Personalized Learning for Cyberbullying Detection

Lu Cheng, Yasin Silva, Deborah Hall, and Huan Liu

Arizona State University, Tempe, Arizona, U.S.,
{lcheng35,ysilva,d.hall,huanliu}@asu.edu

1 Introduction

Cyberbullying has become one of the most pressing online risks for adolescents and has raised serious concerns in society. Traditional efforts are primarily devoted to building a single generic classification model for all users to differentiate bullying behaviors from the normal content [6, 3, 1, 2, 4]. Despite its empirical success, these models treat users equally and inevitably ignore the idiosyncrasies of users. Recent studies from psychology and sociology suggest that the occurrence of cyberbullying has a strong connection with the personality of victims and bullies embedded in the user-generated content, and the peer influence from like-minded users. In this paper, we propose a personalized cyberbullying detection framework PI-Bully with peer influence in a collaborative environment to tailor the prediction for each individual. In particular, the personalized classifier of each individual consists of three components: a global model that captures the commonality shared by all users, a personalized model that expresses the idiosyncratic personality of each specific user, and a third component that encodes the peer influence received from like-minded users. Most of the existing methods adopt a two-stage approach: they first apply feature engineering to capture the cyberbullying patterns and then employ machine learning classifiers to detect cyberbullying behaviors.

However, building a personalized cyberbullying detection framework that is customized to each individual remains a challenging task, in large part because: (1) Social media data is often sparse, noisy and high-dimensional (2) It is important to capture the commonality shared by all users as well as idiosyncratic aspects of the personality of each individual for automatic cyberbullying detection; (3) In reality, a potential victim of cyberbullying is often influenced by peers and the influences from different users could be quite diverse. Hence, it is imperative to develop a way to encode the diversity of peer influence for cyberbullying detection.

To summarize, we study a novel problem of personalized cyberbullying detection with peer influence in a collaborative environment, which is able to jointly model users' common features, unique personalities and peer influence to identify cyberbullying cases.

2 Approach

Building the Personalized Model Previous efforts in cyberbullying detection have been primarily devoted to the development of a single classification model to capture the commonalities of users. Nevertheless, it is important to design additional personalized components to capture the individuality of each person. We assume each user u_i has a personalized model $\mathbf{M}_i \in \mathbb{R}^D$ in addition to the global model $\mathbf{w} \in \mathbb{R}^D$. Moreover, we impose an ℓ_1-norm sparse regularization term on each personalized model $\mathbf{M}_i$ to alleviate the curse of dimensionality. Hence, we obtain the following optimization framework:

$$\min_{\mathbf{w},\mathbf{M}_i} \sum_{i=1}^{U} \sum_{j=1}^{N_i} f(\mathbf{x}_j^i, y_j^i, \mathbf{w} + \mathbf{M}_i) + \lambda_1(\|\mathbf{w}\|_1 + \sum_{i=1}^{U} \|\mathbf{M}_i\|_1). \tag{1}$$

where $f(\cdot)$ is a loss function.

Characterizing Peer Influence The model parameter learning of the above personalized model can be problematic due to the limited amount of training data for each user. To address this problem, we decompose the personalized model $\mathbf{M}_i$ of each user into a personalized component, $\mathbf{P}_i \in \mathbb{R}^D$, which encodes a user's inherent traits, and a collaborative/peer influence component, $\mathbf{Q}_i \in \mathbb{R}^D$, which quantifies the influence received from like-minded users. Then the objective function in Eq. (1) can be reformulated as follows:

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

where λ_2 is a parameter that balances the contribution of collaborative/peer influence for personalized cyberbullying detection. Specifically, s_{ji} denotes how user u_i is influenced by user u_j and the influence from different peers could vary significantly.

3 Results

The dataset is crawled via the Twitter streaming API [5] and 20,000 tweets were manually labeled by two well-trained human annotators with backgrounds in psychology. We compare PI-Bully with several text classifcation machine learning models as well as two cyberbullying detection mode, *Bully* [6] and *SICD* (the state-of-the-art model) [2]. Training datasets are generated by extracting increasing fractions (5% to 30%) of the overall dataset and using the rest part of the datasets as the test datasets. The experimental results show that our proposed PI-Bully framework outperforms all other compared methods, including commonly used text classification models and existing cyberbullying detection models. We also study the impact of the three components with the following variants:

- The personalized model (P): a variant of PI-Bully that eliminates the global model $\mathbf{w}$ and the peer influence component and only keeps the personalized component $\mathbf{P}_i$ for each user.
- The global model (G): a variant of PI-Bully that only keeps the global model $\mathbf{w}$ and ignores the other components.
- Global+Personalized (G+P): a variant of PI-Bullying without the peer influence component.
- Global+Influence (G+I): a variant of the proposed model that eliminates the personalized component $\mathbf{P}_i$ for each user.

Results show that (1) The personalized model P is inferior to the global model ; (2) Both the G+I model and the G+P model outperform the global model G and (3) The proposed PI-Bully framework achieves the best performance and the dominance becomes more obvious as the size of the training dataset increases. This shows the benefits of combining the three proposed components.

4 Future Work

Future work should be directed towards integrating cross-modal features, such as social network features, sentiments, and images from social media platforms, to better understand and predict cyberbullying behaviors. It should also continue building on empirical findings in psychology that identify new correlates and determinants of cyberbullying behaviors, given the tremendous potential for unique interdisciplinary studies between computer science and psychology to address this major social issue.

References

1. Dadvar, M., De Jong, F.: Cyberbullying detection: a step toward a safer internet yard. In: Proceedings of the 21st International Conference on World Wide Web. pp. 121–126. ACM (2012)
2. Dani, H., Li, J., Liu, H.: Sentiment informed cyberbullying detection in social media. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 52–67. Springer (2017)
3. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. The Social Mobile Web **11**(02) (2011)
4. Huang, Q., Singh, V.K., Atrey, P.K.: Cyber bullying detection using social and textual analysis. In: Proceedings of the 3rd International Workshop on Socially-Aware Multimedia. pp. 3–6. ACM (2014)
5. Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In: Proceedings of the 7th International AAAI Conference on Web and Social Media. ICWSM (2013)
6. Xu, J.M., Jun, K.S., Zhu, X., Bellmore, A.: 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. pp. 656–666. Association for Computational Linguistics (2012)