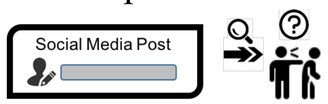
Personalized Learning for Cyberbullying Detection

Lu Cheng, Yasin Silva, Deborah Hall, Huan Liu
Arizona State University
{lcheng35, ysilva, d.hall, huan.liu}@asu.edu

Background

- Given a group of social media posts, cyberbullying detection aims to train a binary classification model to predict the labels of online social media posts.



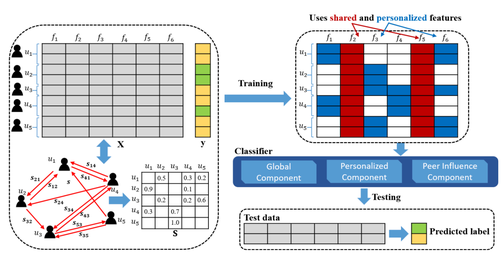
- Limitations of existing approaches:
  - ❖ A generic binary classification model for all users which fails to capture the unique aspects of cyberbullying behavior.
  - ❖ Existing works ignore the patterns of similarity in bullying behavior and victimization within adolescent peer groups.
- Target: Model idiosyncratic characteristics of users and quantify the peer influence from similar users to facilitate cyberbullying detection.
- Social media posts are:
  - ❖ Massive, short and noisy
  - ❖ Sparse and high-dimensional
  - ❖ Without formal forms

Challenges

- Three challenges of personalized learning for cyberbullying detection:
  - ❖ User-generated content is often sparse, high-dimensional with redundant features that may jeopardize the learning performance due to the curse of dimensionality.
  - ❖ In spite of users' idiosyncrasies, they also share common attributes and behaviors.
  - ❖ Bullying victims and perpetrators are influenced by peers, and the influence from users can be diverse.

The Proposed Framework

- Key ideas:
  - ❖ Besides the global model, the proposed framework also learns a personalized model for each individual.
  - ❖ The framework includes a peer influence component to extrapolate information from like-minded users.
- Framework



- Model

$$\min_{\mathbf{w},\mathbf{P}_i,\mathbf{Q}_i} \sum_{i=1}^{U}\sum_{j=1}^{N_i} f(\mathbf{x}_j^i, y_j^i, \mathbf{w}+\mathbf{P}_i+\mathbf{Q}_i) + \lambda_1(\|\mathbf{w}\|_1 + \sum_{i=1}^{U}\|\mathbf{P}_i\|_1) + \lambda_2 \sum_{i=1}^{U}\|\mathbf{Q}_i - \sum_{j=1}^{U} s_{ji}\mathbf{P}_j\|_2^2,$$

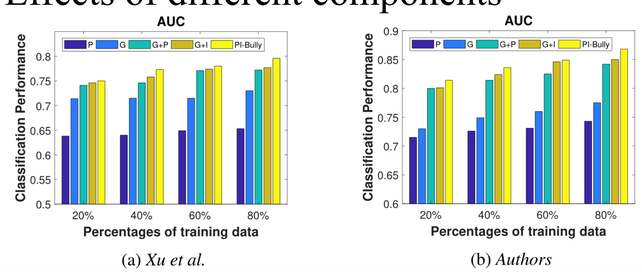
Experimental Results

- Two real-world datasets crawled from Twitter: *Xu. et.al* and *Authors* with the sizes of 3095 and 20000 respectively.
- Baseline methods
  - ❖ Text classification models: k-NN, linear SVM, Random Forest, Nearest Centroid
  - ❖ Existing cyberbullying detection models: *Bully* (2012) and *SICD* (2017)
- Comparison Results

Table 1: Performance comparison of various methods w.r.t. different percentages of training data.

Percentages		20%		40%		60%		80%	
Datasets		*Xu et al.*	*Authors*	*Xu et al.*	*Authors*	*Xu et al.*	*Authors*	*Xu et al.*	*Authors*
Text Classification Models	*k*NN	0.606	0.520	0.638	0.535	0.645	0.543	0.667	0.561
	Linear SVM	0.516	0.539	0.519	0.522	0.531	0.581	0.579	0.602
	RF	0.660	0.737	0.675	0.730	0.659	0.724	0.678	0.752
	NC	0.639	0.595	0.637	0.613	0.646	0.612	0.653	0.617
Cyberbullying Models	*Bully*	0.705	0.680	0.737	0.701	0.752	0.711	0.769	0.725
	SICD	0.718	0.751	0.723	0.768	0.763	0.781	0.777	0.801
	PI-Bully	**0.750**	**0.814**	**0.758**	**0.836**	**0.780**	**0.849**	**0.796**	**0.868**

- Effects of different components



(a) *Xu et al.* (b) *Authors*

Acknowledgements: Sponsored by the NSF grant #1719722