OPTIMAL ONLINE CYBERBULLYING DETECTION

Daphney–Stavroula Zois[⋆] *Angeliki Kapodistria*[⋆] *Mengfan Yao*[†] *Charalampos Chelmis*[†]

[⋆]Electrical and Computer Engineering Department
[†]Department of Computer Science
University at Albany, SUNY, Albany, NY, USA
Emails: {dzois, akapodistria, myao, cchelmis}@albany.edu

ABSTRACT

Cyberbullying has emerged as a serious societal and public health problem that demands accurate methods for the detection of cyberbullying instances in an effort to mitigate the consequences. While techniques to automatically detect cyberbullying incidents have been developed, the scalability and timeliness of existing cyberbullying detection approaches have largely been ignored. We address this gap by formulating cyberbullying detection as a sequential hypothesis testing problem. Based on this formulation, we propose a novel algorithm designed to reduce the time to raise a cyberbullying alert by drastically reducing the number of feature evaluations necessary for a decision to be made. We demonstrate the effectiveness of our approach using a real–world dataset from Twitter, one of the top five networks with the highest percentage of users reporting cyberbullying instances. We show that our approach is highly scalable while not sacrificing accuracy for scalability.

Index Terms— cyber harassment, social media, optimization algorithm, selection process, classification

1. INTRODUCTION

Bullying, once limited to physical spaces (e.g., schools, workplaces or sports fields) and particular times of the day (e.g., school hours), can now occur anytime, anywhere. Cyberbullying can take many forms, however, it typically refers to repeated and hostile behavior (e.g., hurtful comments, videos and images) performed in an effort to intentionally and repeatedly harass or harm individuals [1]. The consequences can be devastating: learning difficulties, psychological suffering and isolation, escalated physical confrontations, suicide. Statistics are staggering: over half of adolescents have engaged in or have been cyberbullied, while $10\% - 20\%$ experience it daily[1]. The seriousness of the problem has led to a number of approaches to detect abusive behavior in online social networks based on text features, such as N–grams [2, 3, 4, 5], part–of–speech tags [2, 5, 6], statistical text–based features [7, 3, 6] including density of profane words, and word embeddings [5, 8, 9]. However, two key practical issues have thus far been largely ignored. First, cyberbullying detection solutions must be scalable to the staggering rates at which content is generated (e.g., $350,000$ tweets per minute[2]). Second, the timeliness of cyberbullying detection is critical in developing mitigation strategies [10, 11, 12].

[1]Bullying Statistics: http://www.bullyingstatistics.org/category/bullying-statistics

[2]Twitter Usage Statistics: http://www.internetlivestats.com/twitter-statistics/

In this paper, we formulate cyberbullying detection as a sequential hypothesis testing problem, and propose a novel algorithm designed to reduce the time to raise a cyberbullying alert by minimizing the number of feature evaluations necessary for a decision to be made. We show that the optimal strategy in this decision problem is an optimal stopping rule: our algorithm sequentially reviews features starting from the most informative, and decides when to stop. Once stopped, it can classify a message as cyberbullying or non–cyberbullying based on the features examined thus far. The optimal number of features used is a function of the cost corresponding to the time and effort spent evaluating each feature, and the classification quality. A key property of our solution is that in accomplishing these two goals it does not adversely impact classification quality. We demonstrate the utility, scalability and responsiveness of our proposed solution in a large–scale real–world dataset obtained from the Twitter online social network.

The remainder of this paper is organized as follows. In Section 2, we formulate the problem and define our optimization function. In Section 3, we derive the optimal stopping and classification strategies. In Section 4, we propose a novel algorithm for timely cyberbullying detection in online social networks. Section 5 describes our evaluation methodology and results on a real–world Twitter dataset. We conclude the paper and discuss possible future work in Section 6.

2. PROBLEM FORMULATION

In this Section, we formalize the cyberbullying detection problem. We describe our model and define our optimization function.

2.1. Description

Automatic detection of cyberbullying requires computational approaches that can take advantage of multifaceted attributes, both linguistic and non–verbal. In order to reduce the burden on human experts, we propose a framework that automatically computes the probability of a message to be indicative of harassment with high accuracy while accounting for the effort of the framework in improving its chances of reaching a highly accurate conclusion.

We use a general data representation, applicable to a wide variety of social media platforms as follows. We consider a set of messages $\mathcal{M}$ and a set of users $\mathcal{U}$. Each message $m \in \mathcal{M}$ is sent from user $s \in \mathcal{U}$ to $r \in \mathcal{U}$. Each message is described by a set of feature occurrences $f(m) = \{y_1, y_2, \ldots, y_K\}$, where K is the total number of features. Each feature denotes the existence of some descriptor in the message (e.g., the presence of profanity). An equivalent framing is that the value of feature $y_n \in \{0, 1\}$, where $n = 1, \ldots, K$, is

determined by comparing y_n with empirically determined thresholds that separate bullying from non–bullying messages.

Each message m can belong to one of two classes, i.e., $\mathcal{B}$ for bullying, and $\mathcal{N}$ for non–bullying. We pose the challenge of automatic cuberbullying detection of each message m sent from user s to r as a sequential hypothesis testing problem and use an additive feature score to encode our belief that m is a bullying instance. Specifically, in our hypothesis testing formulation, only two hypotheses exist: (i) $H_{\mathcal{B}}$, which denotes the true hypothesis that m is a bullying message, and (ii) $H_{\mathcal{N}}$, which represents the case where m is a non–bullying message. For each feature y_n, the probability $p(y_n|H_{\mathcal{B}})$ (similarly $p(y_n|H_{\mathcal{N}})$) of the evaluation of the nth feature to produce outcome y_n when the true hypothesis is $H_{\mathcal{B}}$ (similarly when the true hypothesis is $H_{\mathcal{N}}$) is empirically computed. Similarly, the *a priori* probability $P(H_{\mathcal{B}}) = p$ of m being a bullying message is also estimated empirically. The probability of m being a non–bullying message can be computed as $P(H_{\mathcal{N}}) = 1-p$. In our experiments, we use widely popular text and emotion features [13, 14, 7]. Note that, multiple messages can be sent from or to a user, and more than one messages can involve the same pair of users (e.g., user Alice sents 5 messages to Bob, and 1 message to John, and receives 3 messages from Mary).

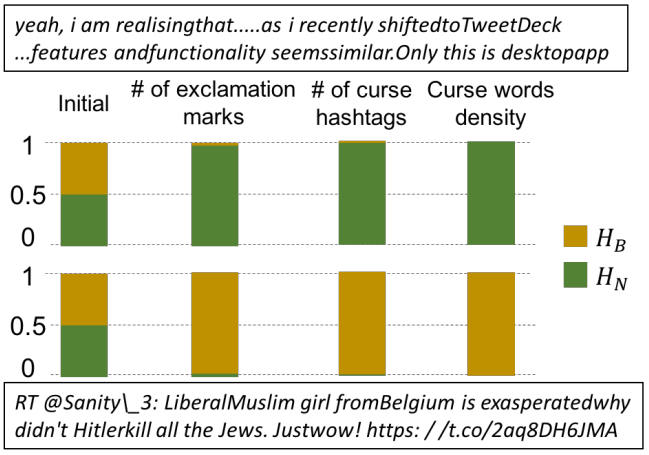

Fig. 1. Posterior probability evolution as more features are extracted and evaluated for a cyberbullying (lower plot) and a non–cyberbullying message (upper plot).

To calculate the bullying belief for m, the framework evaluates features $f(m)$ sequentially as illustrated in Fig. 1. At each step, the framework has to select between stopping and continuing the evaluation process based on the accumulated information thus far and the cost of reviewing additional features. The cost coefficient $c_n > 0$, where $n = 1, \ldots, K$ represents the value of time and effort spent evaluating the nth feature. We also consider misclassification costs $C_{ij} \geqslant 0, i = \mathcal{B}, \mathcal{N}, j = 1, \ldots, L$, where C_{ij} denotes the cost of selecting possibility j when the true hypothesis is H_i, and L denotes the number of decision choices (e.g., bullying or non–bullying). We factor misclassification costs into our approach to quantify the relative importance of detection errors. Note that a model that includes costs may not produce fewer errors than one that does not, and may not rank any higher in terms of overall accuracy, but it is likely to perform better in practical terms because it has a built–in bias in favor of less expensive errors towards one class versus another.

We now formally describe our proposed sequential evaluation process to minimize the number of features used to accurately classify each message m. Specifically, our proposed sequential evaluation process comprises a pair (R, D_R) of random variables. Random variable R takes values in the set $\{0, \ldots, K\}$, and indicates the feature that the framework stops at. Hence it is referred to as *stopping time* in decision theory. Random variable D_R denotes the possibility to select among L possible choices. It depends on R and takes values in the set $\{1, \ldots, L\}$. As an example, consider a case where $L = 3$. In this context, $D_R = 1$ corresponds to "bullying message", $D_R = 2$ denotes "normal message", and $D_R = 3$ indicates "human expert inspection required". Assuming that the random variables y_n are *independent under each hypothesis* $H_i, i = \{\mathcal{B}, \mathcal{N}\}$, the conditional joint probability of $\{y_1, \ldots, y_n\}$ is given as follows:

$$P(y_1, \ldots, y_n|H_i) = \prod_{k=1}^{n} p(y_k|H_i), i = \mathcal{B}, \mathcal{N}. \quad (1)$$

Both the decision to stop at stage n (i.e., the event $\{R = n\}$), and the selection of possibility j (i.e., $D_R = j$) depend only on the accumulated information $\{y_1, \ldots, y_R\}$ by the stopping time R. Equivalently, features that may be examined in the future are not used.

2.2. Optimization Setup

Our goal is to use the least number of features for detection of cyberbullying messages without sacrificing accuracy. To minimize the number of features considered, the stopping time R and the classification rule D_R have to be selected. To this end, we first define the following cost function

$$J(R, D_R) = \mathbb{E}\Big\{ \sum_{n=1}^{R} c_n \Big\} + \sum_{j=1}^{L} \sum_{i=\mathcal{B},\mathcal{N}} C_{ij} P(D_R = j, H_i). \quad (2)$$

The first expression in the cost function regularizes the number of features, whereas the second expression penalizes the average cost of our classification rule. To prove that the optimal strategy is to stop at stage R, we must first show how to obtain the optimum classification rule D_R for any given stopping time R. Once the optimal classification rule has been established, the resulting cost becomes only a function of R, and can thus be optimized with respect to R.

Since D_R depends only on the accumulated information $\{y_1, \ldots, y_R\}$ by stopping time R, the *a posteriori probability* $\pi_n \triangleq P(H_{\mathcal{B}}|y_1, \ldots, y_n)$ must be updated as more features are extracted and evaluated. Lemma 1 shows how to compute π_n iteratively.

Lemma 1 *The posterior probability at stage n where the nth feature is extracted and evaluated to generate outcome y_n is*

$$\pi_n = \frac{p(y_n|H_{\mathcal{B}})\pi_{n-1}}{\pi_{n-1}p(y_n|H_{\mathcal{B}}) + (1-\pi_{n-1})p(y_n|H_{\mathcal{N}})}, \quad (3)$$

where π_{n-1} is the posterior probability at stage $n-1$, and $\pi_0 = p$.

Using Lemma 1 and the fact that $x_R = \sum_{n=0}^{K} x_n \mathbf{1}_{\{R=n\}}$ for any sequence of random variables $\{x_n\}$, where $\mathbf{1}_A$ is the indicator function for event A (*i.e.,* $\mathbf{1}_A = 1$ when A occurs, and $\mathbf{1}_A = 0$ otherwise), the average cost in Eq. (2) can be written compactly as:

$$J(R, D_R) = \mathbb{E}\Big\{ \sum_{n=1}^{R} c_n \Big\} + \mathbb{E}\Big\{ \sum_{j=1}^{L} \big(C_{\mathcal{B}j}\pi_R + C_{\mathcal{N}j}(1-\pi_R)\big)\mathbf{1}_{\{D_R=j\}} \Big\}. \quad (4)$$

3. OPTIMAL STRATEGIES

Here, we solve the optimization problem defined in the Section 2.2 to derive the optimal stopping and classification strategies.

3.1. Classification Strategy

In order to obtain the optimal classification rule D_R for any stopping time R, an independent of D_R lower bound for the second part of Eq. (4) is needed. Since D_R contributes only to this portion of the average cost, the optimal classification rule D_R for a given stopping time R can then be derived. Theorem 2 provides such bound.

Theorem 2 *For any classification rule D_R given stopping time R,*

$$\sum_{j=1}^{L} \big(C_{\mathcal{B}j}\pi_R + C_{\mathcal{N}j}(1-\pi_R)\big)\mathbf{1}_{\{D_R=j\}} \geqslant g(\pi_R), \qquad (5)$$

where $g(\pi_R) \triangleq \min_{1\leqslant j\leqslant L}\big[C_{\mathcal{B}j}\pi_R + C_{\mathcal{N}j}(1-\pi_R)\big]$. The optimal rule is defined as follows:

$$\mathcal{D}_R^{optimal} = \arg\min_{1\leqslant j\leqslant L}\big[C_{\mathcal{B}j}\pi_R + C_{\mathcal{N}j}(1-\pi_R)\big]. \qquad (6)$$

From Theorem 2, we deduce that $J(R, D_R^{optimal}) \leqslant J(R, D_R)$, since the optimal classification rule results to the smallest average cost. Based on the last observation, Eq. (4) can be written as follows:

$$\widetilde{J} \triangleq J(R, D_R^{optimal}) = \min_{D_R} J(R, D_R) = \mathbb{E}\bigg[\sum_{n=1}^{R} c_n + g(\pi_R)\bigg], \qquad (7)$$

which depends only on the stopping time R.

3.2. Stopping Strategy

The solution for optimizing $\widetilde{J}$ with respect to R can be determined by solving the optimization problem

$$\min_{R\geqslant 0} \widetilde{J}(R) = \min_{R\geqslant 0} \mathbb{E}\bigg[\sum_{n=1}^{R} c_n + g(\pi_R)\bigg], \qquad (8)$$

which constitutes a classical problem in optimal stopping theory for Markov processes [15]. Since every stopping time R can take values in $\{0, 1, \ldots, K\}$, the optimum strategy will consist of a maximum of $K+1$ stages. In addition, Bellman's principle of optimality [16] states that the solution we seek must also be optimum, if instead of the first stage we start from any intermediate stage and continue toward the final stage. We derive our optimal stopping strategy as described in Theorem 3 based on the above principle.

Theorem 3 *For $n = K-1, \ldots, 0$, the function $\bar{J}_n(\pi_n)$ is related to $\bar{J}_{n+1}(\pi_{n+1})$ through the equation:*

$$\bar{J}_n(\pi_n) = \min\bigg[g(\pi_n), c_{n+1} + \sum_{y_{n+1}} A_n(y_{n+1})\times \bar{J}_{n+1}\bigg(\frac{p(y_{n+1}|H_{\mathcal{B}})\pi_n}{A_n(y_{n+1})}\bigg)\bigg], \qquad (9)$$

where $A_n(y_{n+1}) \triangleq \pi_n p(y_{n+1}|H_{\mathcal{B}}) + (1-\pi_n)p(y_{n+1}|H_{\mathcal{N}})$ and $\bar{J}_K(\pi_K) = g(\pi_K)$.

The optimal stopping strategy derived by Eq. (9) has a very intuitive structure, *i.e.,* stop at the stage where the cost of stopping is smaller than the cost of continuing. Specifically, at each stage n, our method faces two options given π_n: (i) stop evaluating features and select optimally between the L possibilities, or (ii) continue and evaluate the next feature. The cost of stopping is $g(\pi_n)$, whereas the cost of continuing is $c_{n+1} + \sum_{y_{n+1}} A_n(y_{n+1})\bar{J}_{n+1}\Big(\frac{p(y_{n+1}|H_{\mathcal{B}})\pi_n}{A_n(y_{n+1})}\Big)$.

Finally, we want to emphasize the high scalability of our approach. Indeed, the K functions $\bar{J}_n(\pi_n), n = 0, 1, \ldots, K-1$, are calculated using Eq. (9) by quantizing the interval $[0, 1]$ and computing the corresponding values. This computation relies only on *a priori* information to produce a $K \times d$ matrix, where each row corresponds to the value of the $\bar{J}_n(\cdot)$ function for different values of $\pi_n \in [0, 1]$. This computation needs to be performed only once and can be pre–calculated. Furthermore, probabilities $p(y_n|H_{\mathcal{B}}), p(y_n|H_{\mathcal{N}}), n = 1, \ldots, K, y_n \in \{0, 1\}$, are empirically estimated from training data as follows:

$$\hat{p}(y_n|H_{\mathcal{B}}) = \frac{N(y_n, \mathcal{B})}{\sum_{y'_n} N(y'_n, \mathcal{B})} \text{ and } \hat{p}(y_n|H_{\mathcal{N}}) = \frac{N(y_n, \mathcal{N})}{\sum_{y'_n} N(y'_n, \mathcal{N})}, \qquad (10)$$

where $N(y_n, \mathcal{B})$ and $N(y_n, \mathcal{N})$ denote the number of messages that give rise to outcome y_n after extracting and evaluating the $n$the feature and constitute cyberbullying and non–bullying messages, respectively. We also estimate the *a priori* probabilities as follows:

$$[P(H_{\mathcal{B}}), P(H_{\mathcal{N}})]^T = [p, 1-p]^T = \bigg[\frac{N_{\mathcal{B}}}{N_{\mathcal{B}}+N_{\mathcal{N}}}, \frac{N_{\mathcal{N}}}{N_{\mathcal{B}}+N_{\mathcal{N}}}\bigg]^T, \qquad (11)$$

where $N_{\mathcal{B}}$ and $N_{\mathcal{N}}$ denote the number of messages in the training set that constitute cyberbullying and non–bullying messages, respectively. Hence the complexity of calculating $\bar{J}_n(\pi_n)$ is independent from the actual number of messages, which can be huge (e.g., 350K tweets per minute).

4. AVOID ALGORITHM

In this section, we describe AvOID, a novel algorithm for optimal online cyberbullying detection. AvOID relies on Theorems 2 and 3 to timely and optimally identify cyberbullying content in messages exchanged in online social networks.

In the online cyberbullying detection optimization problem, features about a message m are examined sequentially ordered one after another. The ordering of features is crucial to the computation of the optimum average cost $\bar{J}_0(\pi_0)$. Consider for example the case of two features $f(m) = \{y_1, y_2\}$, where y_1 is the number of bad words, and y_2 is the number of exclamation marks in a message. The number of bad words has been shown to be more informative in discriminating bullying from non–bullying content [1, 13]. Thus, if AvOID was to examine y_2 first, it would be very probable that it would need to evaluate y_1 as well to improve its chance of accurate classification. On the other hand, if y_1 was to be evaluated first, it could be possible for AvOID to reach a decision using one feature only. To avoid the computational complexity of evaluating all $K!$ possible orderings of features, we propose a simple heuristic. Specifically, we sort features in increasing order of $c_n(p(y_n = 0|H_{\mathcal{B}}) + p(y_n = 1|H_{\mathcal{B}}))$. We select this heuristic due to its ability to promote low cost (c_n) features that at the same

Table 1. Features considered in this work.

Type	Features
Text (in total: 8)	# of exclamation marks, # of uppercase letters, # of emoticons, # of acronyms, # of second person pronouns, # of curse hashtags, # of curse words, density of curse words
Emotion (in total: 3)	mean value of valence, arousal and dominance respectively

time result in few errors ($p(y_n = 0|H_{\mathcal{B}}) + p(y_n = 1|H_{\mathcal{B}})$). Our framework can be easily extended to accommodate other heuristics.

Initially, the posterior probability π_0 is set to the prior probability p of a message being an instance of cyberbullying, and the two terms inside the minimization of Eq. (9) are compared. If the first term is less than or equal to the second term, AvOID stops and classifies the message based on the optimal strategy of Eq. (6). Conversely, when the first term is larger than the second term, the first feature is extracted from the message and evaluated; as a result, a comparison outcome y_1 is generated and used to update the posterior probability π_0 to π_1 using the update rule of Eq. (3). Consider for example the case where the first feature denotes the number of bad words in the message. This number is computed and a 0 or 1 is returned to indicate the evaluation outcome. AvOID repeats these steps until either it decides to stop, at which case it classifies the message using $< K$ features, or all features are computed and examined, in which case the message is classified using all K features.

5. NUMERICAL RESULTS

In this section, we provide numerical results to illustrate the performance of AvOID on the detection of cyberbullying messages. We evaluate our algorithm on a real–world Twitter dataset consisting of $10,600$ tweets, half of which constitute harassment tweets. In particular, we construct this dataset by extracting $5,300$ manually labeled harassment tweets from [17] and $5,300$ randomly selected normal tweets from the Twitter corpus of the CAW 2.0 dataset. We focus on $L = 2$ choices (*i.e.,* cyberbullying and non–cyberbullying content) and extract 11 features, shown in Table 1. To extract the number of acronyms, we created a lexicon of offensive acronyms [3], while features curse hashtags and words were extracted based on a curse word lexicon[4]. The mean values of valence, arousal and dominance were extracted based on [18]. We discretize the values of the above features to 0 and 1 based on thresholds empirically determined from their histograms such that 90% of the mass of the observations for each class lies above/below this threshold depending on the class of interest. Five–fold cross validation was performed, and experiments were conducted for different values of $c_n, C_{\mathcal{B}2}$ and $C_{\mathcal{N}1}$.

Fig. 2 illustrates the error probability achieved by AvOID as the average number of features used by the algorithm increases. Results are reported for constant misclassification costs (*i.e.,* $C_{\mathcal{B}1} = C_{\mathcal{N}2} = 0$ and $C_{\mathcal{B}2} = C_{\mathcal{N}1} = 1$) and for varying values of $c_n \in [0, 10]$ when all features have the same cost (*i.e.,* $c_n = c$). The error probability achieved by a non–sequential hypothesis testing method that uses all 11 features is also included for comparison. As expected, when the average number of features used is small, AvOID exhibits large error probability. However, as this number increases, performance improves dramatically. In fact, AvOID attains approximately the same error probability as the non–sequential hypothesis testing method using only $R = 4$ features. This corresponds to a 64% reduction on average in the number of features used without sacrificing accuracy. Different values of costs c_n and misclassification costs $C_{\mathcal{B}2}$ and $C_{\mathcal{N}1}$ result in different error probability values, while trading–off false alarm and misdetection probabilities. The inset in Fig. 2 shows the number of features used by AvOID to classify messages for each tweet in the testing dataset when $R = 4.02$ features. In most cases, 3 to 4 features are enough to reach a classification decision.

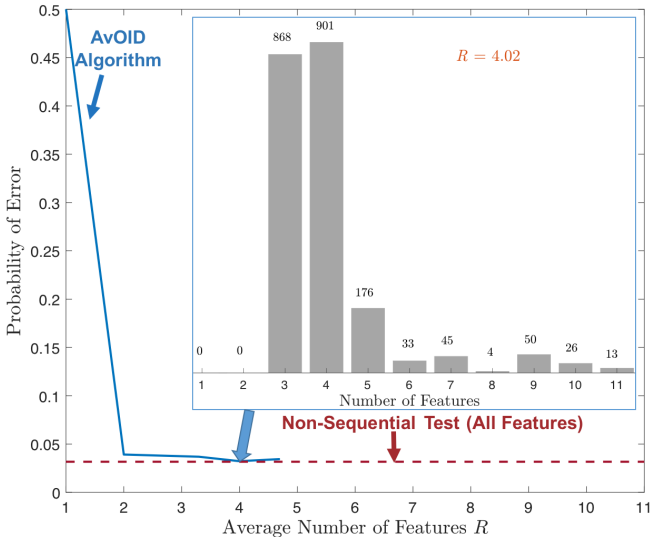

Fig. 2. Probability of error as a function of the expected number of features. Inset shows the distribution of number of features used by AvOID to classify messages for an average of ~ 4 features.

6. CONCLUSIONS AND FUTURE WORK

In this paper, a sequential hypothesis testing formulation was proposed to address the problem of cyberbullying detection. More specifically, each message can belong in one of two classes (*i.e.,* cyberbullying or normal) and the goal is to decide when to stop extracting and evaluating features from the message and declare a decision. To this end, an optimization function was defined in terms of the cost of features and the average cost of the classification strategy and the optimal solution was determined. Our proposed algorithm implements this optimal solution and achieves a 64% reduction on the average number of features used to reach a classification decision without sacrificing accuracy. In our ongoing work, we focus on text and emotion features of messages, as such features have been shown to being informative for cyberbullying classification. In future work, we plan to extend our framework so as to exploit user, network and content information as well.

7. REFERENCES

[1] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana, "Cybercrime detection in online communica-

[3] https://www.noslang.com/dictionary/
[4] http://www.cs.cmu.edu/biglou/resources/bad-words.txt

tions: The experimental case of cyberbullying detection in the Twitter network," *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016.

[2] Karthik Dinakar, Emily Weinstein, Henry Lieberman, and Robert Louis Selman, "Stacked generalization learning to analyze teenage distress.," in *ICWSM*, 2014.

[3] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra, "Detection of cyberbullying incidents on the instagram social network," *arXiv preprint arXiv:1503.03909*, 2015.

[4] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo 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*. ACM, 2015, pp. 617–622.

[5] Mifta Sintaha, Shahed Bin Satter, Niamat Zawad, Chaity Swarnaker, and Ahanaf Hassan, *Cyberbullying detection using sentiment analysis in social media*, Ph.D. thesis, BRAC University, 2016.

[6] Vivek K Singh, Qianjia Huang, and Pradeep K Atrey, "Cyberbullying detection using probabilistic socio-textual information fusion," in *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 2016, pp. 884–887.

[7] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong, "Improving cyberbullying detection with user context.," in *ECIR*. Springer, 2013, pp. 693–696.

[8] Xiang Zhang, Jonathan Tong, Nishant Vishwamitra, Elizabeth Whittaker, Joseph P Mazer, Robin Kowalski, Hongxin Hu, Feng Luo, Jamie Macbeth, and Edward Dillon, "Cyberbullying detection with a pronunciation based convolutional neural network," in *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*. IEEE, 2016, pp. 740–745.

[9] Rui Zhao and Kezhi Mao, "Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 328–339, 2017.

[10] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 3, pp. 18, 2012.

[11] Zahra Ashktorab, Srijan Kumar, Soham De, and Jennifer Golbeck, "ianon: Leveraging social network big data to mitigate behavioral symptoms of cyberbullying," *iConference 2014 (Social Media Expo)*, 2014.

[12] Zahra Ashktorab, "A study of cyberbullying detection and mitigation on instagram," in *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. ACM, 2016, pp. 126–130.

[13] Kelly Reynolds, April Kontostathis, and Lynne Edwards, "Using machine learning to detect cyberbullying," in *10th International IEEE Conference on Machine Learning and Applications and Workshops (ICMLA)*, 2011, vol. 2, pp. 241–244.

[14] Karthik Dinakar, Roi Reichart, and Henry Lieberman, "Modeling the detection of Textual Cyberbullying," *The Social Mobile Web*, vol. 11, no. 2, 2011.

[15] Albert N Shiryaev, *Optimal Stopping Rules*, vol. 8, Springer Science & Business Media, 2007.

[16] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1, Athena Scientific, 2005.

[17] Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, et al., "A Large Labeled Corpus for Online Harassment Research," in *ACM on Web Science Conference*, 2017, pp. 229–233.

[18] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert, "Norms of Valence, Arousal, and Dominance for 13,915 English lemmas," *Behavior Research Methods*, vol. 45, no. 4, pp. 1191–1207, 2013.