Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model

Lu Cheng[†], Kai Shu[‡], Siqi Wu[⋆], Yasin N. Silva[†], Deborah L. Hall[†], Huan Liu[†]

[†]Arizona State University, [‡]Illinois Institute of Technology, [⋆]Australian National University
{lcheng35,ysilva,d.hall,huanliu}@asu.edu,kshu@iit.edu,siqi.wu@anu.edu.au

ABSTRACT

Social media is a vital means for information-sharing due to its easy access, low cost, and fast dissemination characteristics. However, increases in social media usage have corresponded with a rise in the prevalence of cyberbullying. Most existing cyberbullying detection methods are *supervised* and, thus, have two key drawbacks: (1) The data labeling process is often time-consuming and labor-intensive; (2) Current labeling guidelines may not be generalized to future instances because of different language usage and evolving social networks. To address these limitations, this work introduces a principled approach for *unsupervised* cyberbullying detection. The proposed model consists of two main components: (1) A *representation learning* network that encodes the social media session by exploiting multi-modal features, e.g., text, network, and time. (2) A *multi-task learning* network that simultaneously fits the comment inter-arrival times and estimates the bullying likelihood based on a Gaussian Mixture Model. The proposed model jointly optimizes the parameters of both components to overcome the shortcomings of decoupled training. Our core contribution is an unsupervised cyberbullying detection model that not only experimentally outperforms the state-of-the-art unsupervised models, but also achieves competitive performance compared to supervised models.

KEYWORDS

Cyberbullying Detection; Gaussian Mixture Model; Representation Learning; Social Media

ACM Reference Format:

Lu Cheng, Kai Shu, Siqi Wu, Yasin N. Silva, Deborah L. Hall, Huan Liu. 2020. Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340531.3411934

1 INTRODUCTION

Cyberbullying, defined as "aggressively intentional acts carried out by a group or an individual using electronic forms of contact, *repeatedly* or *over time* against victims who cannot easily defend themselves" [39], has been rising at an alarming rate. Previous research has found that nearly 43% of teens in the United States have been victims of cyberbullying [27]. In light of this, efforts aimed at automatically detecting cyberbullying – which seeks to predict whether or not human interactions within a social media session constitute cyberbullying – have a profound societal impact. However, detecting cyberbullying on social platforms is particularly challenging given that a social media session often consists of multi-modal information, for instance, an initial post, a sequence of comments, images/videos, and other social content such as the number of likes and shares. Fig. 1 illustrates an Instagram cyberbullying session where multiple bullying comments are posted.

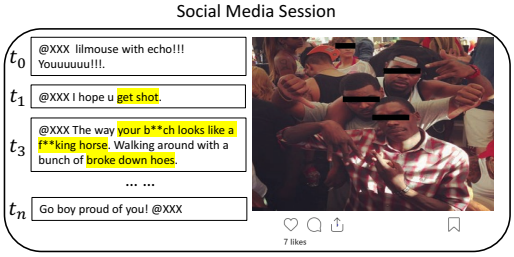


Figure 1: Illustration of a social media session on Instagram. Cyberbullying comments are repetitively posted by multiple users. Bullying words are highlighted. This work seeks to predict whether a given social media session is bullying.

Existing work on cyberbullying detection is mainly based on supervised methods, which often require a large annotated dataset for training. Although these approaches have shown promising results, they suffer from two major limitations: (1) Obtaining a large number of high-quality annotations for cyberbullying is time-consuming, labor-intensive, and error-prone because it requires circumspect examinations of multiple information sources such as images, videos, and numerous comments [18]; (2) Current guidelines for labeling a session as cyberbullying may not be effective in the future due to the dynamic nature of language usage and social networks. Hence, we study alternative mechanisms for *unsupervised* cyberbullying detection, which draws inferences from input social media data but without labeled responses.

Despite potential benefits, unsupervised cyberbullying detection also encounters several challenges: (1) Because cyberbullying typically consists of repetitive acts (as shown in Fig. 1), the temporal dynamics of users' commenting behaviors adds nuanced understandings to the text-based methods that consider each comment as a distinct event over time. Such temporal characterization have

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3411934

been shown to be useful in distinguishing cyberbullying from non-bullying instances [6, 17, 41]. Therefore, a key challenge is how to simultaneously model temporal dynamics and cyberbullying detection such that the two tasks mutually improve each other. (2) Social media sessions inherently present a *hierarchical structure* where words form a comment and comments form a session. Previous studies [6, 47] have revealed that modeling the hierarchical structure is useful for learning high-quality representations. Additionally, because meanings of words and comments are largely context-dependent, the sequential structure of words and comments need to be properly modeled for identifying relevant ones (e.g., the highlighted words in Fig. 1); (3) A straightforward approach for unsupervised cyberbullying detection is to use the off-the-shelf clustering algorithms (e.g., k-means). The effectiveness of this approach largely relies on the quality of input data, however, social media data is notorious for its noise, sparsity, and high-dimensionality. Applying dimensionality reduction to the input data still presents the drawback of *decoupled training*, i.e., representation learning and clustering are carried out separately.

To address these challenges, we propose a principled unsupervised learning framework – Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model (UCD). A central feature of UCD is that it incorporates the comment inter-arrival times of a social media session, which enables the classification of cyberbullying instances using the full commenting history. UCD consists of two main components: a *representation learning* network, which learns the compact multi-modal representations of a session; and a *multi-task learning* network, which predicts whether or not a session contains bullying behaviors while modeling the temporal dynamics of all comments. Specifically, the representation learning network models social media sessions using a Hierarchical Attention Network (HAN) [47] for textual features and a Graph Auto-Encoder (GAE) [23] for user and network features. The multi-task learning network then takes the multi-modal representations (e.g., text, user, and social network) as input to estimate the bullying likelihood using a time-informed Gaussian Mixture Model (GMM). The two UCD components are jointly optimized to mutually boost their learning effectiveness.

The main contributions of this paper are:

- We address the problem of unsupervised cyberbullying detection in social media platforms, which automatically identifies bullying instances without labeled data.
- We propose a principled framework for unsupervised cyberbullying detection, which includes two components that jointly learn low-dimensional representations and predict bullying instances.
- We conduct experiments on two real-world social datasets from Instagram and Vine. Our results show that UCD not only outperforms the state-of-the-art unsupervised models, but also achieves competitive performance against supervised models[1].

2 RELATED WORK

We review related work on automatic cyberbullying detection models and clustering algorithms based on deep neural networks.

[1]Code available at https://github.com/GitHubLuCheng/UCD

2.1 Cyberbullying Detection

To date, cyberbullying has received a significant amount of attention within psychology and social science fields. It has only more recently become a focus of computer science research, where much of the work has been aimed at developing models that automatically identify bullying behaviors. For instance, existing work on automatic cyberbullying detection has used manually labeled data to mine patterns from text [6, 9, 13, 28, 35, 45], social network [3, 20, 26], and other media sources such as images and videos [8, 18, 19, 31, 32]. Xu et al. [45] explored several natural language processing (NLP) techniques to identify bullying traces and further defined the structure of a bullying episode and the associated roles (e.g., victims and bullies) on Twitter. Dinakar et al. [13] concatenated TF-IDF scores, POS tags of frequent bigrams, and profane words as content features to detect cyberbullying on a manually-labeled corpus of YouTube comments. Dani et al. [9] sought to incorporate sentiment into the content features by capturing the sentiment consistency of bullying and non-bullying posts. Most recently, Ziems et al. [49] characterized cyberbullying using five explicit factors to represent its social and linguistic aspects.

Although many researchers define cyberbullying as a harmful behavior that is repeated over time, relatively little work has examined the temporal aspects of cyberbullying. Among the few studies that have, Soni and Singh [41] modeled the commenting behaviors as Poisson point processes and identified several temporal features that help distinguish bullying sessions. Cheng et al. [6] employed a hierarchical attention network to capture the sequence-aware structure of words and comments in a social media session and integrated time interval prediction into the detection model. From a causality perspective, Cheng et al. [5] sought to discover the potential confounders among bullying texts so that the resulting classifiers can be transferred between different domains.

Crucially, most existing work on cyberbullying detection has focused on supervised learning models that require large-scale labeled datasets. To reduce this dependency on human-coded data, Raisi and Huang [34] proposed a weakly-supervised model that starts with a small seed vocabulary of bullying indicators. They then extracts bullying roles and additional bullying indicators based on an unlabeled corpus of social media interactions. Another work [33] studied cyberbullying detection with an ensemble of two learners that co-train one another; one learner examines the language content in the messages while the other considers the social structure. To our knowledge, the only unsupervised cyberbullying detection model GHSOM [10] inputs several NLP and social features into the Growing Hierarchical SOM using the SOMToolbox framework.

2.2 Deep Clustering

Clustering methods based on deep neural networks have shown promising results in real-world applications (e.g., anomaly detection [50]) due to their high representational power. Standard clustering-friendly representations are learned with a two-phase training procedure. In the first phase, the auto-encoder is trained with the mean squared error reconstruction loss. In the second phase, the auto-encoder is further fine-tuned with a combined loss function consisting of the reconstruction loss and a clustering-specific loss.

For example, Song et al. [40] applied an auto-encoder in the clustering tasks and introduced a new objective function that includes the reconstruction error and the distance between data and their corresponding cluster centers in the latent space. Similarly, the Deep Embedded Clustering model in [44] projected the data from an original space to a lower-dimensional feature space and then jointly optimized a clustering objective using stochastic gradient descent via backpropagation. Relevant to the present work, multiple studies have employed unsupervised anomaly detection [2]. For instance, Zong et.al [50] jointly optimized the parameters of a deep auto-encoder and a mixture model with the two components mutually improving each other's performance.

In contrast to most existing cyberbullying detection models, UCD focuses on the unsupervised approach where labeled data is not available during training. To achieve good performance, we exploit multi-modal data and relevant information such as the temporal patterns of comments and the hierarchical structure of social media sessions. Our evaluation results show that the integration of this additional information can significantly improve the effectiveness of unsupervised cyberbullying detection.

3 UCD: THE PROPOSED FRAMEWORK

The framework overview in Fig. 2 shows that our model consists of two major components: (1) a representation learning network that leverages HAN and GAE to obtain multi-modal representations, and (2) a multi-task learning network that jointly optimizes a GMM-based energy estimation task to detect cyberbullying instances and a temporal prediction task to further refine the session representations with the comment inter-arrival times.

3.1 Representation Learning Network

Social media sessions usually consist of multi-modal information, such as text (e.g., comments) and social content (e.g., friendship networks, number of likes and shares). The representation learning network aims to transform these sparse and high-dimensional features into a low-dimensional session representation.

HAN for Text. The majority of prior literature on cyberbullying detection considered the comments in a social media session as independent events and directly extracted textual features from a chunk of combined comments. Notwithstanding its simplicity, this method largely overlooks the hierarchical structure of a social media session and the long-term dependencies among the sequentially posted comments. Previous studies showed that i) modeling document structure can significantly improve the quality of document representations [47]; and ii) capturing long-term dependencies is particularly useful for sequential data modeling [11]. In addition, different words and comments in a post are not equally relevant for cyberbullying detection, i.e., some words/comments are more important than others. For example, *"You're a f**king loser!"* and *"Yeah, I'm a loser."* both include the word *loser*, the former is, however, more likely to represent an instance of bullying. Therefore, we also integrate attention mechanisms to distinguish important words and comments. Following [6], we employ a hierarchical attention network to generate the textual representation for a social media session. The HAN approach is a particularly good fit in cyberbullying detection as it models the two main levels of social media sessions (sequences of words and comments) and at each level, the model captures the long-term dependencies and integrates mechanisms to differentiate the importance of specific words and comments based on their context.

The hierarchical structure of the textual content can be described as follows: a social media session consists of a sequence of comments and each comment includes a sequence of words. Given a session with C comments where each comment i has L_i words $\{w_{it} | t = 1, 2, ..., L_i\}$, we use the bi-directional Gated Recurrent Units (GRUs) [1] to model both the word sequence in a comment and the comment sequence in a session:

$$\begin{aligned} \overrightarrow{s}_{it} &= \overrightarrow{GRU}(W_e w_{it}), \quad \forall t \in [1, L_i], i \in [1, C] \\ \overleftarrow{s}_{it} &= \overleftarrow{GRU}(W_e w_{it}), \quad \forall t \in [L_i, 1], i \in [1, C] \end{aligned} \tag{1}$$

where each word w_{it} is first mapped to a latent space with parameter W_e. The resulting annotation for word w_{it} is a concatenation of the forward and backward hidden states, $s_{it} = [\overrightarrow{s}_{it}, \overleftarrow{s}_{it}]$. To differentiate the word importance, we adopt the attention mechanism [1, 47] to automatically detect words that are more relevant and then aggregate the representation of weighted words to form a comment vector c_i:

$$\alpha_{it} = \frac{\exp(h_{it}^T u_w)}{\sum_t \exp(h_{it}^T u_w)}; \quad c_i = \sum_t \alpha_{it} s_{it}, \tag{2}$$

where h_{it} is the output of a fully connected layer of s_{it} and u_w denotes a word-level context vector [47]. α_{it} denotes a normalized weight describing the importance of word w_{it}. Similarly, the final textual representation v of a social media session can be computed using the encoded comment vectors (i.e., replacing w_{it} of Eq.1 with c_i). Further, we include a dense layer to project the social content, i.e., number of likes and shares, into a latent space. We later concatenate the resulting vector p with v to form the multi-modal representation of a social media session $o = [v, p]$.

GAE for Attributed Social Networks. Self-selection bias (grouping with similar others) and peer influence are closely connected with bullying behaviors in offline environments [7, 14, 37, 43]. Research in human communication [15] reveals a similar observation that online social network positioning is a comparably strong predictor for cyberbullying detection. Hence, it is important to consider the social network structure and peer influence from similar users for improving the performance of cyberbullying detection.

The representation learning network learns user representation by exploiting information from social networks where nodes denote social media users with corresponding profile information being the node attributes, and edges denote the follower/followee relationships. Here, we employ GAE to embed users' attributes as low-dimensional vectors such that users with structural proximity in the social network are close. As one of the most powerful node embedding approaches, GAE has been applied to several challenging learning tasks such as link prediction [16, 23] and node clustering [36]. GAE can effectively incorporate node features and learn more interpretable user representations [23]. The key of GAE is the encoding-decoding scheme, i.e., GAE encodes nodes into low-dimensional vectors which are then decoded to reconstruct the original network structure. Suppose we are given a social network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $U = |\mathcal{V}|$ users. The adjacency matrix of this

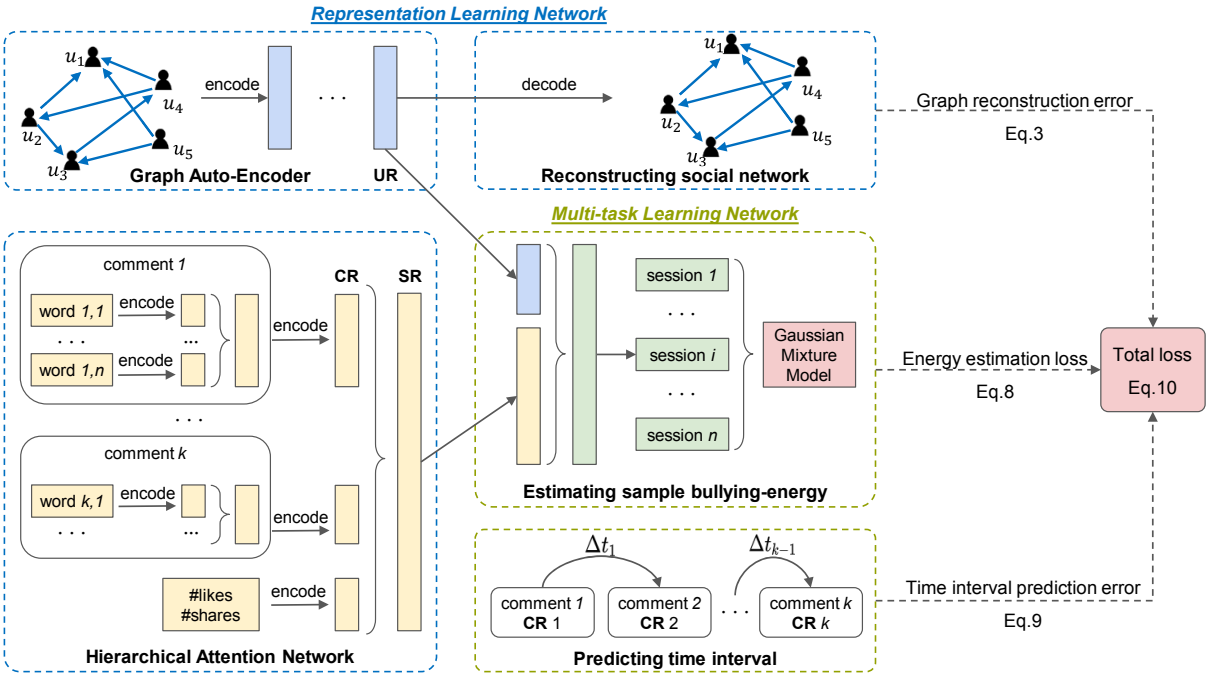


Figure 2: Overview of the proposed framework. UCD consists of two components: (1) The *representation learning network* (the blue dashed rectangles) constructs multi-modal representations of social media sessions (the green solid rectangles in the middle); (2) The *multi-task learning network* (the green dashed rectangles) that simultaneously estimates the energy/likelihood of input samples and predicts time intervals between comments. Observe that the representation learning network combines user (session owner) representation (UR) in the Graph Auto-Encoder (top part) and social representation (SR) in the Hierarchical Attention Network (HAN, bottom left) to form the session representation. The constructed session representation is the input of the sample bullying energy estimation task. Meanwhile, the comment representations (CR) in HAN are fed into the time interval prediction task (bottom part). The overall loss comes from three sources: graph reconstruction error, energy estimation loss, and time interval prediction error. Best viewed in colors.

graph is $A \in \mathbb{R}^{U \times U}$. The User-Feature matrix is $X \in \mathbb{R}^{U \times D}$ with D being the feature dimension. GAE then uses a graph convolutional network (GCN) [22] encoder and an inner product decoder to learn a latent matrix Z by minimizing the following reconstruction error:

$$g = \frac{1}{2}\|A - \hat{A}\|_2^2, \quad \text{with } \hat{A} = \sigma(ZZ^T), Z = \text{GCN}(X, A) \tag{3}$$

where $\sigma(\cdot)$ is the logistic sigmoid function.

The final representation of a session is the concatenation of user (owner) representation and the representation output from HAN, i.e., $ss = [z, o]$, where z is a row vector of Z. This multi-modal representation is then fed into the multi-task learning network.

3.2 Multi-Task Learning Network

Given the multi-modal representation of input sessions, the multi-task learning network simultaneously (1) estimates the sample bullying-energy/likelihood; and (2) models the inter-arrival times of a sequence of comments in a social media session. These two tasks can mutually enhance each other's performance in the training stage. To this end, the multi-task learning network enables the proposed framework to jointly learn session representations and discover cyberbullying instances.

Bullying-energy estimation. The first task of the multi-task learning network is to estimate the sample energy (likelihood) and classify samples with high energy (low likelihood) as bullying instances. A primary benefit of energy-based models is the flexibility to specify the energy expression [48]. Here, we construct a GMM-based density estimator to infer the underlying probability density function. GMM, a widely used unsupervised learning method, seeks to fit a multi-modal distribution with multiple unimodal Gaussian distributions which are the most commonly used distributions for modeling real-world unimodal data. Previous work [48, 50] has shown that GMM is more effective than simple models for data with complex structures. Given the complexity and multi-modal nature of social media data, we leverage GMM to perform density estimation tasks over multi-modal representations.

Let the number of mixture components be K and the latent representation of a social media session be ss, we first generate the mixture membership predictions for ss. We then estimate the parameters of GMM using the predicted membership to obtain the energy estimation of ss. Specifically, we first feed ss into a multi-layer network (MLN) [42] parameterized by θ_m. The output

is denoted as p_{MLN}:

$$p_{MLN} = \text{MLN}(ss; \theta_m) \tag{4}$$

The probability of ss belonging to each component can be estimated as follows:

$$\hat{m} = \text{softmax}(p_{MLN}) \tag{5}$$

where $\hat{m}$ is a K-dimensional vector. Given a batch of N social media session representations $\{ss_1, ss_2, ..., ss_N\}$, together with the corresponding predicted memberships, we can further estimate the parameters in GMM as follows:

$$\hat{\phi}_k = \sum_{i=1}^{N} \frac{\hat{m}_{ik}}{N}; \quad \hat{\mu}_k = \frac{\sum_{i=1}^{N} \hat{m}_{ik} ss_i}{\sum_{i=1}^{N} \hat{m}_{ik}} \tag{6}$$

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{N} \hat{m}_{ik}(ss_i - \hat{\mu}_k)(ss_i - \hat{\mu}_k)^T}{\sum_{i=1}^{N} \hat{m}_{ik}} \tag{7}$$

where $\hat{\phi}_k$, $\hat{\mu}_k$ and $\hat{\Sigma}_k$ denote the mixture probability, mean, and covariance of component $k \in \{1, 2, ..., K\}$ in GMM, respectively. $\hat{m}_{ik}$ denotes the probability of ss_i in the k-th component of GMM. To build the probability density function, we leverage the energy-based model [24] which relies on a specific parameterization of the energy (negative log likelihood). The energy level of a session is defined as:

$$E(ss_i; \theta_m) = -\log\Big(\sum_{k=1}^{K} \hat{\phi}_k \frac{\exp\big(-\frac{1}{2}(ss_i - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1}(ss_i - \hat{\mu}_k)\big)}{\sqrt{|2\pi\hat{\Sigma}_k|}}\Big) \tag{8}$$

where $|\cdot|$ is the determinant of a matrix. The model then classifies a session as cyberbullying if its energy is above a predefined threshold $\tau \in (0,1)$ in the testing phase. In practice, τ is typically set to a comparatively large value, i.e., a cyberbullying session is in general associated with high energy (hence low likelihood). This is because bullying samples are less frequently observed in real-world datasets, as suggested by the statistics in Table 1 as well as in previous literature [7, 12].

Temporal dynamics fitting. Cyberbullying is commonly defined as a *repeated act* of aggression that develops over time [6, 12, 41]. However, most of the existing computational models consider each comment in a social media session as an isolated event. Therefore, they largely overlook the temporal dynamics of users' commenting behavior. Here, we seek to predict the inter-arrival times between comments for obtaining additional feedback from the temporal dynamics. This feature enables the model to exploit the commonalities and differences across bullying-energy estimation and temporal-dynamics prediction for improving the final cyberbullying detection performance.

We first obtain the output e_{in} of the comment encoder for comment i in session n from the HAN module and then conduct a time interval prediction task as follows.

$$\ell = \sum_{i=1}^{C} \frac{1}{2} \|f(e_{in}; \theta_\ell) - \Delta t_i\|^2, \tag{9}$$

where f represents a regression model, θ_ℓ denotes the associated parameters, and $\Delta t_i = t_i - t_{i-1}$ is the time interval between comment $i-1$ and i. We set t_0 to be 0. Let d denote the dimensions of the latent representation of social media sessions, θ_h the parameters of HAN and θ_g the parameters of GAE, the final objective function of UCD can be constructed as:

$$\begin{aligned} J = &\sum_{n=1}^{N} \sum_{i=1}^{C} \frac{1}{2} \|f(e_{in}; \theta_\ell) - \Delta t_i\|^2 + \frac{\lambda_1}{N} \sum_{i=1}^{N} E(ss_i; \theta_m) \\ &+ \frac{\lambda_2}{2} \|A - \hat{A}\|_2^2 + \lambda_3 P(\hat{\Sigma}); \text{ with } P(\hat{\Sigma}) = \sum_{k=1}^{K} \sum_{j=1}^{d} \frac{1}{\hat{\Sigma}_{kjj}} \end{aligned} \tag{10}$$

$P(\hat{\Sigma})$ accounts for the singularity issue in GMM, λ_1, λ_2, and λ_3 are the hyperparameters that control the balance among time interval prediction error, energy estimation loss, graph reconstruction error and regularization for GMM. Specifically, the objective function consists of four components (ordered as presented in Eq. 10):

- The first component is the loss function that describes the prediction error of time interval prediction.
- The second component $E(ss_i; \theta_m)$ models the likelihood (sample energy) that session i is observed. Here, minimizing the energy level of an input session will maximize the likelihood of observing the session.
- The third component is the reconstruction error of GAE in the representation learning network. A lower error indicates that the learned user representations better preserve the structure of the original attributed social network.
- Due to the singularity issue in GMM, we penalize small values on the diagonal entries of the covariance matrices $\hat{\Sigma}$.

The proposed model jointly optimizes the representation learning network and the multi-task learning network to learn effective representations for cyberbullying detection. We train the model by minimizing Eq. 10 using the Adam optimization algorithm [21], where the error backpropagates through the representation learning network, the bullying-energy estimation task, and the time-interval prediction task.

4 EVALUATION

In this section, we present both quantitative and qualitative analyses to evaluate the proposed UCD framework. Specifically, we answer the following research questions:

(1) *Effectiveness*: **a.** How effective is UCD compared to existing unsupervised learning approaches and supervised classification models? **b.** How does each module, i.e., HAN, GAE, and temporal modeling, affects the cyberbullying detection performance of UCD?

(2) *Robustness*: How robust is UCD when varying model parameters?

4.1 Datasets

Our experiments use two public datasets crawled from Instagram[2] and Vine[3] (now in archive status). The datasets were introduced and released in [18] and [31], respectively. The basic statistics of these datasets are presented in Fig. 1.

Instagram: Instagram is a popular social media platform. It is also the platform on which the highest prevalence of cyberbullying has been reported [29]. Using a snowball sampling method, the authors in [18] identified 41K Instagram users, 61% of whom had public profiles. For each public user, the collected data includes

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

[3]https://vine.co/

Table 1: Basic statistics for *Instagram* and *Vine* datasets.

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

the media objects the user had posted, the comments of session, the list of user followers/followees, and the list of users who have commented/liked the media objects. Data labeling (whether the session constituted cyberbullying or not) was conducted on CrowdFlower[4] – a crowdsourcing website – using a procedure whereby each session was labeled by five different contributors. A session is labeled as cyberbullying if three or more contributors had labeled this session as cyberbullying. Overall, the *Instagram* dataset includes 2,218 labeled social media sessions.

Vine: The Vine dataset [31] is used for analyzing cyberbullying in the context of a video-based online platform. It was crawled using a snowball sampling method in which a random user u is first selected as a seed and then the crawling continues with the users that u follows. Each session includes videos, captions, and associated comments (note that social network information was not available for this dataset). All sessions in the dataset have at least 15 comments. Similar to the labeling process used for the *Instagram* data, a total of 970 *Vine* sessions were labeled (as cyberbullying vs. non-bullying) using CrowdFlower.

We use the following information gathered from a media session:

- Attributed social network: A social network where each node represents a user and has attributes such as the number of total followers and followees. The edges denote the following and followed-by relationships.
- Text: The bag-of-words representation of the captions and comments. Each column indicates a term from the corpus and the entry is the corresponding frequency count.
- Time: The posting timestamps of a media object and its associated comments. We extract the time difference between any two consecutive comments.
- Social content: The number of likes and shares of a post receives.

4.2 Experimental setup

To answer the first research question, we compare UCD with multiple unsupervised learning models:

- **k-means.** k-means is one of the most common clustering algorithms. It iteratively assigns each data point to one of k groups with the smallest distance.
- **HAE** [25]. HAE is an LSTM model that hierarchically builds embeddings for social media sessions from comments and words. We also used k-means to cluster the learned representations.
- **DCN** [46]. DCN is a deep learning-based clustering algorithm that regulates auto-encoder performance by using k-means.
- **DAGMM** [50]. DAGMM jointly optimizes a deep auto-encoder that learns low-dimensional representations and a GMM that estimates the density function of the latent representations.
- **XBully** [8]. XBully learns multi-modal representations of social media sessions and then feeds them into a subsequent classification model. We replaced the classification model with k-means.
- **GHSOM** [10]. To our knowledge, Growing Hierarchical Self-Organizing Map (GHSOM) is the only existing model for unsupervised cyberbullying detection. It extracts sentiment, syntactic, and semantic features from text and social network data. The features are then fed into the GHSOM tool[5] for clustering.

To provide a comprehensive analysis of UCD, we also include the following supervised methods:

- **Naïve Bayes (NB)**. NB is a probabilistic classifier based on Bayes' theorem with strong independence assumptions between the features. It is one of the most popular (baseline) methods for text classification.
- **Random Forest (RF)**. RF consists of several individual decision trees that operate as an ensemble. Each individual tree generates a class prediction and the class with the most votes becomes the model's prediction.
- **Logistic Regression (LR)**. LR is a statistical model that uses a logistic function to model a binary dependent variable. It is a common baseline algorithm for binary classification.

For baselines using k-means, we set the number of clusters to 2, and label the cluster with fewer elements as *bullying* and the other one as *non-bullying*. This assumption is supported by the statistics in Table 1 and also generally evident in other real-world cyberbullying datasets [49]. Note that our proposed method (UCD) does not require this assumption as it optimizes Eq. 10 for clustering bullying and non-bullying instances. We implemented the following variants of UCD to examine the impact of each UCD component.

- **UCDXtext.** UCD without HAN. We do not report this variant for *Vine* given that its social network information is not available.
- **UCDXtime.** UCD without time interval prediction.
- **UCDXgraph.** UCD without GAE.

Following previous literature [7, 41], we use four common evaluation metrics – *Precision*, *Recall*, *F1*, and *AUROC (Area Under the Receiver Operating Characteristic Curve)*. Note that we are more interested in detecting cyberbullying instances, therefore, we report Precision, Recall and F1 corresponding to the bullying (positive) class. While the overall performance can be effectively measured by F1 and AUROC scores, multiple application scenarios of cyberbullying detection could particularly benefit from the identification of as many positive cases as possible, i.e., high Recall.

Parameter Setting. Based on Eq. 10, the UCD framework has five hyperparameters: (1) λ_1, for balancing the sample bullying-energy loss; (2) λ_2, for controlling the weight of the reconstruction error of GAE; (3) λ_3, for controlling the weight of diagonal entries in the covariance matrices; (4) K,[6] the number of mixtures in the GMM; and (5) $\tau \in (0, 1)$, a pre-defined energy threshold. We set the parameters based on sensitivity analysis, which is detailed in Section 4.5. Specifically, we set $\lambda_1 = 1e-4$, $\lambda_3 = 1e-9$ and $K = 5$ for both datasets. The energy threshold τ is set to 65% for *Instagram* and 70% for *Vine*. Therefore, Instagram and Vine test sessions with

[4] http://www.figure-eight.com/

[5] http://www.ifs.tuwien.ac.at/ andi/ghsom/

[6] This is different from the k in k-means, which sets the number of clusters (bullying and non-bullying). K in GMM denotes the number of memberships and relates to computing the sample energy. We use the energy threshold to detect bullying instances.

Table 2: Performance evaluation with *Instagram* data.

Unsupervised Learning Models				
Metrics	Precision	Recall	F1	AUROC
k-means	0.79±0.02	0.29±0.04	0.43±0.05	0.63±0.02
XBully	0.32±0.02	0.47±0.03	0.38±0.02	0.51±0.02
HAE	0.53±0.02	0.27±0.03	0.35±0.03	0.53±0.01
DCN	**0.87±0.02**	0.23±0.02	0.36±0.02	0.61±0.01
DAGMM	0.56±0.18	0.56±0.18	0.56±0.18	0.56±0.03
GHSOM	0.35±0.12	0.38±0.06	0.36±0.08	0.54±0.11
UCDXtext	0.33±0.01	0.34±0.01	0.33±0.01	0.53±0.02
UCDXtime	0.47±0.02	0.48±0.01	0.48±0.01	0.63±0.01
UCDXgraph	0.56±0.02	0.57±0.01	0.57±0.02	0.69±0.01
UCD	0.59±0.02	**0.66±0.02**	**0.63±0.02**	**0.73±0.01**
Supervised Learning Models				
Metrics	Precision	Recall	F1	AUROC
NB	0.40±0.03	**0.69±0.03**	0.51±0.03	0.62±0.02
RF	0.78±0.03	0.53±0.03	0.63±0.03	0.73±0.01
LR	**0.79±0.03**	0.55±0.03	**0.64±0.03**	**0.74±0.03**

Table 3: Performance evaluation with *Vine* data.

Unsupervised Learning Models				
Metrics	Precision	Recall	F1	AUROC
k-means	0.03±0.08	0.00±0.00	0.00±0.01	0.50±0.00
XBully	**0.48±0.08**	0.27±0.03	0.34±0.04	0.57±0.02
HAE	0.18±0.04	0.34±0.08	0.23±0.04	0.57±0.03
DCN	0.29±0.20	0.32±0.39	0.22±0.19	0.50±0.03
DAGMM	0.36±0.09	0.31±0.08	0.33±0.08	0.54±0.00
GHSOM	0.32±0.09	0.38±0.10	0.34±0.08	0.50±0.07
UCDXtime	0.33±0.02	0.39±0.03	0.36±0.02	0.56±0.01
UCDXgraph	0.43±0.02	**0.40±0.03**	**0.41±0.02**	**0.58±0.01**
Supervised Learning Models				
Metrics	Precision	Recall	F1	AUROC
NB	0.49±0.05	**0.72±0.05**	0.58±0.04	0.70±0.04
RF	**0.67±0.05**	0.42±0.05	0.51±0.04	0.66±0.02
LR	0.62± 0.05	0.57±0.05	**0.59±0.04**	**0.71±0.03**

the highest 35% and 30% energy values will be classified as bullying cases and the rest as non-bullying cases, respectively. For *Instagram*, we additionally set $\lambda_2 = 0.01$. For the baseline methods, we conducted similar sensitivity analysis on the key parameters reported in their original papers. For both datasets, we use 80% of the data for training and the rest for testing. Each experiment is run 10 times, mean and standard deviations are reported.

4.3 Quantitative Results

For the *Instagram* dataset, we compare UCD and its variants with all baselines. Due to the lack of social network information in the *Vine* dataset, UCD and UCDXtext cannot be evaluated with *Vine*. The best results for unsupervised and supervised models are highlighted in Table 2 and 3 with bold text. The results we present for RF are different from those reported in [32]. We believe this is the case because the original work: 1) considered additional features such as the percentage of negative comments, emotions exhibited in videos, and latent semantic features (10 topics based on the comments using LDA), and 2) performed oversampling (SMOTE [4]) to balance the *Vine* dataset. We use the original *Vine* dataset to better reflect real-world scenarios.

We observe that (1) UCD achieves the best performance in Recall, F1, AUROC, and competitive Precision compared to the unsupervised baselines for both datasets. For the *Instagram* dataset, UCD shows 15.9%, 19.7%, and 35.2% of improvement on AUROC compared to the results using raw features (i.e., k-means), representation learning (i.e., DCN), and the unsupervised cyberbullying detection model GHSOM, respectively. AUROC considers all possible thresholds for classification and is a more appropriate metric when datasets are imbalanced; (2) Imbalanced datasets affect the trade-off between Recall and Precision. While achieving superior Precision, baseline models DCN and k-means show poor Recall. We infer that these models fail to identify most of the cyberbullying instances, which is undesired in many cyberbullying applications; and (3) UCD achieves competitive Recall, F1 and AUC scores compared to supervised methods using the *Instagram* dataset. For instance, LR improves F1 by 1.6% over UCD whereas NB is outperformed by UCD regarding these three metrics. The Precision of UCD is comparatively low implying that its energy threshold favors identifying cyberbullying instances, therefore, UCD miss-classifies more non-bullying instances than baseline methods. In the *Vine* dataset, the supervised methods show larger advantages over UCDXgraph, reflecting the importance of integrating social network information and using larger datasets in order to maximize the performance of UCD. Of particular interest is that UCD also achieves more balanced Precision and Recall values compared to supervised models.

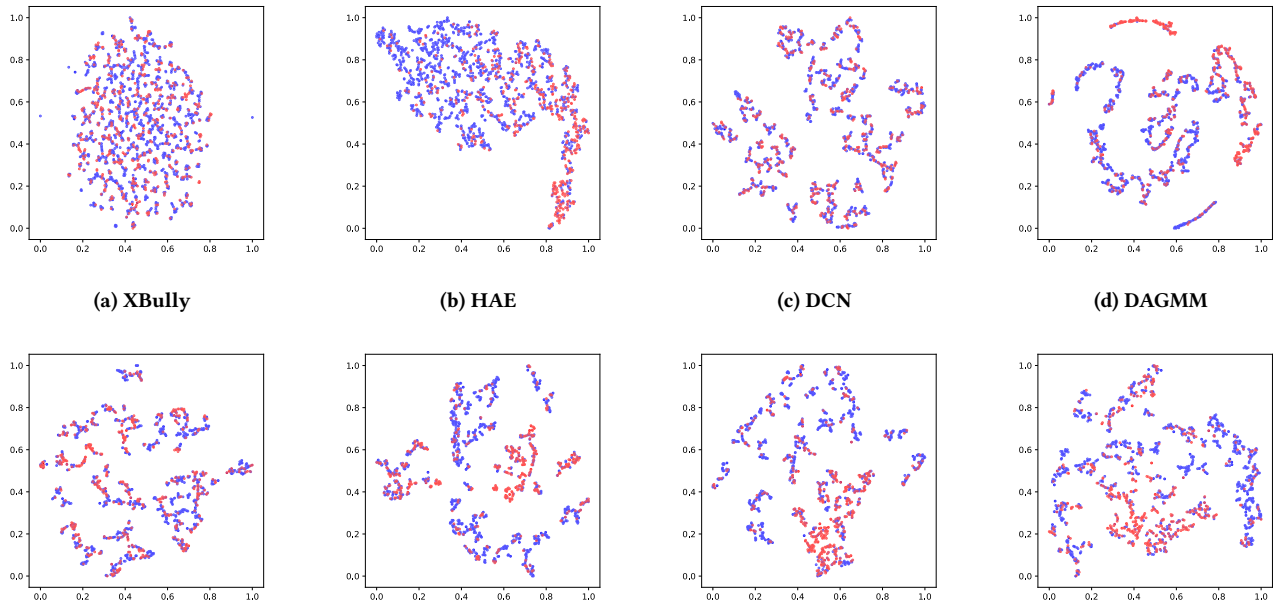
We make the following observations when comparing UCD with its own variants: (1) UCD achieves better performance in all metrics, especially against UCDXtext and UCDXtime, leading us to conclude that each submodule (HAN, GAE, and temporal analysis) has a positive influence on UCD's performance; (2) The performance of UCDXtext drops significantly compared to other variants, highlighting the importance of textual features in cyberbullying detection; (3) UCDXgraph outperforms UCDXtime, indicating that temporal analysis can provide more relevant information for cyberbullying detection than social network properties and thus highlighting the importance of modeling temporal patterns; and (4) the proposed framework performs better on *Instagram* data than on *Vine* data. This is in part due to the smaller sample size and lack of social network information in the *Vine* dataset.

In summary, UCD outperforms unsupervised baselines in terms of identifying cyberbullying instances and the overall performance. Compared to supervised models, it shows competitive performance when the sample size is comparatively large and the social network information is available. None of the evaluated methods achieves high performance in detecting both bullying and non-bullying instances. Future work is encouraged to investigate such methods.

4.4 Qualitative Analysis

We further investigate the qualities of the learned multi-modal representations using t-SNE visualizations in Fig. 3. Taking *Instagram* as an example, we make the following observations:

- As shown in Fig. 3(h), UCD better separates the bullying and non-bullying samples in the latent space. The results of most of



(a) XBully (b) HAE (c) DCN (d) DAGMM

(e) UCDXtext (f) UCDXtime (g) UCDXgraph (h) UCD

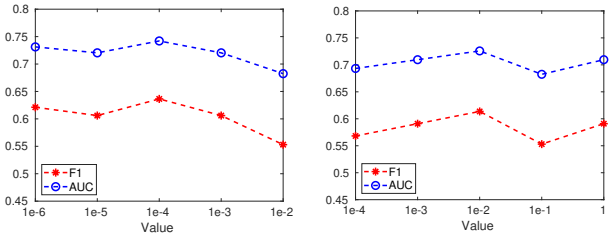
Figure 3: t-SNE visualizations of the low dimensional representations using the *Instagram* dataset. The red dots denote instances of the bullying class and the blue points instances of the non-bullying class. Best viewed in colors.

the other models, particularly XBully, HAE, DCN, and UCDXtext, yield more overlapped clusters.

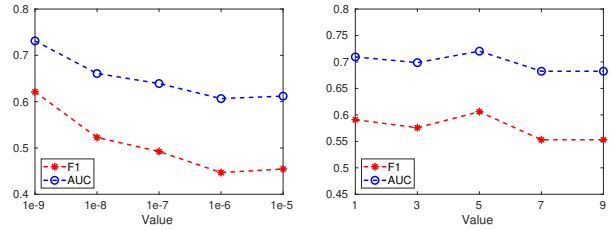
- From the results of DAGMM and UCD, we observe that models with GMM can learn discriminative representations, which is evident by the greater separation between bullying and non-bullying clusters). The overall performance of UCD is better than DAGMM, indicating that UCD benefits from the joint optimization of cyberbullying detection and time interval prediction.
- Both UCD and DAGMM outperform DCN. With a pre-trained auto-encoder, DCN can get easily stuck in a local optimum for achieving lower reconstruction error and could be suboptimal for the subsequent density estimation tasks [50]. A joint optimization of representation learning, bullying-energy estimation, and time interval prediction can help avoid these local optimal cases and achieve better learning performance.
- In contrast to other baseline methods, such as XBully and DCN, HAE in Fig. 3(b) generates large regions that are primarily populated by either bullying or non-bullying samples. This confirms that modeling the hierarchical structure of a session has an important impact in cyberbullying detection.
- UCDXtime produces two main bullying clusters (two red clusters), UCDXgraph generates similar results to UCD, and UCDXtext fails to learn discriminative representations, evidenced by the overlap between the bullying and non-bullying clusters.

4.5 Parameter Analysis

The UCD model has five core parameters ($\lambda_1, \lambda_2, \lambda_3, K, \tau$) for balancing the weights of bullying-energy estimation loss, reconstruction error, regularization of the covariance matrices, the number of mixtures in GMM, and the energy threshold, respectively. Here, we further divide the training data into training (80%) and validation (20%) sets. To investigate the effects of the first four parameters, we run experiments on the *Instagram* dataset varying one parameter at a time and evaluate how it affects the overall performance. We show the sensitivity analysis w.r.t. AUROC and F1 scores in Fig. 4. We observe that large λ_1 that overemphasizes the energy estimation loss can lead to poor performance regarding both F1 and AUROC scores. The trend of varying K is similar to that of λ_1, i.e., the performance drops when the number of components in GMM becomes too large. The best performance is obtained when λ_1 is set to 1e−4 and K is set to 5. In contrast, the performance of varying λ_2 displays an ascending trend in a certain range as shown in Fig. 4(b). The UCD model with a slightly large λ_2 controlling the importance of GAE is more likely to obtain better results. Unsurprisingly, when the covariance matrices in GMM are given too much penalization, i.e., a large λ_3, the F1 and AUROC scores decrease significantly, as shown in Fig. 4(c). The last parameter τ represents the threshold for identifying bullying instances. Given that UCD largely relies on τ for cyberbullying detection, we use both *Instagram* and *Vine* datasets to examine its influence. The results are presented in Fig. 5. It shows that UCD is more robust to τ for *Vine*, whereas its performance slightly decreases for *Instagram* as τ increases. In practice, λ_3 should be set to a small value, and a proper value for parameter τ should be experimentally identified. In general, UCD is robust to most of the model parameters, and consequently can be tuned for various real-world applications.

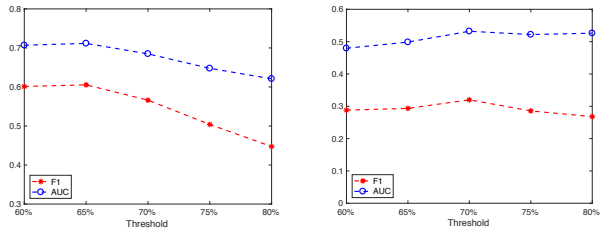


(a) Effect of λ_1 (b) Effect of λ_2



(c) Effect of λ_3 (d) Effect of K

Figure 4: Parameter study w.r.t the AUROC and F1 scores.



(a) *Instagram* (b) *Vine*

Figure 5: Effects of τ on AUROC and F1 scores.

4.6 Case Study

In this subsection, we present two Instagram sessions, one detected as bullying and one detected as non-bullying by UCD. We visualize each with the hierarchical attention information to validate UCD's capability of selecting informative comments and words in a session. The results can be seen in Fig. 6. Every line in each sub-figure is a comment. Shades of blue denote comment weights and shades of red denote word weights. Because both sessions have many comments, only a portion of the content is shown here. Fig. 6(a) shows that UCD can select the words that are more strongly associated with bullying, such as *f*ckin, b*tch, disgusted* and *hell*. In Fig. 6(b), we observe that UCD can also deal with complex cross-comment context. For example, although the session might appear to be a bullying session when looking only at the second comment from the bottom, UCD assigns the session to the non-bullying cluster because it also considers the context of that comment.

5 DISCUSSION

In this section, we elaborate on the reasons behind the performance of UCD, its research impact, and practical considerations. UCD benefits from the following design mechanisms:

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

(a) *Predicted as bullying session.*

how do u get gif i ca nt save them to my phone .
larry zayn being sexy and niall and liam doing something stupid in back .
larry having their little moment there .
are of you actually fans of one direction .
just because ur elounor shipper does n't mean you have to be bitch lol shut up .
i feel like they have changed so many peoples life 's including mine .

(b) *Predicted as non-bullying session.*

Figure 6: Case study using the *Instagram* dataset.

- *Multi-modal features.* UCD actively leverages multi-modal data including text, user information, social network information, and social content. UCD also benefits from deep learning mechanisms specifically designed for each modality, e.g., HAN models the sequence of comments and the hierarchy of a session. Previous work [8] reported the benefits of using multi-modal data to contribute complementary application domain insights and enable better learning performance.
- *Complementary temporal analysis.* In addition to multi-modal representation learning, UCD simultaneously estimates the energy level associated with bullying instances and predicts the time-interval between comments to refine the session representations. Temporal modeling adds nuance to the representation learning network that otherwise would not consider comment evolution [6, 41].
- *Joint optimization.* A key property that differentiates UCD from other approaches is that it jointly optimizes the parameters for representation learning, temporal modeling, and bullying-energy estimation. This approach prevents the drawbacks of decoupled training.

As one of the first attempts to detect cyberbullying in an unsupervised manner, UCD explores the use of deep learning algorithms and shows they can achieve relatively high performance levels. The development of UCD has relevant research and practical impact. UCD addresses key limitations of supervised models: (1) cyberbullying labeled data could be either unavailable or insufficient for training a good supervised classifier, (2) data labeling is often time-consuming and labor-intensive, and (3) the guidelines used for assigning cyberbullying labels in a current session cannot always be generalized to future sessions due to the dynamic nature of language and social networks. We hope this work will motivate further research efforts in unsupervised cyberbullying detection. Regarding the practical use of UCD, it could be integrated into third-party anti-bullying apps, such as Bark[7] and BullyBlocker [38], or as a component of automated mediation tools [30].

6 CONCLUSIONS AND FUTURE WORK

Existing efforts towards detecting cyberbullying have focused primarily on supervised methods that require large amounts of time and labor to annotate datasets. To address this limitation, we propose an unsupervised cyberbullying detection framework, called

[7] https://www.bark.us

UCD, which consists of two major components: a *representation learning* network that encodes multi-modal session representations and a *multi-task learning* network that simultaneously estimates bullying likelihood and models the temporal dynamics of arriving comments. The joint parameter optimization in UCD yields better performance. Our experimental results on two real-world datasets corroborate the effectiveness of UCD.

The presented findings elucidate multiple paths for future work, including a more detailed analysis of the temporal characteristics of cyberbullying behaviors in social media and the study of user-session networks (where users and sessions are the nodes) to directly explore the connection between users and associated posts.

ACKNOWLEDGEMENTS

This work was in part supported by the National Science Foundation (NSF) Grants 1719722 and 1614576.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
[2] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *CSUR* 41, 3 (2009), 15.
[3] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Websci*. ACM, 13–22.
[4] 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.
[5] Lu Cheng, Ruocheng Guo, and Huan Liu. 2019. Robust Cyberbullying Detection with Causal Interpretation. In *WWW' Companion*.
[6] Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, and Huan Liu. 2019. Hierarchical Attention Networks for Cyberbullying Detection on the Instagram Social Network. In *SDM*.
[7] Lu Cheng, Jundong Li, Yasin Silva, Deborah Hall, and Huan Liu. 2019. PI-Bully: Personalized Cyberbullying Detection with Peer Influence. In *IJCAI*. AAAI.
[8] Lu Cheng, Jundong Li, Yasin N Silva, Deborah L Hall, and Huan Liu. 2019. XBully: Cyberbullying Detection within a Multi-Modal Context. In *WSDM*. 339–347.
[9] Harsh Dani, Jundong Li, and Huan Liu. 2017. Sentiment informed cyberbullying detection in social media. In *ECML PKDD*. Springer, 52–67.
[10] Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino. 2016. Unsupervised cyber bullying detection in social networks. In *ICPR*. IEEE, 432–437.
[11] Thomas G Dietterich. 2002. Machine learning for sequential data: A review. In *SSPR*. Springer, 15–30.
[12] 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), 18.
[13] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *ICWSM*.
[14] Dorothy L Espelage, Melissa K Holt, and Rachael R Henkel. 2003. Examination of peer–group contextual effects on aggression during early adolescence. *Child development* 74, 1 (2003), 205–220.
[15] Ruth Festl and Thorsten Quandt. 2013. Social relations and cyberbullying: The influence of individual and structural attributes on victimization and perpetration via the internet. *Human communication research* 39, 1 (2013), 101–126.
[16] Aditya Grover, Aaron Zweig, and Stefano Ermon. 2018. Graphite: Iterative generative modeling of graphs. *arXiv preprint arXiv:1803.10459* (2018).
[17] Aabhaas Gupta, Wenxi Yang, Divya Sivakumar, Yasin N Silva, Deborah L Hall, and Maria Camila Nardini Barioni. 2020. Temporal Properties of Cyberbullying on Instagram. (2020).
[18] 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 *Socinfo*. Springer, 49–66.
[19] 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*. IEEE, 186–192.
[20] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. In *SAM*. ACM, 3–6.
[21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[22] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
[23] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).
[24] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. 2006. A tutorial on energy-based learning. *Predicting structured data* 1, 0 (2006).
[25] Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057* (2015).
[26] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the presence and intensity of hostility on Instagram using linguistic and social features. In *ICWSM*.
[27] C Moessner. 2014. Cyberbullying, Trends and Tudes. *NCPC. org. Accessed* (2014).
[28] Parma Nand, Rivindu Perera, and Abhijeet Kasture. 2016. "How Bullying is this Message?": A Psychometric Thermometer for Bullying.. In *COLING*. 695–706.
[29] online. [n. d.]. Ditch The Label (2013) The Annual Cyberbullying Survey. Availabe from https://www.ditchthelabel.org/wp–content/uploads/2016/07/cyberbullying2013.pdf.
[30] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. *arXiv preprint arXiv:1909.04251* (2019).
[31] 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*. ACM, 617–622.
[32] Rahat Ibn Rafiq, Homa Hosseinmardi, Sabrina Arredondo Mattson, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network. *SNAM* 6, 1 (2016), 88.
[33] Elaheh Raisi and Bert Huang. 2017. Co-trained ensemble models for weakly supervised cyberbullying detection. In *NIPS LLD Workshop*.
[34] Elaheh Raisi and Bert Huang. 2017. Cyberbullying detection with weakly supervised machine learning. In *ASONAM*. ACM, 409–416.
[35] Walisa Romsaiyud, Kodchakorn na Nakornphanom, Pimpaka Prasertsilp, Piyaporn Nurarak, and Pirom Konglerd. 2017. Automated cyberbullying detection using clustering appearance patterns. In *KST*. IEEE, 242–247.
[36] Guillaume Salha, Romain Hennequin, Viet Anh Tran, and Michalis Vazirgiannis. 2019. A degeneracy framework for scalable graph autoencoders. *arXiv preprint arXiv:1902.08813* (2019).
[37] Christina Salmivalli, Arja Huttunen, and Kirsti MJ Lagerspetz. 1997. Peer networks and bullying in schools. *Scandinavian journal of psychology* 38, 4 (1997), 305–312.
[38] Yasin N Silva, Deborah L Hall, and Christopher Rich. 2018. BullyBlocker: toward an interdisciplinary approach to identify cyberbullying. *SNAM* 8, 1 (2018), 18.
[39] 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.
[40] Chunfeng Song, Feng Liu, Yongzhen Huang, Liang Wang, and Tieniu Tan. 2013. Auto-encoder based data clustering. In *CIARP*. Springer, 117–124.
[41] Devin Soni and Vivek Singh. 2018. Time Reveals All Wounds: Modeling Temporal Characteristics of Cyberbullying. In *ICWSM*.
[42] Daniel Svozil, Vladimir Kvasnicka, and Jiri Pospichal. 1997. Introduction to multilayer feed-forward neural networks. *CHEMOMETR INTELL LAB* 39, 1 (1997), 43–62.
[43] Miranda Witvliet, Tjeert Olthof, Jan B Hoeksma, Frits A Goossens, Marieke SI Smits, and Hans M Koot. 2010. Peer group affiliation of children: The role of perceived popularity, likeability, and behavioral similarity in bullying. *Social Development* 19, 2 (2010), 285–303.
[44] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *ICML*. 478–487.
[45] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *NAACL HLT*. ACL, 656–666.
[46] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *ICML*. JMLR. org, 3861–3870.
[47] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL HLT*. 1480–1489.
[48] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. 2016. Deep structured energy based models for anomaly detection. *arXiv preprint arXiv:1605.07717* (2016).
[49] Caleb Ziems, Ymir Vigfusson, and Fred Morstatter. 2020. Aggressive, Repetitive, Intentional, Visible, and Imbalanced: Refining Representations for Cyberbullying Classification. In *ICWSM*.
[50] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*.