Cyberbullying Detection Using Probabilistic Socio-Textual Information Fusion

Vivek K. Singh
School of Communication & Information
Rutgers University, NJ, USA
Email: v.singh@rutgers.edu

Qianjia Huang
School of Communication & Information
Rutgers University, NJ, USA
Email: shy.huang@rutgers.edu

Pradeep K. Atrey
Department of Computer Science
SUNY, Albany, NY, USA
Email: patrey@albany.edu

***Abstract*—Cyberbullying is an important socio-technical challenge in Online Social Networks (OSN). With the growth trends of heterogeneous data in OSN, better network characterization, and textual feature sophistication, recent efforts have realized the value of looking at heterogeneous modes of information including textual features, social features, and image-based features for better cyberbullying detection. These approaches, however, still use these features either individually or combine them 'naively' without considering the different confidence levels associated with each feature or the interdependencies between features. We propose a novel probabilistic information fusion framework that utilizes confidence score and interdependencies associated with different social and textual features and uses those to build better predictors for cyberbullying. The performance of the proposed approach was compared to a recent approach in literature which used a similar dataset and features and the proposed approach resulted in significant improvements in terms of cyberbullying detection.**

1. Introduction

With large aspects of human life, including education (e.g. MOOCs), work (e.g. Mechanical Turk, Task-rabbits), and social interactions (e.g. Twitter, Facebook) moving to the cyber domain, there is a growing need to keep the online environments safe from abuse and exploitation for teenagers and adults alike. According to a recent National Crime Prevention Council report, more than 40% of teenagers in the US have reported being cyberbullied[1]. This is especially worrying, as the multiple studies have reported that the victims of cyberbullying often deal with psychiatric and psychosomatic disorders [3], and a study in Britain has reported that nearly half of suicides among young people are related to bullying[2].

Cyberbullying may be defined as the following: "When the Internet, cellphones or other devices are used to send or post text or images intended to hurt or embarrass another person" [5]. With the growth of Online Social Networks (OSN) such as Twitter and Facebook, it has become more prevalent in recent years. Thus, automatic detection of cyber bullying posts is becoming an increasingly important area of research among social network researchers [8], [10], [13].

1. http://www.ncpc.org/resources/files/pdf/bullying/cyberbullying.pdf
2. http://www.bbc.co.uk/news/10302550

Previous research on automated cyberbullying detection has been dominated by text-mining and analysis. This has included approaches based on specific keywords, emotions in the content, as well as the topics of the conversation (e.g. [11], [6]). With the growth in multimodal online content as well as social network analysis methods, recent efforts have realized the value of looking at heterogeneous modes of information including textual features, social features, and image-based features for better cyberbullying detection [8], [10], [13]. These approaches, however, have still used these features either individually or combined them 'naively' without employing any sophisticated methods for fusing heterogeneous data sources. Building upon the maxim of 'the whole being greater than the sum of its parts', this work proposes the use of a sophisticated probabilistic fusion method to detect cyberbullying incidents. The proposed method leverages the differences in the contributions of heterogeneous data features toward the classification goal and the associations between different features to generate a better classification output.

While similar fusion approaches have been considered in the information fusion literature on audio-video processing before (e.g. [2], [12]), this is the first effort to utilize such fusion approaches to combine heterogeneous (social, text) information sources for cyberbullying detection. We test the proposed approach using multiple text (e.g. density of bad words, part-of-speech tags) and social network based features (e.g. number of nodes, degree centrality) and the results demonstrate the efficacy of the proposed approach. With the growth trends of heterogeneous data in OSN, better network characterization, and textual feature sophistication, the proposed fusion approach could provide the backbone for integrating such features for enhanced cyberbullying detection.

2. Proposed Method

This work proposes a newer method to improve the performance of cyberbullying detection methods beyond that obtainable by using a single modality of data, or its 'naive' assimilation. Naive assimilation here refers to the most common approach of simply combining the features as-is, without identifying the different confidence values for different features or the inter-dependencies between them.

2.1. Fusion Model

We consider a generic cyberbullying classification problem, where the system uses n different data modalities $\{f_i, 1 \leq i \leq n\}$ (socio-textual features in our case) and outputs local decisions about cyberbullying incident C at time instant t in terms of n probability values, $p_1(t), p_2(t), \ldots, p_n(t)$. Here $p_i(t) = P(C|f_{i,t})$, which represents the probability that cyberbullying event C has occurred based on modality f_i at time t. These probabilistic decisions are iteratively fused using a Bayesian approach. Let $P(C|\mathbf{f}_t^{i-1})$ denote probability of occurrence of C at time t based on modalities f_1, f_2, …, f_{i-1}. The individual decision $P(C|f_{i,t})$ based on i^{th} modality is integrated into $P(C|\mathbf{f}_t^{i-1})$ as follows [2]. The fused decision, i.e. probability $P(C|\mathbf{f}_t^{i})$ of occurrence of C at time instant t based on modality set $\mathbf{f}^i$ is given by:

$$P(C|\mathbf{f}_t^{i}) = \frac{P^+ \times \exp(\alpha_{f_i,\mathbf{f}^{i-1}}(t))}{P^+ \times \exp(\alpha_{f_i,\mathbf{f}^{i-1}}(t)) + P^- \times \exp(-\alpha_{f_i,\mathbf{f}^{i-1}}(t))} \quad (1)$$

where, $P^+ = P(C|\mathbf{f}_t^{i-1})^{\mathbf{w}_{i-1}(t)} \times P(C|f_{i,t})^{w_i(t)}$ and $P^- = (1 - P(C|\mathbf{f}_t^{i-1}))^{\mathbf{w}_{i-1}(t)} \times (1 - P(C|f_{i,t}))^{w_i(t)}$ are the weighted combined probabilities of the occurrence and non-occurrence of cyberbullying, respectively, using $\mathbf{f}^{i-1}$ and f_i at time instant t. The terms $\mathbf{w}_{i-1}(t)$ and $w_i(t)$ are confidence scores of $\mathbf{f}^{i-1}$ and f_i at time t, respectively. Note that $\mathbf{w}_{i-1} + w_i = 1$. The exp term denotes the exponential function, whereas $\alpha_{f_i,\mathbf{f}^{i-1}}(t) \in [-1, 1]$ is the measure of agreement/disagreement (called Agreement Coefficient [2]) between two modalities $\mathbf{f}^{i-1}$ and f_i at time t. Note that values -1 and 1 of this measure represent the full disagreement and the full agreement, respectively, between the two modalities. The computation of confidence scores and modeling of $\alpha_{f_i,\mathbf{f}^{i-1}}$ are described in the following subsections.

Adapted from [2], the proposed fusion model uses the logarithmic opinion pool (LOGP) consensus rule satisfying the assumption of conditional (content-wise) independence among different modalities [7], and it normalizes the outcome over the two aspects, the occurrence and non-occurrence, of a cyberbullying incident (see denominator term in Eq. (1)). Note that this fusion model outputs a higher overall probability of the positive output of the classification task when more concurring evidences are combined.

2.1.1. Determining agreement coefficient between modalities. The system computes the agreement coefficient $\alpha_{i,k}(t)$, between the modalities f_i and f_k at time instant t by iteratively averaging the past agreement coefficients with the current observation. Precisely, $\alpha_{i,k}(t)$ is computed as:

$$\alpha_{i,k}(t) = \beta \times (1 - 2 \times |p_i(t) - p_k(t)|) + (1 - \beta) \times \alpha_{i,k}(t-1) \quad (2)$$

where, $p_i(t)$ and $p_k(t)$ are the individual probabilities of the occurrence of cyberbullying based on modalities f_i and f_k, respectively, at time $t \geq 1$; and $\alpha_{i,k}(0) = 1 - 2 \times |p_i(0) - p_k(0)|$. These probabilities represent decisions about the detection tasks. Exactly same probabilities would imply full agreement ($\alpha_i = 1$) whereas totally dissimilar probabilities would mean that the two modalities fully contradict each other ($\alpha_{i,k} = -1$) [2]. In Eq. (2), the term $(1 - 2 \times |p_i(t) - p_k(t)|)$ represents the degree of agreement at the time instant t, and the term $\alpha_{i,k}(t-1)$ is the agreement coefficient at the time instant $t-1$. The term β is the weight of the current agreement coefficient and $1 - \beta$ is same for the past agreement coefficient.

The agreement coefficient between the two sources $\mathbf{f}^{i-1}$ and f_i is modeled as:

$$\alpha_{f_i,\mathbf{f}^{i-1}}(t) = \frac{1}{i-1} \sum_{s=1}^{i-1} \alpha_{s,i}(t) \quad (3)$$

where, $\alpha_{s,i}$ for $1 \leq s \leq i-1$, $1 < i \leq n$ are the agreement coefficients between the s^{th} and i^{th} modalities. The agreement fusion model given in Eq. (3) uses *average-link clustering* [9], which considers the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster. In our case, one cluster is a group $\mathbf{f}^{i-1}$ of $i-1$ modalities and the average distance of a new i^{th} modality with this cluster is calculated.

2.1.2. Determining confidence scores of modalities. The confidence in a modality is related to how accurate it has been in the past. The higher the accuracy of a modality, higher the confidence we would have in it. Using the training data, we compute the accuracy (and therefore the confidence score) of a modality by determining how many times a cyberbullying incident is correctly detected based on it out of the total number of incidents.

While the individual modalities have their own confidence scores, we can compute the overall confidence in a group of modalities as follows. Given that the two modalities f_i and f_k have their confidence scores w_i and w_k, respectively, the system uses a Bayesian method to fuse the confidence scores of individual modalities. The overall confidence w_{ik} in a group of two modalities, f_i and f_k, is computed as follows:

$$w_{ik} = \frac{w_i \times w_k}{w_i \times w_k + (1 - w_i) \times (1 - w_k)} \quad (4)$$

In the above formulation, we assume that although the modalities are correlated in their decisions; they are mutually independent in terms of their confidence scores [2]. The below model for confidence fusion allows for the overall confidence level of a group of modalities to increase monotonically as the more trusted modalities are added into the group.

For n modalities, the overall confidence score is iteratively computed. Let $\mathbf{w}_{i-1}$ be the overall confidence scores of a group of $i-1$ modalities. By fusing the confidence score w_i of i^{th} modality with $\mathbf{w}_{i-1}$, the overall confidence score $\mathbf{w}_i$ in a group of i modalities is computed as:

$$\mathbf{w}_i = \frac{\mathbf{w}_{i-1} \times w_i}{\mathbf{w}_{i-1} \times w_i + (1 - \mathbf{w}_{i-1}) \times (1 - w_i)} \quad (5)$$

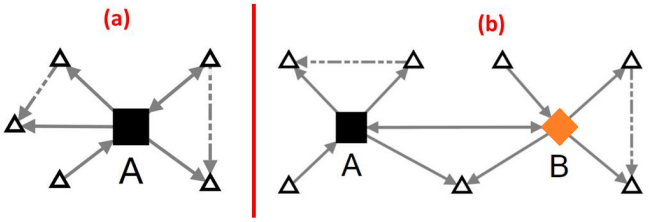

Figure 1. (a): Example of a 1.5-ego-network, (b): A relationship graph created by combining the 1.5-ego-network graphs of a sender (A) and a receiver (B).

2.2. Characterizing Social Network and Textual Content

Similar to [8], the social and textual features have been quantified as follows. The social features were derived using the *1.5 ego-networks* [1], where 'ego' refers to an individual focal node.

Let the 'global social network' be represented as a graph $G =< V; E >$ where V is the set of all the nodes and E is the set of directed edges over those nodes. We define 1-ego-network of a node v as the graph $G_1(V_1; E_1)$ such that it includes all the neighbors of v. We define 1.5-ego-network be the graph $G_{1.5}(V_{1.5}; E_{1.5})$ such that it includes the all the interconnections (edges) between the nodes present in the 1-ego-network defined above.

In Fig. 1(a), the ego-node A has been shown as a solid square and the neighbors of A are marked as triangles. The solid lines represent the edges of the 1-ego-network i.e. E_1 while the dashed lines represent the additional edges in $E_{1.5}$. In this work, we focus on 1.5-ego-networks to characterize an individual's social network 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 [8].

As shown in Fig. 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' (shown as an orange trapezoid). Taken together, these graphs allow us to characterize both the sender and the receiver in terms of their social context. For example, the relationship graph can be used to characterize which users are more 'central' to their social network and also to identify the number of common friends between users A and B.

Specifically, we focus on the following social network features defined for the abovementioned relationship graph: **(1) number of nodes, (2) number of edges, (3) degree centrality**- with variants for in-degree, out-degree, sender and receiver resulting in four different features, **(4) edge betweenness centrality** of the edge between sender and receiver, **(5) tie strength** between a sender and a receiver, and **(6) community embeddedness** measured as a k-core score for the sender and receiver (two features) [8], resulting in a total of ten social features.

Similarly, based on [8] we choose the following textual features: **(1) density of bad words, (2) density of uppercase letters, (3) number of exclamation points and question marks, (4) number of smileys,** and **(5) part-of-speech-tags**.

3. Experiments and Results

We use a labeled cyberbullying dataset as used in [8]. This dataset is a subset of the Twitter corpus from the CAW 2.0 data set, which has been annotated by three labelers for the propensity of cyberbullying. This data set contains 4865 messages with 93 (roughly 2%) of them labeled as bullying messages. The nature of the problem of cyberbullying predicates dealing with such highly imbalanced classes. To mitigate the effects of imbalance, we undertook both minority upsampling and majority down sampling as suggested in the literature [4]. To up sample the minority class we applied the 'SMOTE' method. 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' [4]. The resulting dataset consisted of 186 positive samples and 814 negative samples. While still being imbalanced, this newer dataset allows supervised machine learning to have enough positive samples to learn in a training set and be then applied to a test set.

We split the abovementioned dataset randomly into equal sized training and test sets and tried three different classification approaches. First is the baseline majority based classifier (ZeroR), which simply classifies all rows as the majority class (non-bullying in this case). Second is the "naive fusion" or "early fusion" method, which is the most common approach adopted by practitioners when dealing with multiple features. In this early fusion approach each feature is considered independently and there is no notion of association between features or difference in the confidence levels associated with the features. Specifically, this early fusion was the approach adopted by the authors of [8] using the same set of features and building upon the same dataset. Third approach implemented is the proposed probabilistic "late fusion". In this "late fusion" approach, we assumed the instances of the training data to be available *a-priori* and used them to model the confidence score for the different modalities. Similarly, the various instances were assumed to arrive in a discrete temporal order allowing for the computation of agreement coefficient among features based on the past instances.

Note that the traditional *accuracy* measure may not be sufficient when considering the performance of different classification approaches, as the classes in the dataset are imbalanced and the cost of misclassification varies dramatically between the two classes [4]. As an illustration, in the current setup, a baseline majority based classifier (ZeroR) achieved 81% accuracy on the prediction but would not serve as a useful detector of cyberbullying in practice as it would not detect even a single *positive* example of cyberbullying correctly. Hence, beyond accuracy scores, we also report other metrics such as precision, recall (also called true positive rate (TPR)), and F1 score to compare the performance of different classification approaches considered. (Note that F1 score is the geometric mean of precision and recall.)

TABLE 1. CLASSIFICATION PERFORMANCE OF DIFFERENT APPROACHES FOR CYBERBULLYING DETECTION.

	Accuracy	Precision	Recall	F1-score
Baseline -ZeroR	0.81	0.00	0.00	0.00
Early Fusion	0.76	0.37	0.38	0.37
Proposed Approach	**0.89**	**0.82**	**0.53**	**0.64**

TABLE 2. CLASSIFICATION PERFORMANCE OF THE PROPOSED APPROACH AT DIFFERENT TRAINING SET:TEST SET SPLIT RATIOS

Training:Test Split	Accuracy	Precision	Recall	F1-score
300:700	0.88	0.86	0.45	0.59
500:500	0.89	0.82	0.53	0.64
666:334	0.90	0.92	0.51	0.65

As can be seen in Table 1, there was a clear trend of improving classifier performance as we shift from the baseline ZeroR approach to the proposed approach in terms of the various metrics considered. While the early fusion approach performs better than the baseline ZeroR approach on most metrics - the exception of accuracy has already been discussed above - the proposed "late fusion" approach yields significant improvements over even the early fusion approach. While the accuracy jumps to 89%, more importantly we see noticeable improvements in terms of precision, recall and F1-score over the early fusion based method presented in [8]. These results demonstrate the value of the proposed fusion approach for better cyberbullying detection using heterogeneous social and textual features.

As a next step of analysis, we considered the effect of the relative size of the training set on the performance of the proposed approach. The results for the performance of the proposed approach at different ratios of training and test sizes are shown in Table 2. As can be seen, there is a general consistency in the performance of the proposed approach over different split ratios. There is a small increase in accuracy and F1-score, corresponding with the increase in the size of the training set. While this makes sense as a bigger training set allows for more opportunities to learn the relevant parameters, the relatively small change compared to that observed across approaches in Table 1 lends support to the belief that the proposed approach is robust to reasonable variations in the training:test split ratios and would thus be useful in practical cyberbullying detection problems.

There exist some limitations in the current work. Like many similar efforts on cyberbullying detection, it focuses on a particular social network (Twitter) and does not employ a representative user sample [5], [11]. The balanced dataset used in this work may not mimic the real-world settings. Hence the results presented should be interpreted in terms of the *relative* performance obtained by different approaches rather than as absolute values. In future, we also plan to use more sophisticated features for both social network analysis (e.g. community cliques, Adamic-Adar distance) as well as textual analysis (e.g. topic modeling). However, rather than focusing on individual features the focus of this paper remains on a fusion approach to combine multiple features for improved cyberbullying detection.

4. Conclusion

This paper advances the state of the art in cyberbullying detection beyond individual features to propose a novel method for fusing heterogeneous social and textual features for improved cyberbullying detection. The proposed method leverages the differences in the contributions of heterogeneous data features toward the classification goal and the associations between different features to generate a better classification performance. The obtained results were compared to a recent approach, which used similar dataset and features, and the proposed method resulted in significant improvements in the classification results. With the growth trends in multimodal data, better social network characterization, and textual feature analysis, the proposed fusion approach could provide the backbone for integrating such features for enhanced cyberbullying detection in different settings, thus paving the way for safer online social networks.

Acknowledgment

This material is in part based upon work supported by the National Science Foundation under Grant No. 1464287.

References

[1] Y. Altshuler and et al. The social amplifier - reaction of human communities to emergencies. *Journal of Statistical Physics*, 152(3):399–418, 2013.

[2] P. K. Atrey, M. S. Kankanhalli, and R. Jain. Information assimilation framework for event detection in multimedia surveillance systems. *Multimedia systems*, 12(3):239–253, 2006.

[3] L. Beckman, C. Hagquist, and L. Hellström. Does the association with psychosomatic health problems differ between cyberbullying and traditional bullying? *Emotional and behavioural difficulties*, 17(3-4):421–434, 2012.

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

[5] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18, 2012.

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

[7] C. Genest and J. V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, pages 114–135, 1986.

[8] Q. Huang, V. K. Singh, and P. K. Atrey. Cyber bullying detection using social and textual analysis. In *Proc. Int. Workshop on Socially-Aware Multimedia*, pages 3–6. ACM, 2014.

[9] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[10] R. I. Rafiq and et al. Careful what you share in six seconds: detecting cyberbullying instances in vine. In *Proc. ASONAM*, pages 617–622. ACM, 2015.

[11] K. Reynolds, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying. In *Proc. Int. Conf. Machine Learning and Applications and Workshops (ICMLA)*, volume 2, pages 241–244. IEEE, 2011.

[12] C. G. Snoek, M. Worring, and A. W. Smeulders. Early versus late fusion in semantic video analysis. In *Proc. ACM Multimedia*, pages 399–402. ACM, 2005.

[13] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin. Identification and characterization of cyberbullying dynamics in an online social network. In *Proc. ASONAM*, pages 280–285. ACM, 2015.