BullyBlocker: An Interdisciplinary Approach for the Identification and Prevention of Adolescent Cyberbullying

Brittany Wheeler, Lu Cheng, Deborah Hall, Yasin Silva
Arizona State University

Introduction

Cyberbullying is defined as any behavior performed through electronic or digital media to threaten or cause harm to others [13]. 59% of U.S. teens have experienced at least one instance of abusive online behavior [14]. As a result, cyberbullying is being increasingly identified as a major health concern [6]. The BullyBlocker Project bridges computer, data, and psychological science to design theory-driven methods for identifying and preventing cyberbullying among teens on social media.

Personalized Cyberbullying Detection

Background

- Previous cyberbullying detection methods employed in computer science have mainly focused on the development of global classification models that capture the commonality shared by all users [5,16].
- Research in psychology has identified individual difference characteristics (e.g., personality traits) that are correlated with cyberbullying—e.g., psychopathy, Machiavellianism, and narcissism have been identified as predictors of cyberbullying perpetration [7].

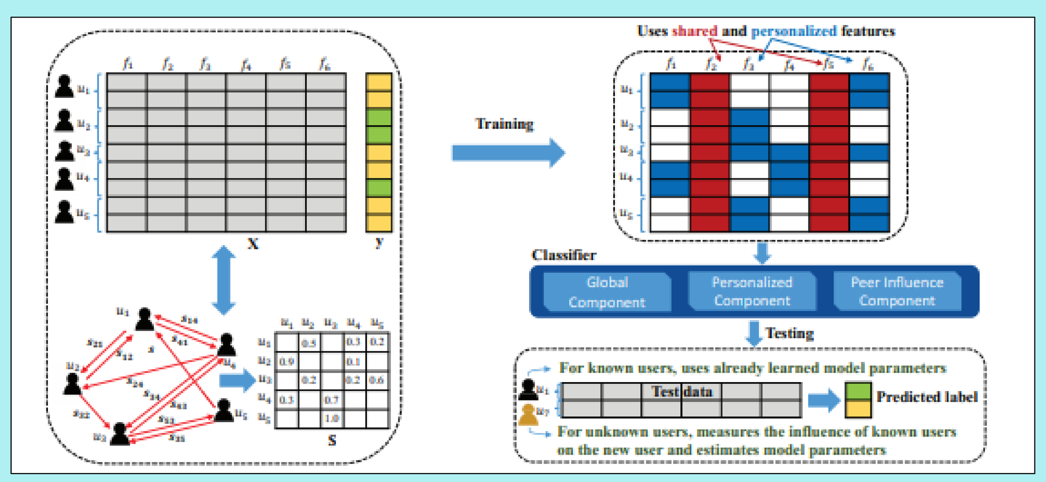

Figure 1.1: The Proposed PI-Bully Framework. The data matrix, X, is used to compute the similarity matrix, S, which quantifies how users are similar to each other. In the training phase, shared and user-specific features train a classifier. Finally, a testing phase with unlabeled data is performed to identify cyberbullying.

Proposed Framework

- The PI-Bully model contains three components as described in Figure 1.1 [4].
- In addition to the global model, a personalized model is included to capture the unique characteristics of the user.
  - A personalized model can suffer from overfitting due to a limited amount of training information.
- A collaborative/peer influence component is included to derive information about cyberbullying experiences from similar users.
  - This component is personalized for each user using a weighted average of the personalized component from other users.

Model Evaluation

- Real world data was crawled via the Twitter streaming API using 25 cyberbullying related keywords (e.g., nerd, gay, freak, and whore) to evaluate the PI-Bully model.
  - 20,000 tweets were extracted to be labeled by human annotators with psychology and computer science backgrounds.
  - After data cleaning and annotator conflict resolution, a total of 19,994 tweets were included. 19.23% of the tweets displayed bullying interactions.

Metrics	Precision	Recall	F1	AUC
*k*NN	0.663	0.364	0.470	0.652
SVM	0.699	0.469	0.562	0.701
RF	0.708	0.478	0.571	0.707
LR	0.680	0.485	0.566	0.705
Bully	0.653	0.508	0.571	0.709
SICD	**0.803**	0.263	0.396	0.791
PI-Bully	0.425	**0.887**	**0.574**	**0.844**

Table 1.1 Performance Comparison.
Values are the average performance of ten runs.

- PI-Bully was compared against several common text classification models (*kNN, Random Forest, Linear SVM, and Logistic Regression)* and two text-based cyberbullying detection models (*Bully* and *SICD*) [5,16].
- PI-Bully achieved the best Recall, F1, and AUC scores compared to the baseline models, indicating improved performance in cyberbullying detection (see Table 1.1).

Impact of Model Components

- By itself, the personal component performed worse than the global model (Figure 1.2).
- The global model with peer influence (G+I) and the global model with a personalized component (G+P) outperformed the global model.
- The PI-Bully framework achieved the best performance indicating the benefits of considering all three components when detecting cyberbullying.

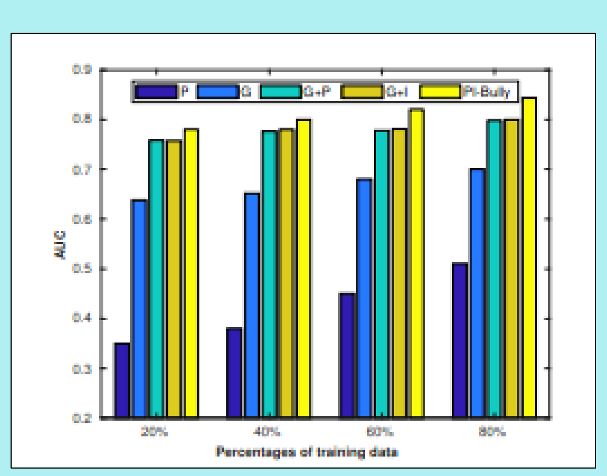

Figure 1.2: Performance Evaluation of Different Model Components

Temporal Characteristics

Background

- Cyberbullying, by nature, is often repetitive, but little is known about how the timing, number, and frequency of cyberbullying messages differ from non-cyberbullying messages [13].

Datasets

- We used two versions of Hosseinmardi et al.'s Instagram dataset, in which cyberbullying labels had been applied at the session level [10].
  - **Human-labeled data**: We extracted a subset of 100 Instagram sessions from the original dataset (50 cyberbullying sessions and 50 normal sessions). Members of our research team manually coded each comment in addition to each session as cyberbullying or normal. Our session labels were similar to the ones by Hosseinmardi et al. for a total of 48 cyberbullying sessions and 52 normal sessions.
  - **ML-labeled data**: We used the eXtreme Gradient Boosting Model to classify all 2,218 Instagram sessions from the original dataset as cyberbullying or normal. Word count vectors, word-level TF-IDF vectors, and psychological features were used in this model. The accuracy of this model was 90%.

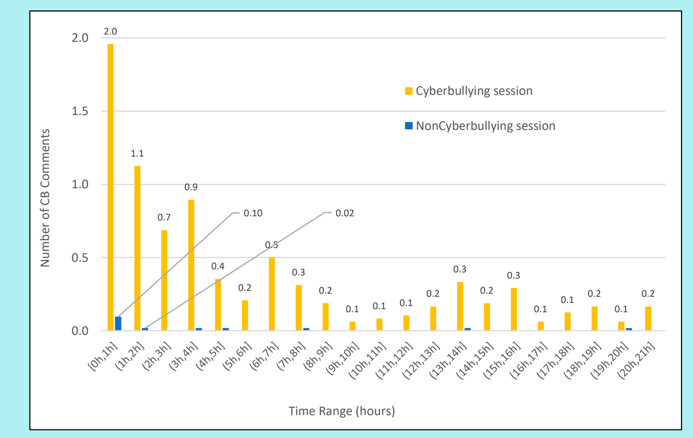

Figure 2.1: Number of CB Comments (Per Session) Over Time with Human-labeled Data

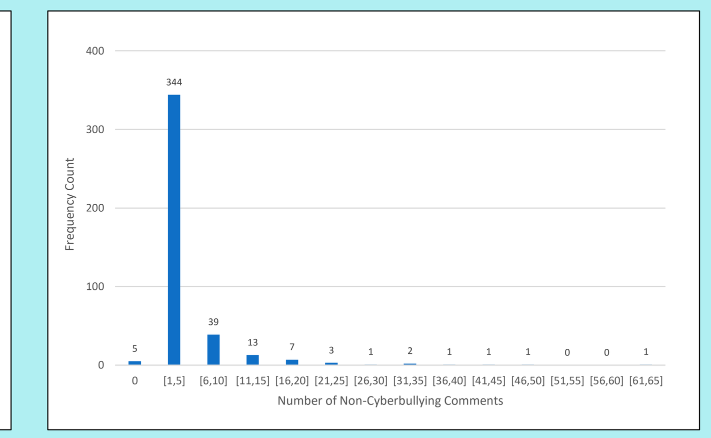

Figure 2.2: Number of Non-CB comments Between Consecutive CB Comments with Human-labeled Data

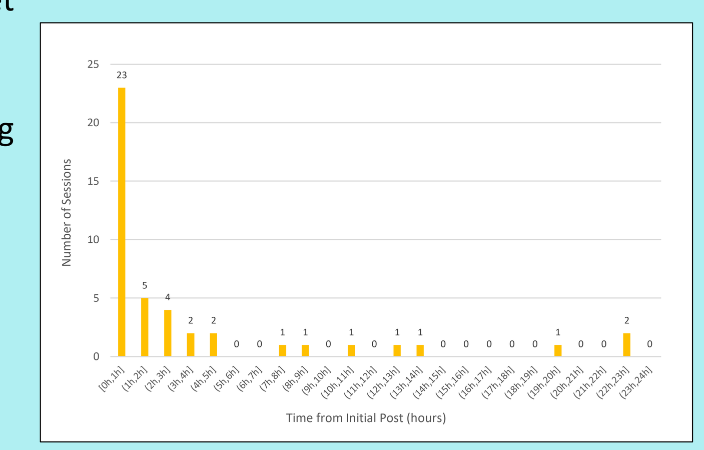

Figure 2.3: Temporal Distribution of First CB Comment with Human-labeled Data

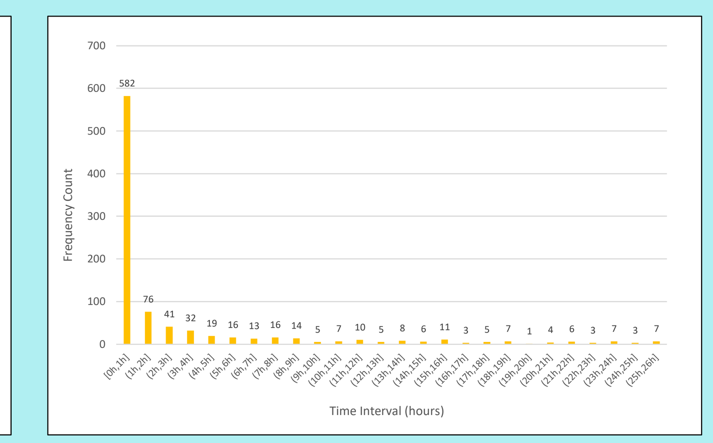

Figure 2.4: Distribution of the Time Interval Between Consecutive CB Comments with Human-labeled Data

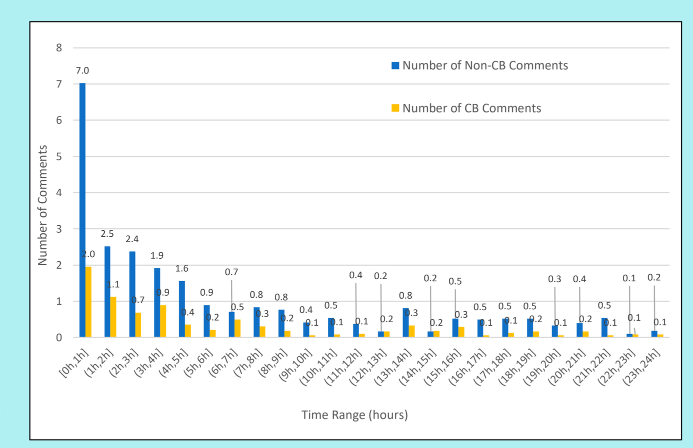

Figure 2.5: Number of CB and Non-CB Comments per CB Session with Human-labeled Data

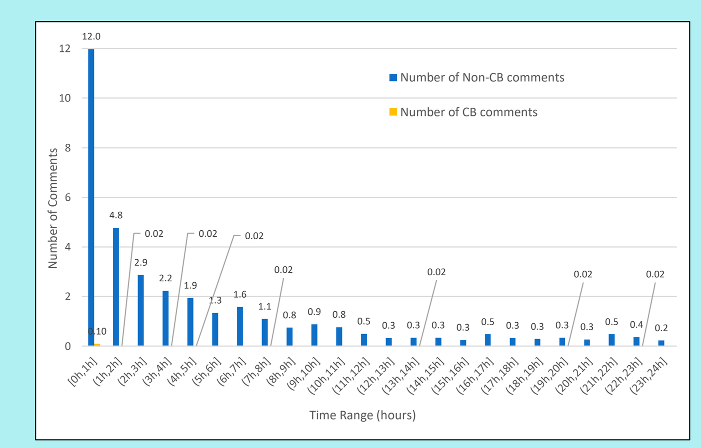

Figure 2.6: Number of CB and Non-CB Comments per Non-CB Session with Human-labeled Data

Analysis

- In non-cyberbullying and cyberbullying sessions, most bullying activity occurred within the first hour after the initial post (Figure 2.1) [8].
  - For non-cyberbullying sessions, on average, 0.1 cyberbullying comment occurred within the first hour.
  - For cyberbullying sessions, on average, 2 cyberbullying comments occurred within the first hour.
  - The first cyberbullying comment occurred within the first hour for about 50% of the sessions (Figure 2.3).
- Non-cyberbullying comments occurred frequently between cyberbullying comments (Figure 2.2).
  - 344 pairs of consecutive cyberbullying comments had 1 to 5 non-cyberbullying comments between them.
  - When the number of non-cyberbullying comments increased, the number of cyberbullying pairs decreased indicating that non-cyberbullying comments might play a protective role.
- The separation between the first and second cyberbullying comment occurred within the first hour for 64% of the sessions (Figure 2.4).
- The number of non-cyberbullying comments in cyberbullying and non-cyberbullying sessions always exceeded the number of cyberbullying comments, but both types of comments decreased over time (Figure 2.5 and 2.6).

Burst Analysis

- Using a burst detection analysis developed by Kleinberg [12], with s = 1.4 and gamma = 0.1, a strong cluster of bursts occurred during the first 5 hours after the initial post, indicating intense cyberbullying comments during this time (Figure 2.7).
  - Less intense bursts also occurred within 6 and 9 hours and between 14 and 15 hours after the initial post.
- Over 30 days, strong bursts of activity occurred within the first 4 days, with the most activity occurring on the first day.
  - Several isolated bursts occurred over the 30 days but were weaker in magnitude.

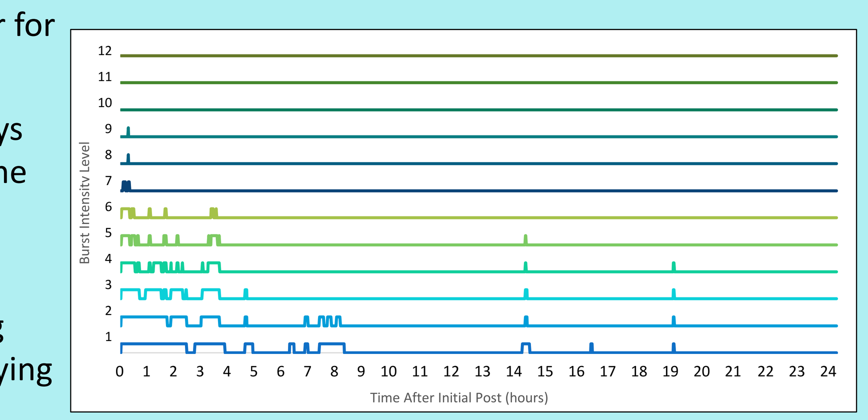

Figure 2.7: Bursts in the Human-labeled Dataset – Short-term (24h)

Hierarchical Attention Network

Background

- Previous approaches to cyberbullying detection have largely overlooked the context and structural properties of social media sessions [3,4,8].
- We developed the Hierarchal Attention Network of Cyberbullying Detection (HANCD) framework to model hierarchical structures, the differential importance of words and comments, temporal characteristics, and social information (e.g., #Likes) to improve cyberbullying detection.
- The Instagram dataset collected by Hosseinmardi et al. (containing 2,218 sessions, with 678 labeled as bullying) was used to evaluate the effectiveness of the HANCD framework [10].
  - 80% of the data was used to train the model.

Proposed Framework

- This framework depicted in Figure 3.1 utilizes [3]:
  - Word sequence encoder
  - Word-level attention layer
  - Comment sequence encoder
  - Comment-level attention layer
  - Contextual information
  - Time interval prediction
- The HANCD framework can classify social media sessions based on text, time, and the social media information provided.

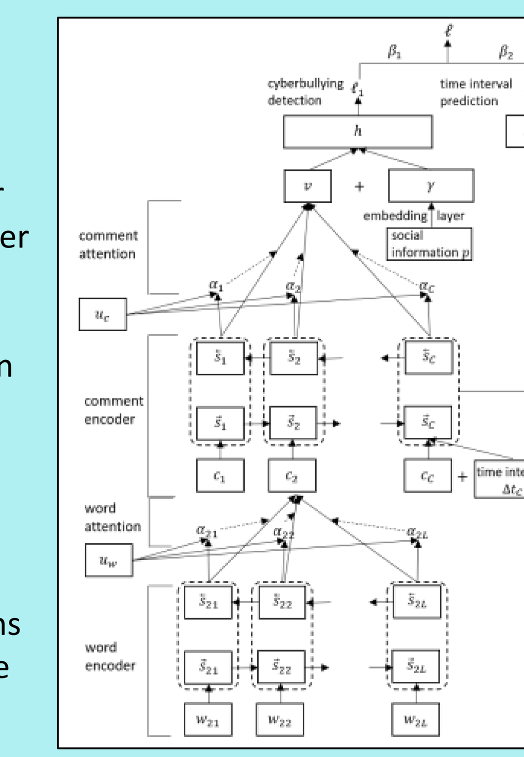

Figure 3.1: Hierarchical Attention Networks for Cyberbullying Detection

Model Evaluation

- The HANCD model was compared to several baseline classification models [1, 2], three end-to-end deep learning models [9,11,17], and two cyberbullying detection models [15,16].
  - As shown in Tables 3.1 and 3.2, the HANCD model provided the best F1 and AUC scores compared to the other models, indicating the advantages of utilizing a time-informed hierarchal framework for cyberbullying detection.

Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.476	0.521	0.501	0.479	0.559	0.236
Naive Bayesian	0.614	0.469	0.607	0.534	0.482	0.355
Logistic Regression	0.700	0.642	0.608	0.677	0.700	0.163
Random Forest	0.585	0.618	0.585	0.617	0.650	0.190
XGBoost	0.715	0.726	0.699	0.674	0.700	0.337
Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	
0.613	0.613	0.708	0.502	0.740	**0.783**	

Table 3.1: Performance Comparisons of Different Models (F1 score).
Values are the average performance of five runs.

Features	Count Vector	Word TF-IDF	N-gram TF-IDF	Char TF-IDF	LIWC	Embedding
KNN	0.770	0.697	0.624	0.708	0.686	0.499
Naive Bayesian	0.706	0.815	0.797	0.786	0.622	0.525
Logistic Regression	0.812	0.825	0.827	0.830	0.776	0.629
Random Forest	0.788	0.804	0.788	0.781	0.743	0.544
XGBoost	0.838	0.828	0.831	0.840	0.772	0.621
Deep Learning Models			Cyberbullying Detection Models			
LSTM	CNN	HAN	Xu et al.	Soni & Singh	HANCD	
0.791	0.781	0.805	0.513	0.810	**0.851**	

Table 3.2: Performance Comparisons of Different Models (AUC score).
Values are the average performance of five runs.

References

1. Breiman, L. (2001), "Random Forests," *Machine Learning,* 45(1), 5-32.
2. Chen, T., and Guestrin, C. (2016), "Xgboost: A Scalable Tree Boosting System," In *KDD*, 785-794.
3. Cheng, L., Guo, R., Silva, Y., Hall, D., and Liu, H. (2019), "Hierarchal Attention Networks for Cyberbullying Detection on the Instagram Social Network," In *SIAM International Conference on Data Mining.*
4. Cheng, L., Li, J., Silva, Y., Hall, D., and Liu, H. (2019), " PI-Bully: Personalized Cyberbullying Detection with Peer Influence," In *28th International Joint Conference on Artificial Intelligence.*
5. Dani, H., Li, J., and Liu, H. (2017), "Sentiment Informed Cyberbullying Detection in Social Media," In *ECML PKDD*, 52-67.
6. Dinakar, K., Reichart, R., and Lieberman, H. (2012), "Modeling the Detection of Textual Cyberbullying," *The Social Mobile Web*, 11(2).
7. Goodboy, A., and Martin, M. (2015), "The Personality Profile of a Cyberbully: Examining the Dark Triad," *Computers in Human Behavior*, 49, 1-4.
8. Gupta, A., Yang, W., Sivakumar, D., Silva, Y., Hall, D., and Barioni, M. (2020), "Temporal Properties of Cyberbullying on Instagram," In *CyberSafety'20.*
9. Hochreiter, S., and Schmidhuber, J. (1997), "Long Short-term Memory," *Neural Computation*, 9(8), 1735-1780.
10. Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., and Mishra, S. (2015), "Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network," In *SocInfo*, 49-66
11. Kim, Y. (2014), "Convolutional Neural Networks for Sentence Classification," *arXiv preprint arXiv:1408.5882.*
12. Kleinberg, K. (2003), "Bursty and Hierarchal Structure in Streams," *Data Mining and Knowledge Discovery*, 7, 373-397.
13. Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., and Lattanner, M. R. (2014), "Bullying in the Digital Age: A Critical Review and Meta-analysis of Cyberbullying Research Among Youth," *Psychological bulletin*, *140*(4), 1073–1137.
14. Pew Research Center (2018), "A Majority of Teens Have Experienced Some Form of Cyberbullying."
15. Soni, D., and Singh, V. (2018), "Time Reveals All Wounds: Modeling Temporal Dynamics of Cyberbullying Sessions," In *ICWSM*.
16. Xu, J. Jun, K., Zhu, X., and Bellmore, A. (2012), "Learning From Bullying Traces in Social Media," In *NAACL HLT*, 656-666.
17. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016), "Hierarchal Attention Networks for Document Classification," In *NAACL HLT*, 1480-1489.

This work was supported by National Science Foundation Award # 1719722