Cyber Bullying Detection Using Social and Textual Analysis

Qianjia Huang
Department of Applied Computer Science
The University of Winnipeg
Winnipeg, MB, Canada
huang-q17@webmail.uwinnipeg.ca

Vivek K. Singh
The Media Lab
Massachusetts Institute of Technology
Cambridge, MA, USA
singhv@mit.edu

Pradeep K. Atrey
Department of Computer Science
University at Albany - SUNY
Albany, NY, USA
patrey@albany.edu

ABSTRACT

Cyber Bullying, which often has a deeply negative impact on the victim, has grown as a serious issue among adolescents. To understand the phenomenon of cyber bullying, experts in social science have focused on personality, social relationships and psychological factors involving both the bully and the victim. Recently computer science researchers have also come up with automated methods to identify cyber bullying messages by identifying bullying-related keywords in cyber conversations. However, the accuracy of these textual feature based methods remains limited. In this work, we investigate whether analyzing social network features can improve the accuracy of cyber bullying detection. By analyzing the social network structure between users and deriving features such as number of friends, network embeddedness, and relationship centrality, we find that the detection of cyber bullying can be significantly improved by integrating the textual features with social network features.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Miscellaneous

Keywords

Cyber bullying detection; Social network features; Cyber crime

1. INTRODUCTION

Cyber bullying is emerging as a serious social problem, especially among teenagers. Cyber bullying is defined as "the use of information technology to harm or harass other people in a deliberate, repeated, and hostile manner"[1]. With the advent of social media networks such as Twitter and Facebook, it has become more prevalent. Thus, automatic detection of cyber bullying posts is becoming an increasingly important area of research among social media researchers.

Previous researches on cyber bullying detection have mostly used text-based methods and employed contextual and sentiment features to improve the text mining system. For instance, Reynolds et al. [8] used the number, density and the value of foul words as features to determine the cyber bullying messages. Similarly, Dinakar et al. [6] found that building individual topic-sensitive classifiers improves the detection of cyber bullying messages. Recently, Dadvar et al. [4] also presented an improved model using the user-based features (i.e: the history of the user's activities). All these works are based on text mining. Recently, Nahar et al. [7] built the cyber bullying network graph model and used a ranking method to identify the most active cyber bullying predators and victims. However, the cyber-bullying detection aspect of this work was still purely text-driven.

The accuracy of text-based cyber bullying detection methods still remains limited. In this paper, we advance the state of the art by adopting a more holistic approach. Our main goal is to explore the value of social information in detecting cyber bullying above and beyond the signals available in the textual content of messages. We believe that since bullying is a social problem, information about the social context surrounding the messages might provide vital clues for their detection. Using a corpus of Twitter messages, our approach identifies both social and textual features and creates a composite model for detecting cyber bullying. The obtained results suggest that the social signals are useful for detecting cyber bullying, and that using multiple channels of information (text plus social features) results in higher detection performance.

To the best of our knowledge, this is the first work that uses both textual and social network features to detect cyber bullying. A brief comparison of the proposed approach with existing works is presented in Table 1.

2. PROPOSED APPROACH

The social information related to the bullying messages and their textual content are quantified as follows.

2.1 Characterizing Social Structure Surrounding Bullying Messages

To understand the impact of social structure on the incidence of cyber bullying, we constructed social network graph(s) and derived a set of features. The social features

[1]http://en.wikipedia.org/wiki/Cyberbullying

SAM'14, November 7, 2014, Orlando, Florida, USA

http://dx.doi.org/10.1145/2661126.2661133.

Table 1: A comparison of the proposed approach with existing works on cyber bullying detection

The work	Textual features	User demography	Social network features
Reynolds et al. [8]	yes	no	no
Dinakar et al. [6]	yes	no	no
Nahar et al. [7]	yes	no	no
Dadvar et al. [4]	yes	yes	no
Proposed work	yes	no	yes

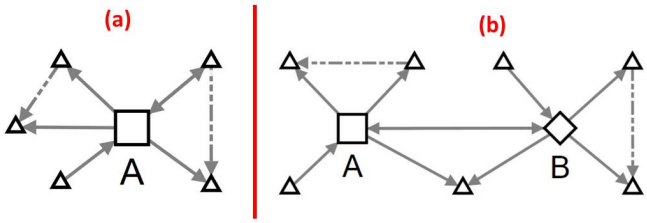

Figure 1: (a): A 1.5 ego-network, (b): A relationship graph defined by combining the 1.5 ego-network graphs of the sender and the receiver.

were derived using the *1.5 ego-networks*[1]. Note that 'ego' refers to an individual focal node. A network has as many egos as it has nodes.

We denote the 'global social network' as a graph $G =< V; E >$ where V is the set of all nodes and E is the set of directed edges over those nodes. We refer to the 1 ego-network of a node v as the graph $G_1(V_1; E_1)$ such that V_1 contains all of the nodes u such that there exists an edge $(v; u)$ in E, and that E_1 contains all the edges from v to the nodes of V_1. Furthermore, we denote by the 1.5 ego-network the graph $G_{1.5}(V_{1.5}; E_{1.5})$ such that $V_{1.5} = V_1$, and $E_{1.5}$ consists of all direct links between nodes that are members of $V_{1.5}$.

In Figure 1(a), the ego node A is marked as a square, and the neighbors are marked as triangles. The edges of the 1-ego network i.e. E_1 are shown using solid lines while the additional edges in $E_{1.5}$ are shown via dashed lines. In this work, we focus on 1.5 ego-networks as they capture a reasonable level of social context (me, my friends, and the relationships between them) while still keeping the data requirements and computational complexity low. This is in congruence with the human processing limits and the social brain hypothesis proposed by Dunbar et al. [5] and also the recent results that have shown the value of 1.5 ego networks in identifying network phenomena [1].

As shown in Figure 1(b), we define the relationship graph of users by combining the 1.5 ego-networks of the two users, the sender 'A' and the receiver 'B'. We represent all communication information for relationship graphs via directed, weighted edges. These graphs allow us to characterize both the sender and the receiver in terms of the position they hold in their respective ego-networks. For example, it allows us to identify which users are more 'central' to their social network and which experience little to no interaction from their peers.

Specifically, we focus on the the following social network features:

Number of nodes is the number of nodes in the resulting 1.5 ego relationship graph. It is an indicator of how large is the (sub)community surrounding the relationship.

Number of edges is the number of nodes in the resulting 1.5 ego relationship graph. It is an indicator of how well-connected is the (sub)community.

Degree centrality is defined as the number of links incident upon a node in a directed network. Prior research has suggested that victims of cyber bullying may have significantly lower self-esteem scores compared to others [2], and may also be more active than others. Hence, we included both in-degree (popularity) and out-degree (activity) centralities in our feature set. Both these features were computed for the sender and the receiver. The *indegree centrality* (denoted by $C_I(i)$) for actor i, is defined as:

$$C_I(i) = \frac{\sum_{j=1}^{n} x_{ji}}{(n-1)}$$

where x_{ji} is the value of the tie from actor j to actor i (the value being either 0 or 1); and n is the number of nodes in the network. The *outdegree centrality* was defined similarly for outgoing connections.

Edge betweenness centrality, denoted by $EB(e)$ for an edge $e \in E$, is defined as a measure of the centrality and influence of edge in network represented by a connected directed graph $G = (V,E)$. It can be calculated as:

$$EB(e) = \sum_{v_i \in V} \sum_{v_j \in V \setminus \{v_i\}} \frac{\sigma_{v_i,v_j}(e)}{\sigma_{v_i,v_j}}$$

Where σ_{v_i,v_j} is the number of shortest paths between nodes v_i and v_j in G, and $\sigma_{v_i,v_j}(e)$ is the number of shortest paths between v_i and v_j which go through e.

In a social network graph, the edges that connect different groups or cliques are more important than other links. The removal of important edges will cause the groups/cliques to be disconnected. In this work, we use edge betweenness of the edge between user A and user B to quantify influence of this particular relationship in the (sub)network.

Links are the number of posts between two users from the labeled conversation.

k-core score of a node is the largest value k such that the node has shared at least k links with at least k different neighbors. More formally it is the largest value k of a 'k-core' containing that node. A 'k-core' is a maximal subgraph that contains nodes of degree k or more. In effect, this feature captures the structural embededness of the node in the network. A user with higher k-core score is more embedded into the network. Note that a similar measure (h-index) is used in different contexts to quantify a user's academic contributions.

2.2 Characterizing Message Content

Based on prior works (e.g. [4, 6, 8]), we choose the following textual features for detecting cyber bullying:

Density of bad words Prior work has suggested that the posts containing curse words are more likely to be bullying messages. We collected 713 curse words (e.g. 'asshole', 'bitch' etc.) and hieroglyphs (such as '5hit', '@ss' etc.) based on online resources and then extended it manually for our purpose.

Density of uppercase letters Based on the finding by Dadvar et al.[4], we chose the ratio of capital letters in a post message as a feature because it might be related to 'shouting' in the online settings.

Number of exclamation points and question marks Just like the uppercase letters, exclamation points and question marks also stand as emotional comments. As cyber bullying is often connected to the strong emotions, we chose

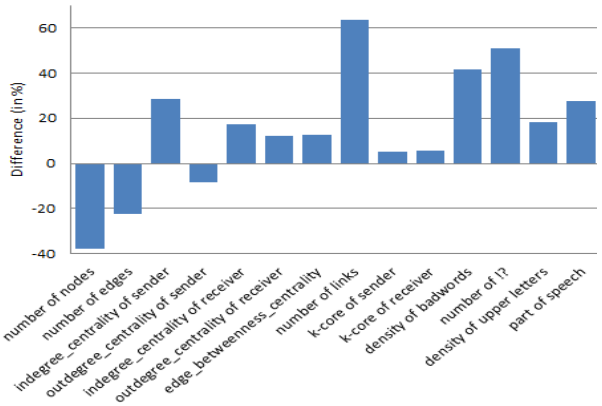

Figure 2: Comparison of textual and social network features for bullying and non-bullying messages.

the number of exclamation points and question marks as a feature for our model.

Number of smileys As smileys are the most commonly used indicators of emotion, we chose the number of smileys as a feature for our model.

Part-of-speech tags Dinakar et al. [6] have suggested part-of-speech tags (JJ-DT, PRP-VBP and VB-PRP) as features to detect commonly occurring bigram pairs in the training data for positive examples, such as 'you are', '... yourself' and so on.

3. EXPERIMENTS AND RESULTS

3.1 Corpus

We used the Twitter corpus from the CAW 2.0 data set. This corpus contains about 900,000 posts of 27135 users (one XML file for each user) from Dec 2008 to Jan 2009. We used this corpus not only because it has been widely used in previous literature, but also because it provides the information for both textual content and social network. We randomly chose 800 files, and kept the comments posted with @, which represent direct paths between two users. This resulted in a data set of approximately 13,000 messages. Then, we asked 3 students to label each message as cyber bullying two times. We labeled each post as 'yes' or 'no' based on whether it was considered to involve cyber bullying. The labeling students were in disagreement or undecided about 36 messages and they were simply removed from the database. This process led to 257 messages marked as bullying messages (around 2 percent of the total). Since the social interaction features of both the sender and the receiver were important for our approach, we decided to focus on the messages for which the interaction history of both the sender and the receiver was available in the CAW 2.0 database. This finally yielded a data set with 2150 pairs of users (relationship graphs similar to Figure 3) and 4865 messages between them (91 of them were labeled as bullying messages). Our results are based on this set of 4865 messages.

3.2 Exploratory Analysis

In Figure 2, we show a comparison of the average values obtained for different features in the two categories as follows: $\frac{mean(bully)-mean(nonbully)}{mean(bully)}$. As can be seen, the values for 4 out of 10 social features showed a difference of at least 20%. Similarly, for 4 out of 5 textual features, the differences between bullying and non-bullying messages were at least 20% . One textual feature, number of smileys, is not shown in the figure as the non-bullying category had zero samples, and the relative difference is undefined, but useful. These differences indicate that both these sets of features could indeed be useful for classification via various machine learning techniques.

Table 2: Classification results(average) for textual, social and composite models using different algorithms.

	Textual features		Social features		All features	
	ROC	TP	ROC	TP	ROC	TP
Bagging	0.578	0.067	0.749	0.374	0.700	0.211
J48	0.492	0.081	0.735	0.378	0.628	0.259
SMO	0.600	0.452	0.689	0.706	0.703	0.733
Dagging	**0.642**	**0.519**	**0.715**	**0.722**	**0.755**	**0.763**
Naive Bayes	0.584	0.489	0.699	0.704	0.695	0.723
ZeroR	0.5	0	0.5	0	0.5	0

Table 3: Top ranked features(average) for textual, social and composite models.

Textual features	Social features	All features
1. Num:Exclamations	1. Num:Links	1. Num:Links
2. Density: Bad words	2. Num:Edges	2. Num:Edges
3. Num:PosBiGrams	3. Num:Nodes	3. Num:Nodes

3.3 Classification

The nature of the problem of cyber bullying predicates dealing with highly imbalanced classes. To mitigate the effects of imbalance, we applied the 'SMOTE' approach to create a balanced data set for training, before testing in the realistic imbalanced settings. SMOTE (Synthetic Minority Oversampling TEchnique) works by under sampling the majority class and over sampling the minority class. However, it mitigates the problem of over fitting caused by simple replication of data points by generating newer (synthetic) examples by operating in 'feature space' rather than 'data space [3].

We used 70% of total messages (both bullying and non-bullying messages) as our training data set. The remaining 30% were used as the testing data set. For accuracy, we had 10 groups of training and testing data sets which were randomly produced.

We used Weka 3.0 as the implementation tool for this work and tried multiple well-known classification algorithms including J48, Naive Bayes, SMO, Bagging and Dagging. We also ranked the features using the Information Gain algorithm. The classification results (average of 10 runs) for the different algorithms are reported in Table 2 in terms of the Receiver operating characteristic (ROC) performance and true positive rate. We have also listed top three features(average of 10 runs) identified for the textual model, the 'social' model and the composite model in Table 3.

Note that the traditional *accuracy* measure is not a good metric when the classes are imbalanced and/or the cost of misclassification varies dramatically between the two classes [3]. For example, in the current setup a baseline majority based classifier (ZeroR) can achieve 98% accuracy on the prediction but would not serve as a useful detector of cyber bullying in practice. Hence other metrics such as Receiver operating characteristic (ROC) and true positive rate for the rarer class are used for evaluating the performance of such classifiers. Similarly, many practitioners focus on the rarer (important) class when making their decision for the classification algorithm, and therefore, we also report the TP rate for the rarer class.

We noticed a clear trend of increasing classifier performance as we shift from 'textual features' models to 'compos-

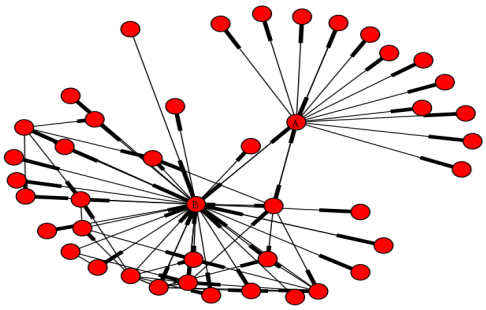

Figure 3: The relationship graph for a bullying message. Here, node A is the bully and node B is the victim. All nodes are anonymized.

ite' models in terms of both ROC and TP rates. In terms of the performance of different classification algorithms, Dagging demonstrates much better results than other methods. It substantially outperforms other methods in terms of the TP rates while also yielding higher ROC values. Focusing on the best obtained results for each column we find that ROC values show a growing trend from a textual features model (0.642), to a social features model (0.715), to a composite model (0.755). We observe a similar trend in terms of the True Positive Rate.

Table 3 shows that the most highly ranked textual features were the number of exclamations, bad words, and PosBiGrams respectively. Each of these has been discussed in prior work. Amongst social features, number of links, number of nodes and the number of edges were found to be the most highly ranked features. Looking at these features together with their relative values in the bullying and non-bullying instances (refer Figure 2), we notice that bullying is much less likely to occur when the 1.5 ego relationship involves many people (nodes) with multiple interconnections (edges). This may allow users to have social support and the potential bulliers may consider it risky to bully a person when there are multiple interconnections surrounding the relationship. A higher number of messages exchanged between users (links) was found to indicate a higher likelihood for a bullying messages. This is an interesting and alarming observation that suggests that bulliers and victims may actually engage in frequent interactions. Lastly, the top three features in the 'composite' model were the same as the 'social' model, which suggests that social aspects of the relationship between the two users are more predictive of a potential bullying scenario than even a content analysis of the messages shared. Note though that the fourth ranked feature for the 'composite' model was the number of exclamation marks in text, which explains the overall better performance of the composite model than only the social model.

There were some cases of cyber bullying, for example, the message "stop farting on people", which were not detected as bullying messages using the textual features models. However, many of these errors were rectified by using social network features. For example, in the abovementioned case, multiple social network features (see Figure 3 for the relationship graph) were closer to those found in the bullying category than the non-bullying category (e.g. *number of nodes*: observed = 43, mean-bully = 43.35, mean-nonbully = 59.68;). Hence a combination of these social features allowed the classification algorithms to learn and classify these messages correctly. Furthermore, note that in this example user B is more active than user A (out-degree centrality scores 0.571 vs 0.333), which is in congruence with the hypothesis from psychology that victims of cyberbullying may be more active than others.

Clearly, current study has some limitations too. Like many other recent efforts in cyber-bullying it focuses on a particular social network (Twitter), which incurs a selection bias. Using Twitter data also meant that we could not consider demographic factors in our analysis. In future, we also plan to use more sophisticated features for social network analysis (e.g. Adamic distance, community cliques) as well as textual analysis (e.g. valence/arousal). However, the focus of this paper remains on testing whether the complimentary layer of information provided by social network analysis is useful for improving cyber-bullying detection.

4. CONCLUSION

This paper advances the state of the art in cyber bullying detection beyond textual analysis to also consider the social relationships in which these bullying messages are exchanged. Our results indicate that social features are useful in detecting cyber bullying. In effect, it suggests that understanding the social context in which a message is exchanged is just as important as the message itself. In future, similar approaches can be applied over more fine grained data about human behavior to detect cyber and physical social bullying in different settings, thus paving the way for a safer environment for bullied individuals in different social settings.

5. ACKNOWLEDGMENTS

This research is supported in parts by the NSERC, Canada, Grant No. 371714 and the University at Albany - SUNY Grant No. 640075.

6. REFERENCES

[1] Y. Altshuler, M. Fire, E. Shmueli, Y. Elovici, A. Bruckstein, A. S. Pentland, and D. Lazer. The social amplifier - reaction of human communities to emergencies. *Journal of Statistical Physics*, 152(3):399–418, 2013.

[2] D. C. Campfield. *Cyber bullying and victimization: Psychosocial characteristics of bullies, victims, and bully/victims*. ProQuest, 2008.

[3] N. V. Chawla. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 853–867. Springer, 2005.

[4] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong. Improving cyberbullying detection with user context. In *Advances in Information Retrieval*, pages 693–696. Springer, 2013.

[5] T. Dávid-Barrett and R. Dunbar. Processing power limits social group size: computational evidence for the cognitive costs of sociality. *Proceedings of the Royal Society B: Biological Sciences*, 280(1765), 2013.

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

[7] V. Nahar, X. Li, and C. Pang. An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering*, 3(5):238–247, 2013.

[8] K. Reynolds, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 241–244. IEEE, 2011.