XBully: Cyberbullying Detection within a Multi-Modal Context

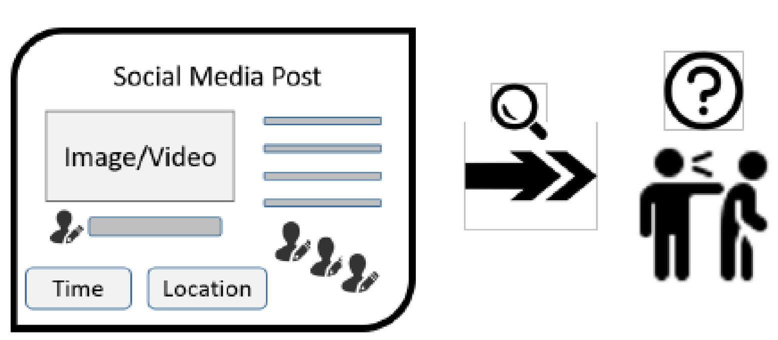
Lu Cheng, Jundong Li, Yasin Silva, Deborah Hall, Huan Liu
Arizona State University

Motivation

- The American Psychological Association and the White House have identified cyberbullying as a serious national health concern, with more than 40% of teenagers in the US reporting that they have been bullied on social media platforms.
- Within the computer science community, existing efforts toward detecting cyberbullying have primarily focused on text analysis.
- These models inevitably ignore critical information included in the various social media modalities such as image, video, user profile, time and location.

Challenges

- Heterogeneous information of different modalities may not be compatible with each other.
- Social media data is often not independent and identically distributed but is intrinsically correlated, either directly or indirectly, limiting the applicability of conventional text analysis approaches.
- Different modalities are often associated with rather diverse feature types (e.g., nominal, ordinal, interval, ratio, etc.)



XBully Framework

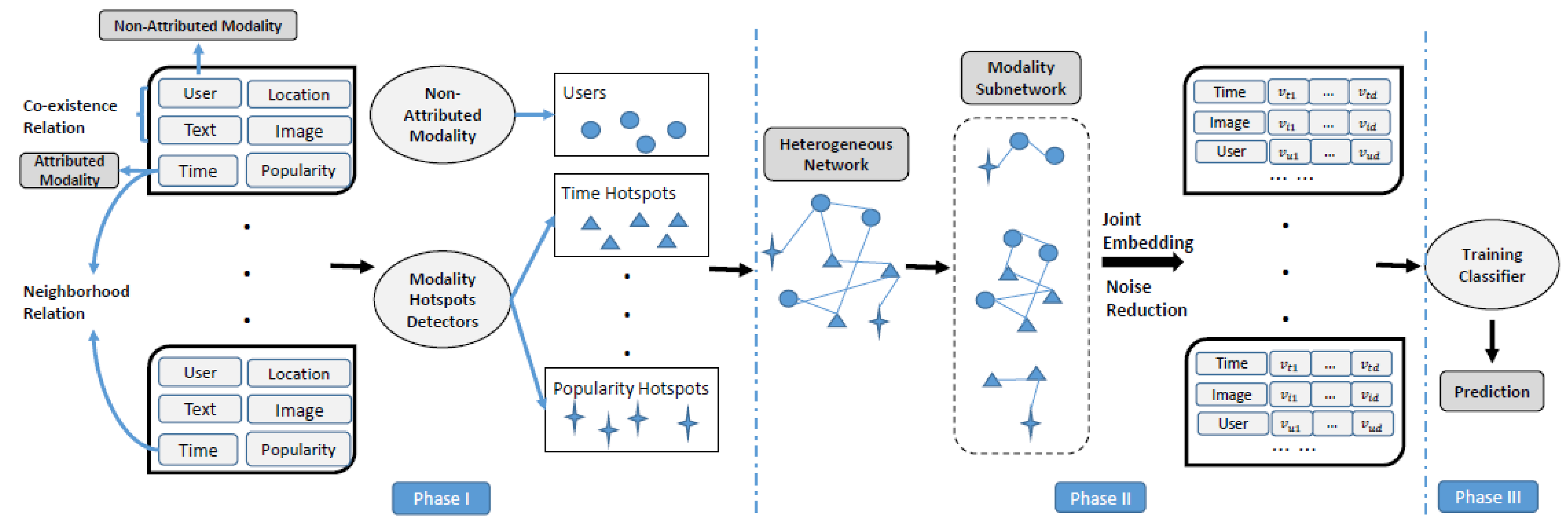


Figure 2: The proposed XBully framework. Given a corpus of social media sessions, we first attempt to discover hotspots for each mode (Phase I); and then based on the detected hotspots and instances of nominals, we leverage the ***co-existence*** and ***neighborhood*** relations to construct a heterogeneous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

Key Components:

1. Detection of Mode Hotspots
2. Heterogeneous Network Embedding
3. Embedding Refinement

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

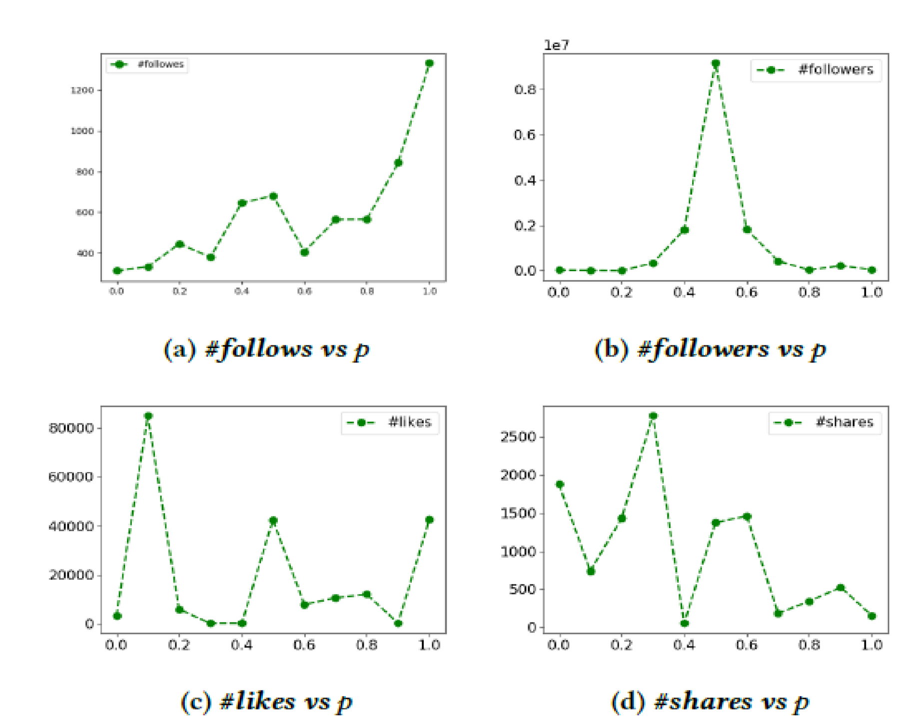
$$Z_{AB} = O_{AA} + O_{AB} + O_{BA} + O_{BB} + \lambda \sum_{i=1}^{|V_{AB}|} R_i,$$

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

Experiments

Table 1: Performance comparison of various methods on the Instagram dataset.

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



(a) #*follows vs p* (b) #*followers vs p*
(c) #*likes vs p* (d) #*shares vs p*

Figure 3: Qualitative Analysis

Contributions

- Problem Formulation
- Algorithms
  - Hotspot detection
  - Heterogeneity
  - Joint embedding
- Experimental Evaluation

Acknowledgement

This work was supported by National Science Foundation Award # 1719722.

BullyBlocker
http://bullyblocker.asu.edu

Data Mining and Machine Learning Lab