Dinakar et al. [7] concatenated TF-IDF features, POS tags of frequent bigrams, and profane words as content features to investigate both explicit and implicit cyberbullying behaviors in negative text comments in YouTube and Formspring profiles [7, 8]. Xu et al. [37] presented several off-the-shelf tools such as Bag-of-Words models and LSA- and LDA-based representation to predict bullying traces on Twitter. Sanchez and Kumar [27] proposed the use of a Naïve Bayes classifier to identify instances of cyberbullying in Twitter. Dani et al. [6] proposed the *SICD* model, which leverages sentiment information to facilitate cyberbullying detection by capturing sentiment consistency of normal and bullying tweets. In [4, 5], the authors made use of gender-specific features and contextual features such as users' previous posts and the use of profane words to improve the performance of cyberbullying detection. Yao et al. [38] formulated cyberbullying detection on Instagram as a sequential hypothesis testing problem and gradually added text-based features based on the feature scores. With the increasing prevalence of social networking platforms, network-based features such as number of friends, network structure, and relational centrality have also been used more frequently to detect cyberbullying behaviors. For example, Homan et al. [12] studied the social structure of LGBT youth with depression in the TrevorSpace social network[6]. Huang et al. [15] studied a number of graph properties and found that cyberbullying detection performance was significantly improved when both network-based features and textual features were exploited. Tomkins et al. [34] proposed a socio-linguistic model that jointly detects cyberbullying content in messages, identifies participant roles, and exploits social interactions. Additional advances have come from newly-developed systems and applications to help identify cyberbullying risk on social network platforms, such as [28, 29]. These models aim to estimate the probability that an individual is experiencing cyberbullying from streams of received messages as well as various vulnerability (i.e., risk) factors. A similar work is discussed in [30], in which the authors studied the effectiveness of simply concatenating visual features and textual features.

Network Embedding Network embedding, which seeks to learn low-dimensional vector representations of nodes by exploiting different properties of the underlying network, has been successful in advancing a number of downstream learning tasks [11]. Significant advances have also resulted from the foundational work of DeepWalk [25], which makes an analogy between truncated random walk in the network and short sliding window across sentences in a text corpus, node2vec [10], which proposes a flexible notion of node neighborhood and employs a biased random walk procedure to explore neighborhoods of each node in a diversified way, and Tang et.al [32], who have proposed embedding large-scale information networks by carefully designing an objective function that preserves the first- and second-order node proximities. Notably, however, these prior contributions have focused on representation learning for homogeneous networks.

In recognition that many real-world information systems can be modeled as a heterogeneous information network, Dong et al. [9] proposed a heterogeneous network embedding model, metapath2vec, that formalizes meta-path-based random walks to construct the heterogeneous neighborhood of a node, and then leverages a heterogeneous Skip-gram model [21] to learn embeddings. Tang et al. [31] first presented a large-scale heterogeneous text network by jointly training a *word-word* co-occurrence network, a *word-document* bipartite network, and a *word-label* bipartite network with both labeled and unlabeled text data. Most recently, there has been growing interest in performing network embedding on attributed networks [16, 17], dynamic networks [19, 40], signed networks [36], and hypernetworks [35]. Because social media platforms have become a leading environment in which cyberbullying occurs, cyberbullying detection frameworks that take the rich multimodal nature of social media data into account are imperative. To our knowledge, we are the first to study the multi-modal cyberbullying detection problem using network representation learning. Due to the simplicity, scalability, and effectiveness of embedding models, our approach can significantly improve the quality of the features for cyberbullying detection.

6 CONCLUSIONS AND FUTURE WORK

With the growing popularity of social media platforms and increased social media use among teens, in particular, cyberbullying has become more prevalent and begun to raise serious societal concerns. Although they mark an important step forward, most previous efforts aimed at detecting cyberbullying have been based primarily on text analysis, and have thus failed to consider the multimodal nature of social media data (e.g., texts, images, likes/shares,

[6]https://www.trevorspace.org/

etc.). Our proposed model is based on the belief that multi-modal information can offer valuable insights for characterizing and detecting cyberbullying behaviors that both complement and extend previous work.

In this paper, we study the novel problem of cyberbullying detection within a multi-modal context. To address the challenges tied to multi-modal social media information, we propose an innovative cyberbullying detection framework, XBully, based on network representation learning. XBully first identifies representative mode hotspots to handle diverse feature types and then jointly maps both attributed and nominal nodes in a heterogeneous network into the same latent space by exploiting the cross-modal correlations and structural dependencies. Extensive experimental results on real-world datasets corroborate the effectiveness of the proposed framework. Future work directed towards building a deeper understanding of different modalities in characterizing cyberbullying behaviors will not only improve cyberbullying detection, but may also shed light on behaviors that are unique to users with different roles (e.g., victims, bullies) within cyberbullying interactions. Furthermore, we believe that the most promising and efficient path forward entails interdisciplinary collaboration among researchers in computer science and psychology to address this major social issue.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Oren Anava and Kfir Levy. 2016. k*-nearest neighbors: From global to local. In *Advances in Neural Information Processing Systems*. 4916–4924.
[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).
[3] Shaosheng Cao, Wei Lu, and Qiongkai Xu. 2015. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 891–900.
[4] Maral Dadvar and Franciska De Jong. 2012. Cyberbullying detection: a step toward a safer internet yard. In *Proceedings of the 21st WWW*. ACM, 121–126.
[5] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*. Springer, 693–696.
[6] Harsh Dani, Jundong Li, and Huan Liu. 2017. Sentiment Informed Cyberbullying Detection in Social Media. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 52–67.
[7] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 3 (2012), 18.
[8] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying. *The Social Mobile Web* 11, 02 (2011).
[9] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD*. ACM, 135–144.
[10] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD*. ACM, 855–864.
[11] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation Learning on Graphs: Methods and Applications. *arXiv preprint arXiv:1709.05584* (2017).
[12] Christopher M Homan, Naiji Lu, Xin Tu, Megan C Lytle, and Vincent Silenzio. 2014. Social structure and depression in TrevorSpace. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 615–625.
[13] 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.
[14] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Prediction of cyberbullying incidents in a media-based social network. In *ASONAM 2016*. IEEE, 186–192.
[15] 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*. ACM, 3–6.
[16] Xiao Huang, Jundong Li, and Xia Hu. 2017. Accelerated attributed network embedding. In *Proceedings of the 2017 SDM*. SIAM, 633–641.
[17] Xiao Huang, Jundong Li, and Xia Hu. 2017. Label informed attributed network embedding. In *Proceedings of the Tenth ICWSM*. ACM, 731–739.
[18] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2017. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 94.
[19] Jundong Li, Harsh Dani, Xia Hu, Jiliang Tang, Yi Chang, and Huan Liu. 2017. Attributed network embedding for learning in a dynamic environment. In *Proceedings of the 2017 CIKM*. ACM, 387–396.
[20] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.
[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[22] Parma Nand, Rivindu Perera, and Abhijeet Kasture. 2016. " How Bullying is this Message?": A Psychometric Thermometer for Bullying.. In *COLING*. 695–706.
[23] Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics* 33, 3 (1962), 1065–1076.
[24] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.
[25] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD*. ACM, 701–710.
[26] 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*. ACM, 617–622.
[27] Huascar Sanchez and Shreyas Kumar. 2011. Twitter bullying detection. *ser. NSDI* 12 (2011), 15–15.
[28] Yasin N. Silva, Christopher Rich, Jaime Chon, and Lisa M. Tsosie. 2016. BullyBlocker: An app to identify cyberbullying in facebook. In *ASONAM 2016*. 1401–1405.
[29] Yasin N. Silva, Christopher Rich, and Deborah Hall. 2016. BullyBlocker: Towards the identification of cyberbullying in social networking sites. In *ASONAM 2016*. 1377–1379.
[30] Vivek K Singh, Souvick Ghosh, and Christin Jose. 2017. Toward multimodal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2090–2099.
[31] Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD*. ACM, 1165–1174.
[32] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th WWW*. International World Wide Web Conferences Steering Committee, 1067–1077.
[33] Waldo R Tobler. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography* 46, sup1 (1970), 234–240.
[34] Sabina Tomkins, Lise Getoor, Yunfei Chen, and Yi Zhang. 2018. A Socio-linguistic Model for Cyberbullying Detection. In *2018 ASONAM*. IEEE, 53–60.
[35] Ke Tu, Peng Cui, Xiao Wang, Fei Wang, and Wenwu Zhu. 2018. Structural Deep Embedding for Hyper-Networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.
[36] Suhang Wang, Jiliang Tang, Charu Aggarwal, Yi Chang, and Huan Liu. 2017. Signed network embedding in social media. In *Proceedings of the 2017 SDM*. SIAM, 327–335.
[37] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics, 656–666.
[38] Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. 2018. Cyberbullying Detection on Instagram with Optimal Online Feature Selection. In *2018 ASONAM*.
[39] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. 2017. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *Proceedings of the 26th WWW*. International World Wide Web Conferences Steering Committee, 361–370.
[40] Lekui Zhou, Yang Yang, Xiang Ren, Fei Wu, and Yueting Zhuang. 2018. Dynamic Network Embedding by Modeling Triadic Closure Process. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.