HENIN: Learning Heterogeneous Neural Interaction Networks for Explainable Cyberbullying Detection on Social Media

Hsin-Yu Chen
Institute of Data Science
National Cheng Kung University
Tainan, Taiwan
d0107330@gmail.com

Cheng-Te Li
Institute of Data Science
National Cheng Kung University
Tainan, Taiwan
chengte@ncku.edu.tw

Abstract

In the computational detection of cyberbullying, existing work largely focused on building generic classifiers that rely exclusively on text analysis of social media sessions. Despite their empirical success, we argue that a critical missing piece is the model explainability, i.e., why a particular piece of media session is detected as cyberbullying. In this paper, therefore, we propose a novel deep model, HEterogeneous Neural Interaction Networks (HENIN), for explainable cyberbullying detection. HENIN contains the following components: a comment encoder, a post-comment co-attention sub-network, and session-session and post-post interaction extractors. Extensive experiments conducted on real datasets exhibit not only the promising performance of HENIN, but also highlight evidential comments so that one can understand why a media session is identified as cyberbullying.

1 Introduction

In recent years, cyberbullying has become one of the most pressing online risks among youth and raised serious concerns in society. Cyberbullying is commonly defined as the electronic transmission of insulting or embarrassing comments, photos or videos, as illustrated in Figure 1. Harmful bullying behavior can include posting rumors, threats, pejorative labels, and sexual remarks. Research from the American Psychological Association and the White House has revealed more than 40% of young people in the US indicate that they have been bullied on social media platforms (Dinakar et al., 2012). Such a growing prevalence of cyberbullying on social media has detrimental societal effects, such as victims may experience lower self-esteem, increased suicidal ideation, and a variety of negative emotional responses (Hinduja and Patchin, 2014). Therefore, it has become critically important to be able to detect and prevent cyberbullying

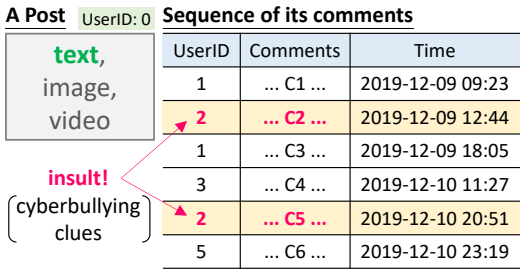


Figure 1: An illustration of a media session containing an image/video/posted text and a sequence of comments. A cyberbullying session is typically composed of multiple insulting comments.

on social media. Research in computer science aimed at identifying, predicting, and ultimately preventing cyberbullying through better understanding the nature and key characteristics of online cyberbullying.

In the literature, existing efforts toward automatically detecting cyberbullying have primarily focused on textual analysis of user comments, including keywords (Dadvar et al., 2012; Nahar et al., 2013; Nand et al., 2016) and sentiments analysis (Dani et al., 2017). These studies 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 largely overlooked temporal information of cyberbullying behaviors. They also ignore user interactions in social networks. Furthermore, the majority of these methods focus on detecting cyberbullying sessions effectively but cannot explain "why" a media session was detected as cyberbullying. Given a sequence of comments with user attributes, we think sequential learning can allow us to better exploit and model the evolution and correlations among individual comments. Besides, graph-based learning can enable us to represent and learn how users interact with each other in a session.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2543–2552,
November 16–20, 2020.

This work aims to detect cyberbullying by jointly exploring explainable information from user comments on social media. To this end, we build an explainable cyberbullying detection framework, **HE**terogeneous **N**eural **I**nteraction **N**etworks (**HENIN**), through a coherent process. HENIN consists of three main components that learn various interactions among heterogeneous information displayed in social media sessions. A comment encoder is created to learn the representations of user comments through a hierarchical self-attention neural network so that the semantic and syntactic cues on cyberbullying can be captured. We create a post-comment co-attention mechanism to learn the interactions between a posted text and its comments. Moreover, two graph convolutional networks are leveraged to learn the latent representations depicting how sessions interact with one another in terms of users, and how posts are correlated with each other in terms of words.

Specifically, we address several challenges in this work: (a) how to perform explainable cyberbullying detection that can boost detection performance, (b) how to highlight explainable comments without the ground truth, (c) how to model the correlation between posted text and user comments, and (d) how to model the interactions between sessions in terms of users, and the interactions between textual posts in terms of words. Our solutions to these challenges result in a novel framework HENIN.

Our contributions are summarized as follows. (1) We study a novel problem of explainable cyberbullying detection on social media. (2) We provide a novel model, HENIN [1], which jointly exploits posted text, user comments, and the interactions between sessions and between posts to learn the latent representations for cyberbullying detection. (3) Experiments conducted on Instagram and Vine datasets exhibit the promising performance of HENIN, and the evidential comments and words highlighted by HENIN, for detecting cyberbullying media sessions with explanations.

2 Related Work

Relevant studies can be categories into social contexts-based and user comment-based approaches. **Social contexts-based approaches** utilize three categories of features, user-based, post-based, and network-based. (a) Post-based features rely on text analysis to identify cyberbullying evidences (e.g., profane words) on social media (Dadvar et al., 2012; Nahar et al., 2013; Nand et al., 2016). Xu et al. (2012) point out Latent Semantic Analysis(LSA) and Latent Dirichlet Allocation (LDA) can be used to learn latent representations of posts. In addition, *SICD* (Dani et al., 2017) further models post sentiments for cyberbullying detection. (b) User-based features are extracted from user profiles to measure their characteristics. Gender-specific features, user's past posts, account registration time, and frequently-used words are useful user-based features (Dadvar and De Jong, 2012; Dadvar et al., 2013). (c) Existing studies (Cheng et al., 2019b; Tu et al., 2018; Wang et al., 2017) also prove that network-based features are effective in detecting cyberbullying. These features are learned by constructing propagation networks or interaction networks that depict how posts are spread and how users interact with each other. **User comment-based approaches** utilize the sequence of user comments to detect cyberbullying of the source post. CONcISE (Yao et al., 2019) is a sequential hypothesis testing method conducted on the comment sequence to select the significant comment features. Raisi and Huang (2018) detect harassment-based cyberbullying by identifying expert-provided key phrases from user comments.

3 Problem Statement

Let $S = \{s_1, s_2, ..., s_M\}$ denote a corpus of M social media sessions. Each media session contains the posted text and its subsequent comments. Let P be a posted text, consisting of N words $\{w_i\}_{i=1}^{N}$. Let $C = \{c_1, c_2, ..., c_T\}$ be a set of T comments related to the post P, where each comment $c_j = \{w_1^j, w_2^j, ...w_{Q_j}^j\}$ contains Q_j words. Let $G_{ss} = (V_S, E_S)$ be a session-session weighted graph, in which we consider each media session as a node $s \in V_S$ and the similarity between sessions as an edge weight $e_{(s_i,s_j)} \in E_S$. Let $G_{pp} = (V_P, E_P)$ be a post-post weighted graph, in which we consider each posted text as a node $p \in V_P$ and the similarity between posts as an edge weight $e_{(p_i,p_j)} \in E_P$. We treat the cyberbullying detection problem as the binary classification problem, i.e., each media session is associated with a binary label $y = \{0, 1\}$ with 1 representing a bullying session, and 0 representing a non-bullying session. At the same time, we aim to learn a rank

[1] The Code of HENIN model is available at: https://github.com/HsinYu7330/HENIN

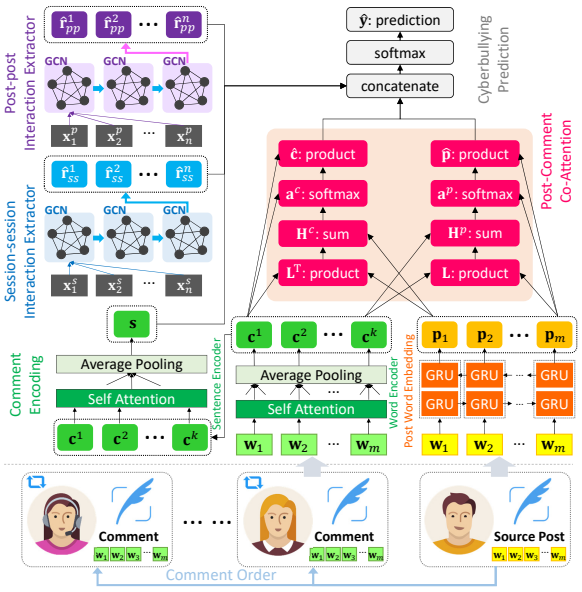


Figure 2: The proposed HENIN model, which contains four components: a joint word-level and sentence-level comment encoder, a post-comment co-attention mechanism, session-session and post-post interaction extractors, and the final cyberbullying prediction.

list RC from all comments in $\{c_j\}_{j=1}^{T}$, according to the degree of explainability, where RC_k denotes the k_{th} most explainable comment. The explainability of comments denotes the impact degree of detecting the media session is cyberbullying or not. Formally, we can represent the problem as *Explainable Cyberbullying Detection*.

Problem: Given a posted text P, a set of related comments C, the session graph G_{ss} and the post graph G_{pp}, the goal is to learn a cyberbullying detection function $f : f(P, C, G_{ss}, G_{pp}) \rightarrow (\hat{y}, RC)$, such that it maximizes the prediction accuracy with explainable comments ranked highest in RC.

4 The proposed HENIN Model

In this section, we present the details of the proposed HENIN, which jointly learns the hierarchical self-attention and graph convolutional neural networks for cyberbullying detection. It consists of four major components (Figure 2): (1) a comment encoder (including word-level and sentence-level), (2) a post-comment co-attention mechanism, (3) session-session and post-post interaction extractors, and (4) a cyberbullying prediction component.

The comment encoder component depicts the modeling from the comment linguistic features to latent representation features through hierarchical word-level and sentence-level self-attention networks. The explainability degree of comments is learned through the attention weights within sentence-level self-attention learning. The post-comment co-attention mechanism is performed in the level of word embeddings. The mutual interactions between the posted text and comments can be learned through the post-comment co-attention. On the other hand, the session-session interaction extractor and the post-post interaction extractor aim at modeling how users interact across media sessions, and how words are correlated across posts, through two graph convolutional neural networks. Finally, the cyberbullying prediction is made by concatenating the representations of the aforementioned three components.

4.1 Comment Encoding

A set of comments related to the given media session contains linguistic cues at the word and sentence levels. Textual usages in comments provide different degrees of importance for explainability of why the session is detected as cyberbullying. For example, in a cyberbullying media session extracted from the Instagram dataset (see Section 5.1), the comment *"how the fuck are you even a fucking fan you cunt if you just talk shit about harry fuck you kaitlyn!"*, the words "fuck" and "shit" contribute more signals to reflect apparent and evidential emotion sense, compared to other ones. Meanwhile, this comment strongly expresses malicious remarks to someone, and therefore it is not only more explainable but also useful to determine whether it is a cyberbullying session.

Several studies have shown that improved document representations with highlighting important words and sentences for classification can be learned by hierarchical attention neural networks (Yang et al., 2016; Cheng et al., 2019a). Inspired by (Yang et al., 2016), we adopt a hierarchical neural network to model word-level and sentence-level representations through self-attention mechanisms. Specifically, we first learn the comment embedding vector by utilizing the word encoder with self-attention. Then we learn the comment representations through the sentence encoder with self-attention.

Word Encoder. Given a comment c_j with m words, we first embed the words to a latent space via the pre-trained word2vec model (Mikolov et al., 2013). Then we capture words' contextual relations among comments by calculating scaled dot-product

attention (Vaswani et al., 2017). Specifically first, let word embeddings as input vectors $\mathbf{x}_i$. The query vector sequence $\mathbf{q}_i$, the key vector sequence $\mathbf{k}_i$, and the value vector sequence $\mathbf{v}_i$ can be obtained by linear transformation, i.e., $\mathbf{q}_i = \mathbf{w}_q\mathbf{x}_i$, $\mathbf{k}_i = \mathbf{w}_k\mathbf{x}_i$, and $\mathbf{v}_i = \mathbf{w}_v\mathbf{x}_i$, where $\mathbf{w}_q, \mathbf{w}_k, \mathbf{w}_v$ are the learnable parameters through the networks. Next we compute the dot products of the query with all keys, divide each by $\sqrt{d_k}$ (d_k is the dimension of keys), and apply a softmax function to obtain the attention weights on the values: $\mathbf{a}_i = \text{softmax}(\frac{\mathbf{q}_i\mathbf{k}_i^\top}{\sqrt{d_k}})$, where $\mathbf{a}_i$ is an attention weight vector that measures the importance of each word in the comment. Finally, each word's hidden representation can be obtained by computing the dot products of attention weights $\mathbf{a}_i$ and the value vector sequence $\mathbf{v}_i$. We take the average of the learned representations to generate the comment vector $\mathbf{c}^j$, given by: $\mathbf{c}^j = \frac{\sum_{i=1}^{m}\mathbf{a}_i\mathbf{v}_i}{m}$.

Sentence Encoder. Similar to the word encoder, we utilize the scaled dot-product attention to encode each media session. The aim is to capture the context information at the sentence level, and to generate the media session representation of post P_i, denoted by $\mathbf{s}^i$, from the learned comment embedding vectors $\{\mathbf{c}^1, \mathbf{c}^2, ..., \mathbf{c}^k\}$. Every post's sentence embedding $\mathbf{s}$ will be used as features for cyberbullying prediction.

4.2 Post-Comment Co-attention Mechanism

To model the interaction between posted text and comments, we propose a post-comment co-attention mechanism that learns the semantic word-level correlation between posted text and comments. That said, we intend to simultaneously learn and derive the attention weights of words on posted text and comments. Specifically first, similar to comment encoding, word embeddings of a posted text are obtained by a pre-trained word2vec model. We adopt recurrent neural networks with bidirectional gated recurrent units (GRU) to model word sequences from both directions of words. The bidirectional GRU contains the forward GRU $\overrightarrow{f}$ that reads posted text p^i from word w_1^i to w_m^i and the backward GRU $\overleftarrow{f}$ that reads posted text p^i from word w_m^i to w_1^i, given by: $\overrightarrow{\mathbf{h}_t^i} = \overrightarrow{GRU}(\mathbf{w}_t^i)(t \in \{1, ..., m\})$ and $\overleftarrow{\mathbf{h}_t^i} = \overleftarrow{GRU}(\mathbf{w}_t^i)(t \in \{m, ..., 1\})$. We obtain the embedding of word p_t^i in a posted text by concatenating its forward and backward hidden states $\overrightarrow{\mathbf{h}_t^i}$ and $\overleftarrow{\mathbf{h}_t^i}$, i.e., $\mathbf{p}_t^i = [\overrightarrow{\mathbf{h}_t^i}, \overleftarrow{\mathbf{h}_t^i}]$. Then we can construct the feature matrix of words of posted text $\mathbf{P} = [\mathbf{p}^1, ..., \mathbf{p}^N]$. Similarly the feature matrix of comments $\mathbf{C} = [\mathbf{c}^1, ..., \mathbf{c}^T]$ can be derived.

The proposed co-attention mechanism attends to the posted text words and the comment simultaneously. By extending the co-attention formulation (Lu et al., 2016; Cui et al., 2019), we first compute the affinity matrix $\mathbf{L} \in \mathbb{R}^{T\times N}$: $\mathbf{L} = \tanh(\mathbf{C}^\top\mathbf{W}_l\mathbf{P})$, where $\mathbf{W}_l$ is a matrix of learnable weights. The affinity matrix $\mathbf{L}$ is used to transform the comment attention space to the posted text attention space, and vice versa for $\mathbf{L}^\top$. As a result, we can consider the affinity matrix as a feature matrix, and learn to predict the posted text and comment attention maps $\mathbf{H}^p$ and $\mathbf{H}^c$, as follows: $\mathbf{H}^p = \tanh(\mathbf{W}_p\mathbf{P} + (\mathbf{W}_c\mathbf{C})\mathbf{L})$, and $\mathbf{H}^c = \tanh(\mathbf{W}_c\mathbf{C} + (\mathbf{W}_p\mathbf{P})\mathbf{L}^\top)$, where $\mathbf{W}_p, \mathbf{W}_c$ are the matrices of learnable parameters. The attention weights of posted text and comments, $\mathbf{a}^p$ and $\mathbf{a}^c$, can be obtained by: $\mathbf{a}^p = \text{softmax}(\mathbf{w}_{hp}^\top\mathbf{H}^p)$, $\mathbf{a}^c = \text{softmax}(\mathbf{w}_{hc}^\top\mathbf{H}^c)$, where $\mathbf{w}_{hp}^\top$ and $\mathbf{w}_{hc}^\top$ are vectors of learnable weight parameters. Based on the above attention weights, the posted text and comment attention vectors are obtained by calculating the weighted sum of the posted text features and comment features via: $\hat{\mathbf{p}} = \sum_{i=1}^{N}\mathbf{a}_i^p\mathbf{p}^i$ and $\hat{\mathbf{c}} = \sum_{i=1}^{T}\mathbf{a}_i^c\mathbf{c}^i$,, where $\hat{\mathbf{p}}$ and $\hat{\mathbf{c}}$ are the learned features vectors for posted text and comments, respectively, through the co-attention mechanism.

4.3 Interaction Extractors

To learn and represent the potential interactions between two sessions as well as two text posts, we utilize multilayer neural networks that operate on graph data based on the layers of graph convolutional networks (GCN) (Kipf and Welling, 2016). GCN is able to induce embedding vectors of nodes based on features of their neighborhoods. We create two multi-layer GCNs to learn the embeddings of the given session s_i and its posted text P_i from the session-session graph G_{ss} and the post-post graph G_{pp}, respectively.

Session-session Interaction Extractor. Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n) \in \mathbb{R}^{n\times p}$ be the vectors of user participation in all sessions, where n is the number of all sessions and p is the number of users. Each vector $\mathbf{x}_i$ is a multi-hot encoding that depicts how session s_i is participated by all users. Let matrix $\hat{\mathbf{R}}_{ss}$ be the representations of all sessions learned from the session-session graph $G_{ss} = (\mathbf{X}, \mathbf{A})$, where $\mathbf{A} \in \mathbb{R}^{n\times n}$ encodes the pairwise relationships (such as cosine similarity, which

is used by default) between sessions. We exploit GCN to learn $\hat{\mathbf{R}}_{ss}$. GCN contains one input layer, several propagation layers, and the final output layer (Kipf and Welling, 2016). At deeper layers, the nodes indirectly receive more information from farther nodes in the graph. Given the input feature matrix $\mathbf{X}^{(0)} = \mathbf{X}$ and the graph structure matrix $\mathbf{A}$, GCN performs the layer-wise propagation in hidden layers via $\mathbf{X}^{(k+1)} = \rho(\hat{\mathbf{A}}\mathbf{X}^{(k)}\mathbf{W}^{(k)})$, where $k = 0, 1, ..., K-1$ and $\mathbf{W}^{(k)}$ is the matrix of learnable parameters in the k-th layer. ρ is a non-linear activation function, such as ReLU, and $\mathbf{X}^{(k+1)}$ denotes the activation output in the k-th layer. $\hat{\mathbf{A}}$ is the normalized symmetric adjacency matrix, $\hat{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D} = diag(d_1, d_2, ..., d_n)$ is a diagonal matrix with $d_i = \sum_{j=1}^{n} \mathbf{A}_{ij}$. Finally, the graph representations $\hat{\mathbf{R}}_{ss} = [\hat{\mathbf{r}}_{ss}]$ can be obtained from the output layer that uses *softmax* as the activation function.

Post-post Interaction Extractor. Similar to *session-session interaction extractor*, we depict each posted text in the graph G_{pp} as a real-valued vector $\mathbf{x}_i$ by using the word embedding vector of post P_i as the initial feature. By performing GCNs as aforementioned, we can derive the graph representations of all posts, denoted by $\hat{\mathbf{R}}_{pp} = [\hat{\mathbf{r}}_{pp}]$.

4.4 Cyberbullying Prediction

By concatenating the sentence embedding vector $\mathbf{s}$, the post-comment co-attention feature vectors $\hat{\mathbf{p}}$ and $\hat{\mathbf{c}}$, the session interaction representation $\hat{\mathbf{r}}_{ss}$, and the post interaction representation $\hat{\mathbf{r}}_{pp}$, we generate the prediction via a fully-connected layer, given by: $\hat{\mathbf{y}} = \sigma([\hat{\mathbf{p}}, \hat{\mathbf{c}}, \mathbf{s}, \hat{\mathbf{r}}_{ss}, \hat{\mathbf{r}}_{pp}]\mathbf{W}_f + \mathbf{b}_f)$, where $\hat{\mathbf{y}}$ is the predicted probability vector indicating the predicted probability of label 1 (i.e., cyberbullying). $\mathbf{W}_f$ and $\mathbf{b}_f$ are the learnable parameters and biases. σ is the sigmoid function. $y \in \{0, 1\}$ denotes the ground-truth label of media sessions. The goal is to minimize the cross-entropy loss function: $\mathcal{L}(\Theta) = -y\log(\hat{\mathbf{y}}) - (1-y)\log(1-\hat{\mathbf{y}})$, where Θ denotes all parameters of the network. The parameters in the network are learned through the *Adam* optimizer (Kingma and Ba, 2014), which is an adaptive learning rate method that uses estimations of first and second moments of gradient to adapt the learning rate for each weight of the neural network. We choose *Adam* since it is generally regarded as being fairly robust and effective to the choice of the hyperparameters, and it is widely used for training neural networks.

Table 1: Statistics of Instagram and Vine datasets.

Datasets	Instagram	Vine
# Sessions	2,211	882
# Bullying	676	283
# Non-Bullying	1,535	599
# Comments	159,277	70,385
# Users	72,176	25,699

5 Experiments

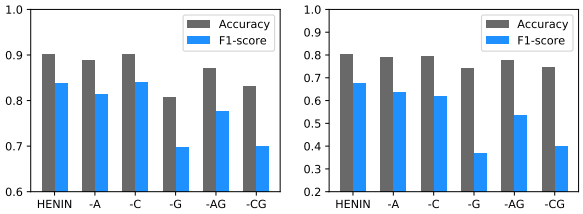
We aim to answer the following evaluation questions. **EQ1:** Can HENIN improve the cyberbullying media session *classification performance*? **EQ2:** How effective is *each component* of HENIN? **EQ3:** Is HENIN able to perform accurate *early* detection of cyberbullying sessions? **EQ4:** Can HENIN highlight comments that can *explain why* a media session is detected as cyberbullying?

5.1 Datasets and Settings

We use two social media datasets whose statistics is shown in Table 1. One is Instagram dataset (Hosseinmardi et al., 2015, 2016), which contains image description and user comments. The other is Vine (Rafiq et al., 2015, 2016), which is a mobile application website that allows users to record and edit a few seconds looping videos. The texts of both datasets are in English.

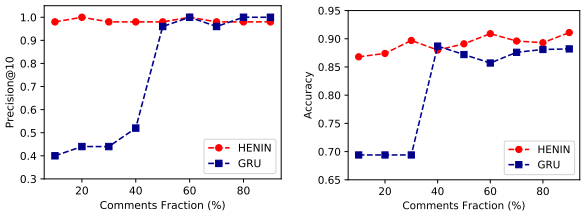
We compare our HENIN model with several methods, including classification models such as Logistic Regression (**LR**) (Hosseinmardi et al., 2015, 2016) and Random Forest (**RF**) (Rafiq et al., 2015, 2016). We collect posted text and all related comments of the session as a document to embed the session to a latent space via pre-trained doc2vec model (Le and Mikolov, 2014). Then we leverage the session representations as input features to train LR and RF classifiers. In addition, we also compare HENIN with three end-to-end deep learning models, including **RNN**, **GRU**, and GRU with attention **GRU+A**. We also compare HENIN with a recent advance **CONcISE** (Yao et al., 2019), which has a sequential hypothesis testing-based mechanism to produce timely and accurate detection of cyberbullying. For a fair comparison with CONcISE, we follow their settings by using their suggested key terms: "ugly", "shut", "suck", "gay", "beautiful", "sick", 'bitch", 'work", "hate", and "fuck."

We provide the hyperparameter settings to enable the reproducibility. (1) The maximum number of words per comment `MAX_COM_WORD_LEN=10`



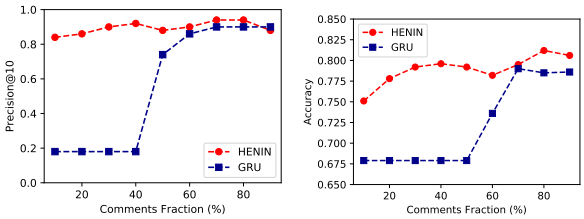
(a) Instagram (b) Vine

Figure 3: Results of ablation analysis for HENIN.



(a) Precision@10 (b) Accuracy

Figure 4: Effect of comments' fraction on Instagram.



(a) Precision@10 (b) Accuracy

Figure 5: Effect of comments' fraction on Vine.

and 6 on Instagram and Vine, respectively, according to the median of all comments' length. (2) The maximum length of user comments MAX_COM_LEN=75 and 80 on Instagram and Vine, respectively. (3) The dimension of word embeddings d=300. (4) The number of GCN layers is 3. (5) The matrix **A** for GCN is constructed by pairwise cosine similarity between posts and sessions.

5.2 Cyberbullying Detection Performance

To answer **EQ1**, we first compare our HENIN with baseline methods. To evaluate the performance of cyberbullying detection methods, we use the following metrics, which are commonly used to evaluate classifiers: Accuracy (Acc), Precision (Pre), Recall (Rec), and F1-Score (F1). To have the experiments be more robust and reliable, we randomly choose 80% of media sessions for training and the remaining 20% for testing. We repeat the process 5 times, and report the average values. The results are shown in Table 2. We can find that the proposed HENIN consistently outperforms the competing methods across two datasets on Accuracy, Recall, and F1, i.e., except for the metric of Precision. Although RF and RNN lead to higher scores in Precision in Instagram and Vince datasets, respectively, their performance in other metrics is not stable. It is also worthwhile to notice that models considering attention mechanisms, i.e., HENIN and GRU+A, tend to produce better performance. This implies the importance of modeling contextual correlation and contribution at either word or sentence level on the detection of cyberbullying.

5.3 Ablation Analysis for HENIN

To answer **EQ2**, we further investigate the effect of each component in the proposed HENIN model. We aim at evaluating the following reduced variants of HENIN. (1) **-A**: HENIN without the Post-Comment co-attention component, (2) **-G**: HENIN without the GCN components, (3) **-C**: HENIN without the Comment Encoder, (4) **-AG**: HENIN without the Post-Comment co-attention and GCN components, and (5) **-CG**: HENIN without the Comment Encoder and GCN components.

The results are shown in Figure 3. The ablation analysis of HENIN brings two insights. First, all of the three components (i.e., comment encoder, session-session and post-post interactions, and posted text-comment co-attention) contribute apparently to the performance improvement. Second, When the model without considering the representations learned from session and post interactions, the performance reduces 14% and 9.6% in terms of F1-Score and Accuracy metrics on Instagram, and 30.7% and 6% on Vine. In other words, "-G" models hurt the performance most. The results suggest that modeling interactions between sessions and between posts through GCNs in HENIN is important.

5.4 Early Detection of Cyberbullying

To answer **EQ3**, we examine whether HENIN can accurately detect cyberbullying sessions at early stages. In other words, we aim to understand how a model performs given only a partial proportion of observed comments. Here we choose GRU as the baseline for comparison. Specifically, for each media session, we sort all comments by response time, then choose various fractions of comments into the training and testing sets. We utilize *Precision@k* and *Accuracy* as the evaluation metrics,

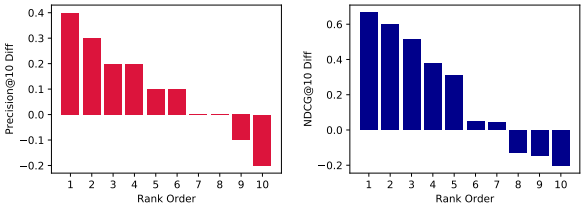
Table 2: The main performance comparison in four metrics for cyberbullying detection on two datasets. Note that the best model and the second model are highlighted by **bold** and underline, respectively.

Datasets	Metrics	CONcISE	RNN	GRU	GRU+A	LR	RF	HENIN
Instagram	Acc	0.627	0.782	0.815	0.884	0.840	0.805	**0.902**
	Pre	0.388	0.817	0.846	0.835	0.792	**0.901**	0.889
	Rec	0.381	0.376	0.496	0.781	0.652	0.405	**0.829**
	F1	0.384	0.507	0.569	0.805	0.715	0.559	**0.838**
Vine	Acc	0.603	0.706	0.747	0.797	0.788	0.786	**0.804**
	Pre	0.363	**0.830**	0.773	0.757	0.748	0.751	0.821
	Rec	0.376	0.190	0.309	0.559	0.512	0.498	**0.643**
	F1	0.369	0.245	0.418	0.636	0.608	0.597	**0.676**

where $k = 10$. The results are shown in Figure 4 and Figure 5. From the figures, we can see that, our proposed HENIN can achieve much better performance when the observed comments are quite a few (i.e., the fraction of comments is low than 40%). In contrast, GRU model needs at least 50% comments on both datasets to obtain the same good performance as HENIN. In short, we prove that HENIN is able to produce quite accurate early detection of cyberbullying sessions.

5.5 Explainability and Case Study

Explainability. To answer **EQ4**, we evaluate the performance of the explainability of our *HENIN* model from the perspective of comments. We choose *GRU+A* as the baselines for comment explainability since it can learn attention weights for comments as a kind of explainability. Specifically, we want to see if the top-ranked explainable comments determined by our HENIN are more likely to be related to the major contexts in cyberbullying media sessions. We randomly choose 10 media sessions, which contains at least 20 but not more than 50 comments, to evaluate the explainability ranking list of the comment *RC*. Then we denote the ground-truth ranking list by rating the explainability score from $\{0, 1, 2, 3, 4\}$ for each comment, where 0 means "not explainable at all", 1 means "not explainable", 2 means "neutral", 3 means "somewhat explainable", and 4 means "highly explainable (highly malicious)." We invite three domain experts to perform the ground-truth ratings for every comment. The average rating scores are used to generate the ranking list. Therefore, for each media session, we have two lists of top-k comments, $L^{(1)} = \{L_1^{(1)}, L_2^{(1)}, ..., L_k^{(1)}\}$ by HENIN, and $L^{(2)} = \{L_1^{(2)}, L_2^{(2)}, ..., L_k^{(2)}\}$ by GRU+A. The top-k comments are ranked and selected using the comment attention weights from high to low. To estimate the rank-aware explainability of comments, we utilize *Normalized Discounted Cumulative Gain* (NDCG) (Järvelin and Kekäläinen, 2002) and *Precision@k* as the evaluation metrics. We empirically set $k = 10$.



(a) Precision@10 (b) NDCG@10

Figure 6: The discrepancy histograms of mean Precision@10 and mean NDCG@10 (in the y-axis) for the results between HENIN and GRU+A in Vine dataset.

The results are shown in Figure 6, where media sessions are sorted by the discrepancy in the metrics between two methods, i.e., NDCG@k(HENIN)−NDCG@k(GRU+A), in a descending order. From the figures, we can have two observations. First, among 10 Vine media sessions, HENIN obtains higher precision scores than GRU+A for 6 cases. The overall mean precision scores over 10 cases for HENIN and GRU+A are 0.51 and 0.41, respectively. Second, similar results can be found on NDCG scores. HENIN is superior to GRU+A on 7 cases, and two cases have equal NDCG scores. The overall mean NDCG scores over 10 cases for HENIN and GRU+A are 0.57 and 0.36, respectively. These results demonstrate that the attention weights of HENIN are able to highlight more evidential comments than GRU+A, and its explainability can be verified.

Case Study. We further demonstrate the explainable comments that HENIN correctly ranks high

Posted text

Lets go in the hallway right now bitch **30** comments

Top-7 comments ranked by HENIN

Rank	AttW	Comments
1	1.421	*What a bitch tell him to hmu and ill kill his bitch ass for hitting a woman*
2	0.219	*if a bitch hit a nigga wit a object damn right we gon retaliate*
3	0.127	*When ugly girl try play flight with cute boiltshanabishh Axi Esete*
4	0.077	*That weak ass punch lmao Michael Featherston*
5	0.074	*She had no business hitting him wit anything period*
6	0.072	*Laurie us to the kid with the mole on his face*
7	0.070	*Court-dawg Jimecia Bandy Donishia Phillips*

Figure 7: The top-7 comments highlighted by HENIN.

but GRU+A misses. These cases are presented in Figure 7. We can find that: (1) our HENIN can rank more evidential comments higher than non-explainable comments. For example, the top-1 comment *"What a bitch tell him to hmu and ill kill his bitch ass for hitting a woman"* contains explicit vulgar and malicious texts that can explain why this media session detected as cyberbullying. (2) We can give higher attention weights to explainable comments than those neutral and unrelated comments. For example, the unrelated comment *"Court-dawg Jimecia Bandy Donishia Phillips"* has an attention weight 0.070, which is lower than an explainable comment *"if a bitch hit a nigga wit a object damn right we gon retaliate"* with attention weight 0.219. Therefore, the latter comment is selected to be a more important evidence for cyberbullying prediction. In short, HENIN is able to not only accurately detect cyberbullying sessions, but also highlight evidential comments as explanations.

5.6 HENIN Hyperparameter Analysis

Since we have shown that the graph-based interactions between sessions and between posts have a great impact on the detection (Section 5.3), we further aim to investigate how different hyperparameters of GCNs affect the performance. Here we study two hyperparameters. One is the number of GCN layers. The other is the choice of similarity measures in constructing the matrix $\mathbf{A}$ for GCN. The results on stacking the different number of GCN layers are shown in Table 3. We can see that stacking more GCN layers leads to performance improvement by around 1.1% in terms of F1 on Instagram and 2.2% on Vine.

The weight matrix $\mathbf{A}$ for GCN is obtained by calculating the similarity for all pairs of nodes in the graph. We compare three commonly similarity measures, Cosine similarity: $\cos(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$, Jaccard similarity: $\text{jac}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i \mathbf{x}_j}{\sum \mathbf{x}_i \sum \mathbf{x}_j - \sum \mathbf{x}_i \mathbf{x}_j}$, and Euclidean similarity: $\text{euc} = 1 - \bar{\text{euc}}(\mathbf{x}_i, \mathbf{x}_j) = 1 - \bar{N}(\sqrt{\sum (\mathbf{x}_i - \mathbf{x}_j)^2})$ ($\bar{\text{euc}}$ and $\bar{N}$ denote normalization to [0,1]). The results are shown in Table 4. We can see that on the Instagram dataset, using Euclidean similarity can improve the performance by 4.9% and 2.8% in terms of F1 and Accuracy, respectively. On the Vine dataset, using Jaccard similarity outperform than the other two measures by improving 1.2% and 1.7% in terms of F1 and Accuracy, respectively. The results suggest that in different datasets, we need to choose the proper similarity measure to construct the weight matrix as the performance can be affected.

Table 3: Effect of the number of GCN layers.

Dataset	Instagram		Vine	
	Acc	F1	Acc	F1
#layers=1	0.896	0.827	0.803	0.672
#layers=2	0.896	0.829	0.797	0.654
#layers=3	**0.902**	**0.838**	**0.804**	**0.676**

Table 4: Effect of similarity measures in constructing matrix $\mathbf{A}$ depicting the graph for GCN.

Dataset	Instagram		Vine	
$\mathbf{A}_{ij}$	Acc	F1	Acc	F1
$\cos(\mathbf{x}_i, \mathbf{x}_j)$	0.894	0.823	0.806	0.668
$\text{jac}(\mathbf{x}_i, \mathbf{x}_j)$	0.893	0.824	**0.811**	**0.673**
$\text{euc}(\mathbf{x}_i, \mathbf{x}_j)$	**0.922**	**0.872**	0.794	0.661

6 Conclusion

Cyberbullying detection on social media attracts growing attention in recent years. It is also crucial to understand why a media session is detected as cyberbullying. Thus we study the novel problem of explainable cyberbullying detection that aims at improving detection performance and highlighting explainable comments. We propose a novel deep learning-based model, HEterogeneous Neural Interaction Networks (HENIN), to learn various feature representations from comment encodings, post-comment co-attention, and graph-based interactions between sessions and posts. Experimental results exhibit both promising performance and evidential explanation of HENIN. We also find that the learning of graph-based session-session and post-post interactions contributes most to the performance. Such results can encourage future studies to develop advanced graph neural networks in better representing the interactions between heterogeneous information. In addition, it is worthwhile to further model information propagation and tem-

poral correlation of comments in the future.

Acknowledgments

This work is supported by Ministry of Science and Technology (MOST) of Taiwan under grants 109-2636-E-006-017 (MOST Young Scholar Fellowship) and 109-2221-E-006-173, and also by Academia Sinica under grant AS-TP-107-M05.

References

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

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

Limeng Cui, Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: A system for explainable fake news detection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2961–2964. ACM.

Maral Dadvar and Franciska De Jong. 2012. Cyberbullying detection: a step toward a safer internet yard. In *Proceedings of the 21st International Conference on World Wide Web*, pages 121–126.

Maral Dadvar, FMG de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer.

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*, pages 52–67. Springer.

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):18.

Sameer Hinduja and Justin W Patchin. 2014. *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Corwin Press.

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*, pages 49–66. Springer.

Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Prediction of cyberbullying incidents in a media-based social network. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 186–192. IEEE.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297.

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*, pages 3111–3119.

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

Parma Nand, Rivindu Perera, and Abhijeet Kasture. 2016. "how bullying is this message?": A psychometric thermometer for bullying. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 695–706.

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 *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 617–622. ACM.

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. *Social Network Analysis and Mining*, 6(1):88.

Elaheh Raisi and Bert Huang. 2018. Weakly supervised cyberbullying detection using co-trained ensembles of embedding models. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 479–486. IEEE.

Ke Tu, Peng Cui, Xiao Wang, Fei Wang, and Wenwu Zhu. 2018. Structural deep embedding for hyper-networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Suhang Wang, Jiliang Tang, Charu Aggarwal, Yi Chang, and Huan Liu. 2017. Signed network embedding in social media. In *Proceedings of the 2017 SIAM international conference on data mining*, pages 327–335. SIAM.

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*, pages 656–666. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Mengfan Yao, Charalampos Chelmis, and Daphney? Stavroula Zois. 2019. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *The World Wide Web Conference*, pages 3427–3433.