Prediction of Cyberbullying Incidents in a Media-based Social Network

Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, Shivakant Mishra
Department of Computer Science, University of Colorado Boulder
Boulder, Colorado, USA 80309
Email:{homa.hosseinmardi, rahat.rafiq, richard.han, qin.lv, mishras}@colorado.edu

***Abstract*—Cyberbullying is a major problem affecting more than half of all American teens. Prior work has largely focused on detecting cyberbullying after the fact. In this paper, we investigate the prediction of cyberbullying incidents in Instagram, a popular media-based social network. The novelty of this work is building a predictor that can anticipate the occurrence of cyberbullying incidents before they happen. The Instagram media-based social network is well-suited to such prediction since there is an initial posting of an image typically with an associated text caption, followed later by the text comments that form the basis of a specific cyberbullying incident. We extract several important features from the initial posting data for automated cyberbullying prediction, including profanity and linguistic content of the text caption, image content, as well as social graph parameters and temporal content behavior. Evaluations using a real-world Instagram dataset demonstrate that our method achieves high performance in predicting the occurrence of cyberbullying incidents.**

I. Introduction

Cyberbullying is an increasingly serious problem in online social networks. While the research community has begun to explore automated detection of cyberbullying in social networks [1], [2], [3], [4], [5], [6], the area of automated prediction of cyberbullying in social networks remains relatively unexplored. We differentiate between the two in that *cyberbullying detection* leverages both initial user data and later comments by other users to determine, after the fact, whether cyberbullying has occurred. In contrast, *cyberbullying prediction* utilizes only initial user data to predict the occurrence of cyberbullying before it even happens via the comments of other users. Initial user data may be derived from a variety of sources, such as the initial post of an image or video, social graph-based properties, and temporal properties, that are available before the subsequent text-based discussions or comments from which cyberbullying may arise.

Cyberbullying prediction is useful in a variety of dimensions. First, prediction can be used to perform targeted, hence efficient and scalable, detection of cyberbullying in large social networks. Classification is often compute-intensive, and executing a cyberbullying classifier every time a new text comment arrives in a large social network with hundreds of millions or even billions of users would be impractical from the point of view of scalability. The social network provider would need a large and costly number of servers devoted just to cyberbullying classification. Instead, if we can target computational resources more efficiently to focus on the most likely discussions that may be prone to cyberbullying, then this can substantially reduce the cost of cyberbullying detection. Cyberbullying prediction provides the ability to estimate in advance those users or media sessions whose discussions may result in cyberbullying. Therefore, we can efficiently focus our computational resources on these most vulnerable users or media sessions, rather than applying a brute force classification approach to all comments.

Cyberbullying prediction is further useful for identifying in advance users who may be the most vulnerable victims of cyberbullying. As a result, such vulnerable users may be forewarned to protect themselves from potentially negative incoming comments. Also, if the vulnerable users are minors, then their parents may be warned a priori that their children may be more likely victims of cyberbullying. Other resources such as counseling and suicide prevention may be provided to users who are predicted as possible cyberbullying victims.

Facebook, Twitter, YouTube, Ask.fm, and Instagram have been listed as the top five networks with the highest percentage of users reporting experience of cyberbullying [7]. Instagram is of particular interest as it is a media-based mobile social network, which allows users to post and comment on images. Cyberbullying in Instagram can happen in different ways, including posting a humiliating image of someone else by perhaps editing the image, posting mean or hateful comments, aggressive captions or hashtags, or creating fake profiles pretending to be someone else [8]. Figure 1 illustrates a typical profile on Instagram that we use for the prediction of cyberbullying. The user names in the figure are anonymized by circles due to privacy reasons. The rectangles in the figure highlight some of the features that we use for cyberbullying prediction, such as post-time (i.e., this particular photo was posted 49 weeks ago), media caption (i.e., "Nia Liked! (Yes, I am active)"), and the first few comments highlighted by the big rectangle. At Instagram, a user first posts an image with an accompanying text caption. We also see that the user exhibits some graph-based linkages to other users in the social network, e.g., through the user's followers and the people followed by the user. All of these a priori information – the image, its caption, graph properties like number of followers and followings, total number of shared media, etc – is known before the first comments are posted in response to the initial

IEEE/ACM ASONAM 2016, August 18-21, 2016
San Francisco, CA, USA
978-1-5090-2846-7/16/$31.00 2016 IEEE

image. These are the types of features that we utilize in our prediction algorithm.

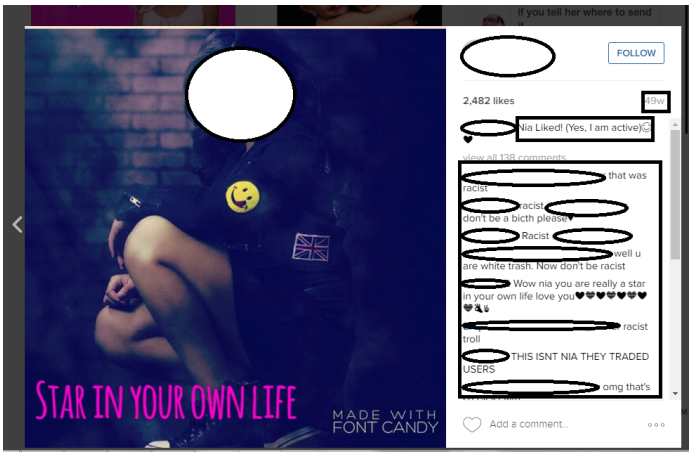

Fig. 1. An example of a typical image post and discussion on Instagram.

We believe that our research is the first to explore prediction of cyberbullying in media-based social networks. In the following, we describe related work in Section II, our data collection and labeling methodology in Section III, design and evaluation results of our prediction algorithm in Section IV. Since accurate prediction enables more targeted detection, we also describe the results from the design of a detection algorithm for cyberbullying in Section V. We finish the paper with a discussion and summary of conclusions.

II. Related Work

The research literature has proposed a variety of approaches for detecting cyberbullying using features such as user context [9], gender information [10], lexical features [11] and graph properties [1]. Textual content has been a major factor in detecting potential cyberbullying instances in online social networks [6], [5], [12], [13], [14], [15], [3], [16]. These works have largely applied a text analysis approach to online comments, since this approach results in higher precision and lower false positives than simpler list-based matching of profane words [17].

Prediction of cyberbullying in the Instagram social network has been proposed in [18] but it focused more on detection rather than prediction because the classifier took as features all the comments associated with a particular media to make a decision. To the best of our knowledge, no previous work has been done in terms of predicting the onset of potential cyberbullying in a multi-modal online social network like Instagram, where a user shares an image/media and other users' comments come streaming in as the discussion unfolds. We believe that it is imperative to predict the potentiality of a shared media session in multi-modal online social networks when it is at its nascent stage (i.e., only the initial posting is available, which on Instagram equates to the image-content and text caption) to facilitate efficient and effective detection and/or monitoring of cyberbullying. Towards this end, it has been observed that certain image-contents like "drug" are highly correlated to cyberbullying and categories such as "tattoo" and "food" are not [19], whereas other researches have reported that gender, age and previous bullying and/or cyberbullying (both perpetrator and victim) experiences are highly correlated to being involved in cyberbullying later [20].

III. Data Collection

Starting from a random seed node, we identified 41K Instagram user ids using a snowball sampling method via the Instagram API. Among these Instagram ids, 25K (61%) users had public profiles while the rest had private profiles. Due to the limitation on the private profiles' lack of shared information, the 25K public user profiles comprise our sample data set. For each public Instagram user, the collected profile data includes the media objects (videos/images) that the user has posted and their associated comments, user id of each user followed by this user, user id of each user who follows this user, and user id of each user who commented on or liked the media objects shared by the user. We consider each media object plus its associated comments as a *media session*.

Labeling data is a costly process and therefore in order to make the labeling of cyberbullying more manageable, we sought to label a smaller subset of these media sessions. First, to have a higher rate of cyberbullying instances, we considered media sessions with at least one profanity word in their associated comments. We tag a comment as "negative" using an approach similar to [21]. For this set of 25K users, 3,165K unique media sessions were collected, and 697K of these sessions have at least one profane word in their comments by users other than the profile owner. The profane words we used were obtained from a dictionary [22], [23].

In addition, we needed media sessions with enough comments so that labelers could adequately assess the frequency or repetition of aggression, which is an important part of the cyberbullying definition. We selected a threshold of 15 as the lower bound on the number of comments in a media session, based on the observation that the average number of comments posted by users other than the profile owner in an Instagram profile is around 16 [24].

Labeling process for media session with non-zero negativity is provided in our previous paper [19] with a thorough analysis of the result. When selecting data for labeling, we chose all 922 media sessions with more than 40% negativity (Set40+), i.e., 40% of the comments of a media session contain at least one profane word. For media sessions whose comments have non-zero but less than 40% negativity, we randomly selected a set of 1,296 media sessions (Set0+). These two sets were pre-filtered based on text analysis to contain profanity. Meaning they were selected based on knowledge about the comments. For the purpose of prediction of cyberbullying incidents, we need to augment these two sets of labeled media sessions to also include media sessions whose comments do not contain any profane word (Set0). This will create a labeled data set that is independent of the content of text comments, namely profanity, and hence usable for prediction, which will not have

access to comment contents. In evaluating the performance of the classifiers, we will look at the result of these three sets separately.

Therefore, we randomly selected 1,164 media sessions from the ones with no profanity usage with the criteria of having more than 15 comments and labeled the new set by the same methodology as provided in our previous paper [19]. This methodology is comprised of first training the labelers on CrowdFlower with the definition of cyberbullying, requiring the labelers to pass a set of questions (quiz mode), subjecting labelers to random test questions while labeling (work mode), and imposing a minimum threshold time for labeling. In total 80 contributors worked on the quiz mode, 68 passed, 11 failed and 1 gave up. Among the 68 contributors, 9 were further filtered out during the labeling process, because they either failed the work mode or rushed through their labeling process (i.e., less than the threshold time for labeling). Labeled data was obtained from the remaining 59 trusted contributors. We also labeled the image contents of the new set of media sessions. Table I shows the statistics related to this labeling process. Furthermore, across all three datasets, we only consider the media sessions whose labeling had a confidence level of more than 60%. More details about the labeling statistics of set40+ and set0+ can be found in our previous paper [19].

TABLE I
LABELING PROCESS STATISTICS FOR MEDIA SESSIONS WITH NO NEGATIVITY. TRUSTED JUDGMENTS ARE THE ONES MADE BY TRUSTED CONTRIBUTORS.

Trusted Judgments	5,638
Untrusted Judgments	72
Average Test Question Accuracy of Trusted Contributors	82%
Labeled Media Sessions per Hour	8

Table II shows the user social graph measurements for the labeled media sessions across all three sets. We differentiate cyberbullying from cyberaggression as a stronger and more specific type of aggressive behavior that is carried out repeatedly against a person who cannot easily defend himself or herself, creating a power imbalance. We observe that the p-values are all less than 0.1 for followers and following, which suggests that these features will be helpful in cyberbullying prediction.

TABLE II
MEAN VALUES OF SOCIAL GRAPH PROPERTIES FOR CYBERBULLYING VERSUS NON-CYBERBULLYING SAMPLES AND CYBERAGGRESSION VERSUS NON CYBERAGGRESSION, ($^{**}p < 0.05, *p < 0.1$ OF APPLYING T-TEST).

Label	*Media objects	*Following	Followers
Non-cyberbullying	1,157.8	721.7	**398,283.7
Cyberbullying	1,198.3	626.5	**465,376.1
Non-cyberaggression	1,152.6	724.4	*393,901.6
Cyberaggression	1,204.3	640.3	*440,403.6

Given this collected labeled data, in the next sections we design and evaluate multimodal classifiers that extend beyond merely the text dimension to further incorporate image-based features and user properties for predicting and detecting cyberbullying.

IV. PREDICTION OF CYBERBULLYING INCIDENTS

Monitoring all social network users for cyberbullying is very costly and not feasible for large social networks. Relatively simple List-based detection based on finding only negative words is relatively simply but have high false positives and low true negatives. As a result, more sophisticated yet more compute-intensive classifiers are needed. Predictive filtering gives us the flexibility to either devote fewer computational resources to the cyberbullying classification task, or for the same number of servers devote more intensive and sophisticated classification.

We consider Instagram as a social network that is especially interesting to study in terms of the benefits of prediction, because of its structure in which the media object is posted first, followed by the discussion comments. What features can we base our prediction upon? In our previous work [19] we observed there is a correlation between non-text features and text features. This finding was our main motivation to consider non-text features for predicting the ensuing comments in a media session. As such, we base our prediction only on the initial posting of the media object, any image features derived from the object, and any properties of the media session observed at or before that posting time. These include the post time itself, the associated text caption's content, the profile owner's social graph properties, including the number of followers and following and the total number of shared media objects.

Each media session in the three introduced sets was labeled by five CrowdFlower contributors. A majority vote criterion was employed to determine whether a media session constituted cyberbullying [19]. To design and train the classifier, five-fold cross-validation was applied to the data such that 80% of the data was used for training in each run and 20% was used for testing. For the dataset Set40+, we found that 49% of the media sessions were labeled as non-cyberbullying and 51% as cyberbullying. For the dataset Set0+, we have 15% non-cyberbullying and 85% cyberbullying examples. The Set0, whose media sessions have no negative words, contained no examples of cyberbullying. To achieve a balanced training data set, we over-sampled the minority class of data labeled as cyberbullying.

We applied a logistic regression classifier to train a predictor with the forward feature selection approach. It is interesting that using only the image content feature, for Set0+, 98% of cyberbullying incidents were captured. The false positive rate for Set0 is 24%. Next adding graph properties will increase the F1-measure for both Set0+ and Set40+. There is a big drop from 24% false positive rate to only 5% within Set0. Adding the number of media objects does not have a big impact on any of the performance metrics for any of the sets. Table III provides the results for each of the three sets. It shows that *cyberbullying incidents can be predicted with 0.99 recall for*

Set0+ and 0.61 recall for Set40+. The best false positive rate over Set0 is 3% , using only the image contents, media and user meta data based on a ridge regression classifier.

In addition, we were interested to explore if prediction could be improved using only a limited set of early comments, not the complete set of comments for a media session. Hence we also show in Table III the results of prediction using the first 15 comments. We see that our predictor is able to improve its recall for Set0+ and Set40+ to 0.72 and 1.0 respectively using a logistic regression classifier. The false positive rate over Set0 has reduced to 1% with the help of text features.

V. Detection of Cyberbullying Incidents

Since effective prediction enables better targeted detection, we were interested in applying a similar methodology as in the prediction section to the training and testing of a detector. This distinction is of course the detection algorithm benefits from having access to the text comments from the discussion.

In this section we only work on the data with non-zero negativity. The idea comes from having a first layer predictor with near to zero false positive. Based on our analysis in our previous paper [19], we evaluated three types of features, namely those features obtained from the content of comments, those derived from shared media objects, and those obtained from user graph properties of the profile owner, such as the number of followers or followings. For the text features, we removed characters such as "!", ">", etc, as a preprocessing step. We first focused on unigrams and bigrams. I will remove this sentence: LIWC categories are derived from unigrams and hence implicitly included as part of this text analysis. Next, we removed stop words such as "and","or", and "for". Finally, each feature vector was normalized by removing the mean zero and scaling to unit variance. In this section we focus on using a ridge regression classifier to be able to interpret the capability of the features in detection of cyberbullying.

Table IV illustrates the improvement of the results for different features. *As we can see, the ridge regression classifier based on text-based unigrams and bigrams achieves the highest recall (0.54) for the Set0+ set and 0.83 recall on Set40+ with more than 200,000 dimensions. However, many of the features with high score have high correlation. For example, there is 0.55 correlation between "shit" and "bitch", 0.47 correlation between "u" and "ugly", and 0.58 correlation between "dumb" and "stupid". These high correlations caused the colliniarity problem in logistic regression classification. By reducing the number of dimensions to 10 (see Table V), we obtained better results in both sets*. More specifically, the reason is due to the high correlation of unigram features, which causes the collinearity problem. To choose the final 10 dimensions, we first remove the words with close-to-zero coefficients. Next, in a backward feature selection approach, we keep removing the variables with high correlation, high p-value and small coefficient, and check the F1-measure after each removal to make sure it will not degrade the performance of the model.

In addition, a variety of non-text features were evaluated, including those features extracted from user behavior (number of shared media objects, following, followers), media properties (likes, post time, caption) and image content. For example, we investigate the feature corresponding to the number of words. However, adding this feature does not provide any value to the classifier performance. It was observed that the number of words is considerably higher for examples of cyberbullying. The reason is the high correlation between the number of words and a set of variables with positive coefficients, namely "bitch", "fuck", "gay", "hate", "shut", "suck", "ugly". Similarly, we considered the "time interval" variable, i.e., the mean time between posts. This variable also has high correlation with cyberbullying indicator words and does not add to the classifier performance. Both of these support our correlation analysis for "time interval" and "word count". Another feature related to the media session is the number of likes the image has received, however it does not provide any improvement with a very small coefficient in the model.

We then considered the image content features. We first remove the features with small coefficient, then examine the correlations. "sick" has positive coefficient and has been seen a lot with Tattoo images. We observe positive correlation between Tattoo and sick. Drugs, Bike and Tattoo are the features with highest percentage for cyberbullying/non-cyberbullying sessions. However, these features are also correlated highly with text features and were unable to improve the classifier's performance.

Besides media session related features, we also considered user related features, including total shared media objects, followers and followings. Total number of media is negatively correlated with words with positive coefficients, meaning it has a higher percentage for cyberbullying incidents. Due to the high correlation and collinearity problem, it causes the performance of the classifier to decrease. Considering follows and followings does not provide any improvement either. There was no significant correlation between these two features and other text features. We should recall the t-test over the mean value of these two features had high p-value, suggesting that there is no significant difference between the values of these features for cyberbullying and non-cyberbullying classes.

For the results provided in Table IV, threshold 0.5 was used for assigning a label to each class. However, by changing the threshold we can tune it to obtain appropriate precision and recall depending on how we are going to use the output of the classifier. Figures 2 and 3 provide the ROC curve for Set40+ and Set0+. The AUC (Area Under Curve) for each set is 0.91 and 0.87 respectively which are reasonably good results.

VI. Discussion

While this paper has introduced prediction of cyberbullying in a media-based mobile social network, there remain a number of areas for improvement. One theme for future work is to improve the performance of our classifier. New algorithms should be considered, such as deep learning and

TABLE III
CYBERBULLYING PREDICTION'S CLASSIFIER PERFORMANCE. USER PROPERTIES ARE FOLLOWERS, FOLLOWING AND TOTAL NUMBER OF SHARED MEDIA OBJECTS.

Features	Set	F1-measure	Precision	Recall	False Positive
Image content	Set40+	0.56	0.62	0.51	0.37
Image content	Set0+	0.27	0.15	0.98	0.83
Image content	Set0	-	-	-	0.24
Following, Image content	Set40+	0.62	0.68	0.51	0.18
Following, Image content	Set0+	0.37	0.23	0.91	0.48
Following, Image content	Set0	-	-	-	0.03
Followers, Following, Image content	Set40+	0.68	0.75	0.60	0.22
Followers, Following , Image content	Set0+	0.42	0.28	0.88	0.34
Followers, Following, Image content	Set0	-	-	-	0.05
Media objects ,Followers, Following, Image content	Set40+	0.69	0.77	0.62	0.21
Media objects ,Followers, Following, Image content	Set0+	0.45	0.31	0.87	0.3
Media objects ,Followers, Following, Image content	Set0	-	-	-	0.04
Post time ,User properties, Image content	Set40+	0.67	0.76	0.61	0.22
Post time ,User properties, Image content	Set0+	0.52	0.38	0.88	0.23
Post time ,User properties, Image content	Set0	-	-	-	0.04
Caption ,Post time ,User properties, Image content	Set40+	0.67	0.76	0.61	0.22
Caption,Post time ,User properties, Image content	Set0+	0.57	0.40	0.99	0.23
Caption ,Post time ,User properties, Image content	Set0	-	-	-	0.03
Early Comments, Caption,Post time ,User properties, Image content	Set40+	0.75	0.78	0.72	0.22
Early Comments, Caption,Post time ,User properties, Image content	Set0+	0.66	0.50	1.00	0.14
Early Comments, Caption,Post time ,User properties, Image content	Set0	-	-	-	0.01

TABLE IV
CYBERBULLYING DETECTION'S CLASSIFIER PERFORMANCE

Features	Dataset	F1-measure	Precision	Recall
unigram	Set0+	0.45	0.54	0.40
unigram	Set40+	0.81	0.87	0.77
unigram, bigram	Set0+	0.5	0.46	0.54
unigram, bigram	Set40+	*0.84*	0.85	0.83
Reduced feature set (83 words)	Set0+	0.53	0.47	0.60
Reduced feature set (83 words)	Set40+	0.83	0.88	0.77
Reduced feature set (10 words)	Set0+	*0.58*	0.50	0.68
Reduced feature set (10 words)	Set40+	*0.85*	0.89	0.81

TABLE V
SELECTED WORDS AS THE INPUT VARIABLES FOR THE CYBERBULLYING CLASSIFIER.

beautiful	sick	work	bitch	fuck	gay	hate	shut	suck	ugly

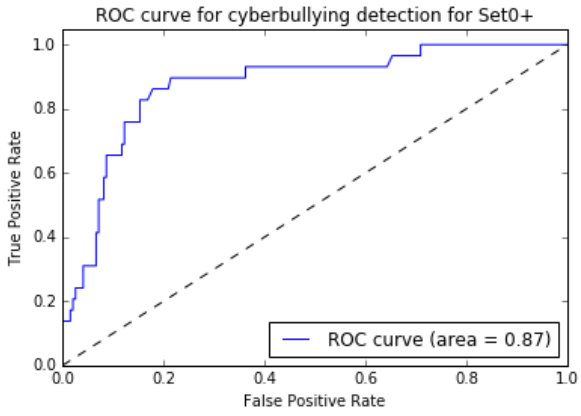

Fig. 2. ROC curve for detection of cyberbullying incidents with more than 0% negativity.

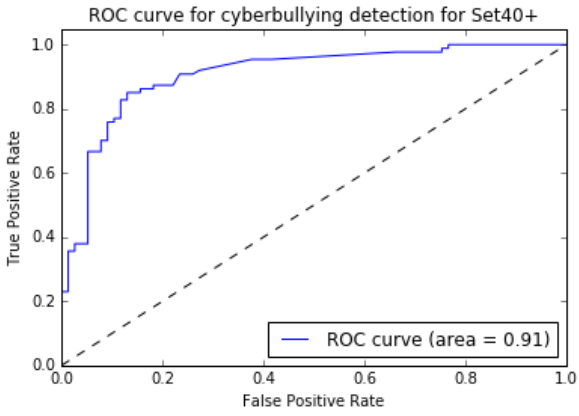

Fig. 3. ROC curve for detection of cyberbullying incidents with more than 40% negativity.

neural networks. More input features should be evaluated, such as new image features, mobile sensor data, etc. Incorporating image features needs to be automated by applying image recognition algorithms. Temporal behavior of comments for a posted media should be taken into account in designing the detection classifier.

In this work we have only considered the image content and image and user metadata for prediction of cyberbullying.

However, based on the improvement seen in using a small number of text comments, we think that considering the commenting history of users in previously shared media can prove to be useful.

Another theme for future work is to obtain greater detail from the labeling surveys. Our experience was that streamlining the survey improved the response rate, quality and speed. However, we desire more detailed labeling, such as for different roles in cyberbullying identifying and differentiating the role of a victim's defender, who may also spew negativity, from a victim's bully or bullies. Finally we can cascade our predictor with a more complicated detection algorithm to make examining cyberbullying-prone media sessions more scalable.

VII. Conclusion

In this work, we have investigated the problem of cyberbullying prediction in the Instagram media-based social network. Using Instagram labeled data, a logistic regression classifier was used to examine the predication power of diverse features. We show that the non-text features such as image and user meta data were central to cyberbullying prediction, where a logistic regression classifier achieved 0.72 recall and 0.78 precision for Set40+, 1.0 recall and 0.50 precision for Set0+. The false positive rate for Set0 is as low as 0.01. Our model aims to detect cyberbullying incidents with non-zero negativity, Specifically, media sessions with high negativity can be detected with 0.81 recall and 0.89 precision using only a 10-dimension feature set extracted from comments. Using the same model, we achieved 0.68 recall and 0.50 precision in predicting cyberbullying incidents with less negativity. Furthermore, none-text features (e.g., image content) are available once a user posts an image at Instagram. Even though non-text features were not very helpful in cyberbullying detection, some of them had high correlation with text features, which gave us the insight of using non-text features in prediction of cyberbullying incidents.

Acknowledgment

We wish to acknowledge financial support for this research from the US National Science Foundation (NSF) through grant CNS 1528138.

References

[1] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*. ACM, 2014, pp. 3–6.

[2] T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, and K. Araki, "Cyberbullying detection based on category relevance maximization," 2013.

[3] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," in *Communications in Information Science and Management Engineering*, ser. CISME'13, 2013.

[4] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying." in *The Social Mobile Web*, 2011.

[5] B. S. Nandhini and J. Sheeba, "Online social network bullying detection using intelligence techniques," *Procedia Computer Science*, vol. 45, pp. 485 – 492, 2015, international Conference on Advanced Computing Technologies and Applications (ICACTA). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S187705091500321X

[6] B. S. Nandhini and J. I. Sheeba, "Cyberbullying detection and classification using information retrieval algorithm," in *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*, ser. ICARCSET '15. New York, NY, USA: ACM, 2015, pp. 20:1–20:5. [Online]. Available: http://doi.acm.org/10.1145/2743065.2743085

[7] D. the Label Anti Bullying Charity, "The annual cyberbullying survey 2013," 2013. [Online]. Available: http://www.ditchthelabel.org/annual-cyber-bullying-survey-cyber-bullying-statistics/

[8] S. Hinduja, "Cyberbullying on Instagram," 2013. [Online]. Available: http://cyberbullying.us/cyberbullying-on-instagram/

[9] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. Jong, *Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, ch. Improving Cyberbullying Detection with User Context, pp. 693–696.

[10] M. Dadvar, F. M. G. de Jong, R. J. F. Ordelman, and R. B. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), Ghent, Belgium*. Ghent: University of Ghent, February 2012, pp. 23–25.

[11] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, ser. SOCIALCOM-PASSAT '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 71–80. [Online]. Available: http://dx.doi.org/10.1109/SocialCom-PASSAT.2012.55

[12] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: query terms and techniques," in *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 2013, pp. 195–204.

[13] V. Nahar, S. Unankard, X. Li, and C. Pang, "Semi-supervised learning for cyberbullying detection in social networks," in *Databases Theory and Applications*, ser. LNCS'12, 2014, pp. 160–171.

[14] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, "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*, ser. ASONAM '15. New York, NY, USA: ACM, 2015, pp. 617–622. [Online]. Available: http://doi.acm.org/10.1145/2808797.2809381

[15] A. K. K. Reynolds and L. Edwards, "Using machine learning to detect cyberbullying," *Machine Learning and Applications, Fourth International Conference on*, vol. 2, pp. 241–244, 2011.

[16] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, pp. 18:1–18:30, Sep. 2012. [Online]. Available: http://doi.acm.org/10.1145/2362394.2362400

[17] S. Sood, J. Antin, and E. Churchill, "Profanity use in online communities," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1481–1490.

[18] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Prediction of cyberbullying incidents on the instagram social network," *CoRR*, vol. abs/1508.06257, 2015. [Online]. Available: http://arxiv.org/abs/1508.06257

[19] H. Hosseinmardi, S. A. Mattson, R. Ibn Rafiq, R. Han, Q. Lv, and S. Mishra, *Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings*. Cham: Springer International Publishing, 2015, ch. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network, pp. 49–66. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-27433-1_4

[20] T. L. Zacchilli and C. Y. Valerio, "The knowledge and prevalence of cyberbullying in a college sample," *Journal of Scientific Psychology*, vol. 5, pp. 12–23, 2011.

[21] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra, "Towards understanding cyberbullying behavior in a semi-anonymous social network," in *Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 2014, pp. 244 – 252.

[22] "Bad word list and swear filter," [accessed 10-November-2014]. [Online]. Available: http://www.noswearing.com/dictionary

[23] N. words list, "Negative words list form, luis von ahn's research group," 2014. [Online]. Available: http://www.cs.cmu.edu/ biglou/resources/

[24] H. Hosseinmardi, R. I. Rafiq, S. Li, Z. Yang, R. Han, S. Mishra, and Q. Lv, "Comparison of common users across Instagram and Ask.fm to better understand cyberbullying," in *The 7th IEEE international Conference on Social Computing and Networking (SocialCom)*, 2014.