Robust Detection of Cyberbullying in Social Media

Mengfan Yao
University at Albany
Department of Computer Science
myao@albany.edu

ABSTRACT

The potentially detrimental effects of cyberbullying have led to the development of numerous automated, data–driven approaches, with an emphasis on classification accuracy. Cyberbullying, as a form of abusive online behavior, although not well–defined, is a repetitive process, i.e., a sequence of aggressive messages sent from a bully to a victim over a period of time with the intent to harm the victim. Existing work has focused on aggression (i.e., using profanity to classify toxic comments independently) as an indicator of cyberbullying, disregarding the repetitive nature of this harassing process. However, raising a cyberbullying alert immediately after an aggressive comment is detected can lead to a high number of false positives. At the same time, three key practical challenges remain unaddressed: (i) detection timeliness, which is necessary to support victims as early as possible, (ii) scalability to the staggering rates at which content is generated in online social networks, (iii) reliance on high quality annotations from human experts for training of highly accurate supervised classifiers.

To overcome the challenges associated with cyberbullying detection in online social networks, my PhD thesis focuses on a novel formulation of the online classification problem as sequential hypothesis testing that seeks to drastically reduce the number of features used while maintaining high classification accuracy. To reduce the dependency on labeled datasets, I seek to develop efficient semisupervised methods that extrapolate from a small seed set of expert annotations. Preliminary results are very encouraging, showing significant improvements over the state–of–the–art.

CCS CONCEPTS

• **Information systems** → *Collaborative and social computing systems and tools*; *Social networking sites*; *Data mining*; *World Wide Web*; *Social networks*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Classification; cyberharassment; optimization; sequential selection; social networks

ACM Reference Format:
Mengfan Yao. 2019. Robust Detection of Cyberbullying in Social Media. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3314196

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314196

1 PROBLEM STATEMENT

The potentially devastating real–world consequences to victims have resulted in numerous cyberharassmet classification methods with a focus on detection accuracy (e.g. [6]. While high accuracy is undoubted, the state–of–the–art relies on a **fixed** set of features learned during training for **offline** detection (i.e., after all correspondence has become available), hindering the ability to respond in a **timely** manner (i.e., as soon as possible) to cyberbullying events. Moreover, the **scalability** of existing methods to the staggering rates at which content is generated (e.g., 95 million photos and videos are shared on Instagram per day[1]) remains unaddressed [8].

To achieve timely, scalable, and accurate detection of cyberbullying, in my PhD thesis, I would like to address the following 3 research **problems**: (RQ1) Can the time to detection be reduced given that cyberbullying is a repetitive process that unfolds over time? (RQ2) Is it possible to maintain high classification accuracy (i.e., minimize false negatives) while solving for problem 1? (RQ3) Can novel unsupervised learning methods be devised to achieve the same goal?

2 STATE OF THE ART

State-of-the-art mainly comprises 3 bodies of research: (i) cyberbullying detection, (ii) online streaming feature selection (OSFS) methods and (iii) online learning algorithms for classification.

Cyberbullying Detection. Although many approaches have been proposed recently to detect cyberbullying (e.g. [14]), all existing works focus on classification accuracy, neglecting the equally important aspects of scalability and timeliness [8]. With the exception of an incremental cyberbullying detection method to improve detection responsiveness [8], no prior or related work has studied cyberbullying as a repetitive process.

Online Classification Methods. Machine learning methods are usually trained offline and deployed without further updates once training is complete. Once deployed, the accuracy of offline–trained models often deteriorates with time, whereas their retraining may be prohibitive if data is large and/or evolving. At the same time, online learning algorithms that examine data points one at a time, updating their model parameters as new samples arrive, are facing scalability constraints [5]. Although related, my framework fundamentally differs from online classification methods in that the belief on a classification outcome is updated at each time step, allowing detection in real–time.

Online Streaming Feature Selection Methods. Recently, online streaming feature selection (OSFS) methods have been proposed to handle streaming features (e.g.[11]). In contrast to conventional feature selection methods, which are often employed to identify a

[1] 33 Mind–Boggling Instagram Stats & Facts for 2018: https://www.wordstream.com/blog/ws/2017/04/20/instagram-statistics

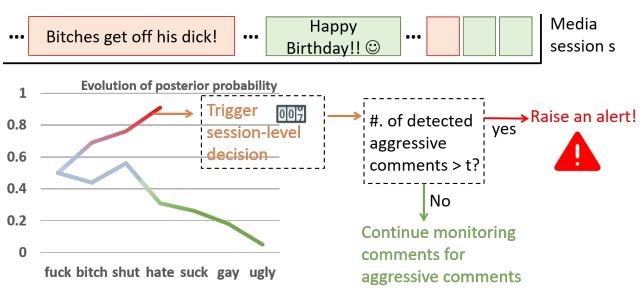

Figure 1: Overview of my proposed approach. Given a media sessions s, and an alert threshold t, my approach examines comments as they become available over time and raises an alert only after the number of detected aggressive comments surpasses the threshold. The posterior probability evolution of an aggressive (red) and non–aggressive (green) comment as more features are examined is provided for illustration purposes. Notice that the number of features used to make a decision in each case differs.

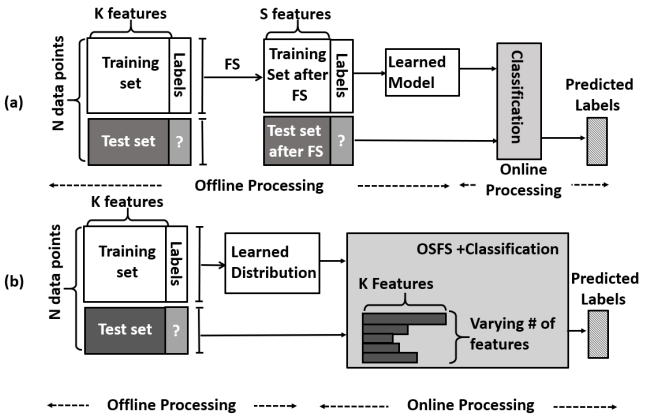

Figure 2: Given a $N \times K$ data matrix and column vector Y of labels, feature selection (FS) is applied to select $S << K$ features before training and classification in (a) existing classification methods. Compare this with (b) my framework, where a model is trained offline once on all features, and online streaming feature selection (OSFS) and classification are performed simultaneously, resulting in a variable number of features used for each data point.

small subset of imformative features to improve classification accuracy but require all features to be available upfront, OSFS methods assume that the size of training data is fixed, and strive to maintain a feature subset that is sequentially updated with the arrival of each new feature based on relevancy and redundancy [11, 15] and group membership [7, 13].

3 PROPOSED APPROACH

In my completed work, I have formulated cyberbullying detection on Instagram media sessions as a two–stage, online framework (illustrated in Figure 1) that (i) sequentially examines comments as they become available over time, and (ii) minimizes the number of feature evaluations necessary for a decision to be made for each comment. This work has been accepted as a short paper to be included in the proceedings of TheWebConf'19 [12].

In my formulation, I use a general data representation, applicable to a wide variety of social media platforms as follows. Each media session $s \in \mathcal{M}$ belongs to a user $u \in \mathcal{U}$, has an associated media object (i.e., image or video) along with its corresponding caption and hashtags, and a set of comments $\{m_1, \ldots, m_{N_s}\}$ from users in $\mathcal{U}$, where N_s denotes the number of comments in s.

In the first stage, comments in s are examined and classified sequentially as aggressive or non–aggressive. Whenever an aggressive comment is detected, a comment–level decision is made. One novelty aspect of my proposed approach is that in this stage, when a given comment is being examined, unlike any existing feature selection method that outputs a global subset of discriminatory features to be used for classification, my approach conducts online streaming feature selection and comment–level classification simultaneously, resulting in a potentially different number of features used to classification and a potential higher classification accuracy. Figure 2 illustrates this difference between existing classification methods (Figure 2. a) and my framework (Figure 2. b).

In the second stage, before a new comment is examined, the number of comment–level detections is compared to a predefined threshold t, and the media session is classified as cyberbullying when the threshold is exceeded. This is another novelty of my proposed approach – comment level decisions are combined to reduce the number of false alarms.

In order to address RQ2, my plan is to introduce a cost into the optimization function to account for the number of comments examined. Such addition will enable my framework to optimally stop examining comments when a decision can be "safely" reached based on accumulated knowledge. As a result, false negative is expected to be largely reduced. Because of the inherent difficulty of RQ3, currently, the work for this task is in an early stage. Specifically, the problem has been formulated, and the first related works have been identified. Possible solutions are yet to be determined.

4 METHODOLOGY

In this section, I introduce the methodology of my proposed optimization framework.

4.1 Problem Formulation

In my hypothesis testing formulation, only two hypotheses exist: (i) $H_{\mathcal{B}}$, which denotes the true hypothesis that m is an aggressive comment, and (ii) $H_{\mathcal{N}}$, which represents the case where m is a non–aggressive comment. Each comment m is described by a vector of features $f(m) = \{y_1, y_2, \ldots, y_K\}$, where K is the total number of features, and $y_k \in \mathcal{Y}$. 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 observe value 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 an aggressive comment is also estimated empirically. The probability of m being a non–aggressive comment can be computed as $P(H_{\mathcal{N}}) = 1 - p$.

To calculate the belief for m, the framework evaluates features sequentially as illustrated in Figure 1. When examining each comment, 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. I additionally consider misclassification costs $C_{ij} \geq 0, i = \mathcal{B}, \mathcal{N}, j = 1, \ldots, L$, 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., aggressive or non–aggressive). I factor misclassification costs into my approach to quantify the relative importance of detection errors.

I now formally describe my proposed sequential evaluation process to minimize the number of features used to accurately classify each comment m. Specifically, my 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, and 0 indicates that no features were evaluated. 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 "aggressive comment", $D_R = 2$ denotes "non–aggressive comment", 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 $P(y_1, \ldots, y_n|H_i) = \prod_{k=1}^{n} p(y_k|H_i), i = \mathcal{B}, \mathcal{N}$. 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.

4.2 Optimization Setup

My goal is to achieve high accuracy using the least number of features for detecting aggression at the comment–level. To minimize the number of features considered, the stopping time R and the classification rule D_R have to be selected. To this end, I first define the following cost function:

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

The first expression in the cost function regularizes the number of features, whereas the second expression, commonly referred to as Bayes Risk, penalizes the average cost of the classification rule. My goal can be interpreted as finding the minimum average cost with respect to both random variables R and D_R, i.e., $\min_{R, D_R} J(R, D_R)$, to derive the optimal stopping and classification rules. Intuitively, the optimal rule is to stop at corresponding stopping time R using optimum classification rule D_R.

4.3 Classification Rule

If denote *a posteriori probability* as $\pi_n \triangleq P(H_{\mathcal{B}}|y_1, \ldots, y_R)$, it is not difficult to see that π_n can be iteratively computed as

$$\pi_n = \frac{p(y_n|H_{\mathcal{B}})\pi_{n-1}}{\pi_{n-1}p(y_n|H_{\mathcal{B}}) + (1 - \pi_{n-1})p(y_n|H_{\mathcal{N}})}, \quad (2)$$

where π_{n-1} denotes the posterior probability of a comment being a cyberbullying comment given the first $n-1$ features, and $\pi_0 = p$, i.e., *a priori* probability. Consequently, Eq. (1) can be written compactly as:

$$J(R, D_R) = \mathbb{E}\left\{ \sum_{n=1}^{R} c_n \right\} + \mathbb{E}\left\{ \sum_{j=1}^{L} (C_{Bj}\pi_R + C_{Nj}(1 - \pi_R)) \mathbb{1}_{\{D_R = j\}} \right\}. \quad (3)$$

In order to obtain the optimal classification rule D_R for any stopping time R (i.e., find the optimum Bayes test given that the values of the first R features $y_1, \ldots, y_R$ are observed), an independent of D_R lower bound for the second part of Eq. (3) 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. Specifically, for any classification rule D_R given stopping time R, $\sum_{j=1}^{L} (C_{Bj}\pi_R + C_{Nj}(1 - \pi_R)) \mathbb{1}_{\{D_R = j\}} \geq g(\pi_R)$, where $g(\pi_R) \triangleq \min_{1 \leqslant j \leqslant L} \left[C_{\mathcal{B}j}\pi_R + C_{\mathcal{N}j}(1 - \pi_R) \right]$. The optimal rule is thus defined as follows:

$$\mathcal{D}_R^{optimal} = \arg\min_{1 \leqslant j \leqslant L} \left[C_{\mathcal{B}j}\pi_R + C_{\mathcal{N}j}(1 - \pi_R) \right]. \quad (4)$$

From Eq. (4), it follows that $J(R, \mathcal{D}_R^{optimal}) \leq J(R, D_R)$ since the optimal classification rule results to the smallest average cost. Based on the last observation, Eq. (3) can be written as:

$$\min \tilde{J} \triangleq \min_{D_R} J(R, D_R) = \min \mathbb{E}\left\{ \sum_{n=1}^{R} c_n + g(\pi_R) \right\}, \quad (5)$$

which depends only on the stopping time R.

4.4 Stopping Rule

The solution for optimizing $\tilde{J}$ in Eq. (5) with respect to R can be determined by solving the optimization problem $\min_{R \geq 0} \tilde{J}_R = \min_{R \geq 0} \mathbb{E}\left\{ \sum_{n=1}^{R} c_n + g(\pi_R) \right\}$. This constitutes a classical problem in optimal stopping theory for Markov processes [9]. Since the 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 [1] 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 the optimal stopping rule, $R^{optimal}$, as follows:

$$R^{optimal} = \min \left\{ 0 \leq n \leq K | S_n = \tilde{J}_n \right\}, \quad (6)$$

where for $n = K - 1, \ldots, 0$, the n–th stage cost $S_n(\pi_n)$ is related to $S_{n+1}(\pi_{n+1})$ through the following recursion:

$$S_n(\pi_n) = \min \left[g(\pi_n), c_n + \int A_n(y_{n+1}) \times S_{n+1}\left(\frac{p(y_{n+1}|H_{\mathcal{B}})\pi_n}{A_n(y_{n+1})} \right) dy_{n+1} \right], \quad (7)$$

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

The optimal stopping rule derived by Eq. (7) has a very intuitive structure. *i.e.,* stop at the stage where the cost of stopping (the first expression in the minimization) is no greater than the expected cost of continuing given all information accumulated at the current stage (the second expression in the minimization). Specifically, at each stage n, my method faces two options given π_n: (i) stop evaluating features and classify, i.e. compute D_n, or (ii) continue and evaluate the next feature. Note that $\tilde{J}_n$ can be interpreted as the cost if the system stops at stage n, whereas S_n is the optimal cost when the system is at stage n (regardless of continuing or stopping). From Eq. (7), it is easy to verify that $S_n \geq \tilde{J}_n$ for all $n = 0, \ldots, K$. Thus, the optimal time to stop is when the equality is first obtained.

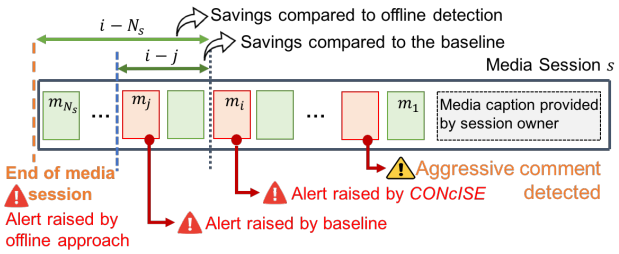

Figure 3: Illustration of timeliness, i.e., the number of comments "saved" as compared to (i) an offline approach that has to wait until all comments related to a media session become available, or (ii) an online baseline that raises an alert after a threshold.

5 RESULTS

In this section I provide experimental analysis to evaluate my proposed approach (termed CONcISE). For a fair comparison with the state–of–the–art, I use the Instagram data set collected by Hosseinmardi et al. [6]. In total, the dataset contains 3,829,756 users and 9,828,760 comments. Of all media sessions which contain at least 40% profanities. I augmented this dataset with $\sim 10K$ comment–level labels obtained from 10 experts. The comments span in total 22.1% of the media sessions with 40% or more profanity. I use this small subset of labeled comments for training, and the remaining unlabeled comments (i.e., 77.9% of labeled media sessions in the original datasetf from [6]) for testing.

As for feature representations, I consider the 10 unigrams [6], the 1,384 profane unigrams and bigrams [10] and 394 offensive unigrams and bigrams [4] that are shown to be imformative in state-of-the-art approaches [6, 8], to examine the robustness of *CONcISE* on the dictionary used for detection. This results in three variants of *CONcISE*: **CONcISE–10**, **CONcISE–Profane**, and **CONcISE–Noswearing**, respectively, that use the corresponding dictionaries.)

Comparison with Cyberbullying Detection State-of-the-art. Table 1 summarizes the performance of *CONcISE* as compared to the state–of–the–art for cyberbullying detection, in terms of accuracy, recall and precision of the bullying class, area under the ROC curve (AUC), the average number (and standard deviation) of features used by my approach in detecting aggressive comments (similarly for the number of comments needed to detect a cyberbullying session). The best performing method is marked in bold. All in all, *CONcISE* and its variations outperform the baselines almost invariably across the board of evaluation metrics. Clearly, all three variants of *CONcISE* outperform the baselines, often by a considerable margin. At the same time, *CONcISE* significantly reduces both the average number of features used to classify individual comments as well as the number of comments examined before determining if a session is an instance of cyberbullying.

Comparison with Online Streaming Feature Selection State-of-the-art. I compare *CONcISE* to the best performing OFSF methods in terms of feature selection quality, namely SAOLA [13] and OFS–Density [15], over the feature space defined by the corresponding disctionary used, i.e., 10 unigrams, Noswearing and Profane dictionaries respectively. As SAOLA and OFS-Density do not perform classification in conjuction to feature selection, I use KNN, SVM, Random Forest and Standard Bayes for both baselines; KNN and SVM have previously been shown to achieve the highest classification performance in [13, 15]. I report the accuracy of *CONcISE*, as well as all combinations of baselines and classifiers. Figure 4 shows that *CONcISE* consistently outperforms OFS-Density and SAOLA over all 3 feature spaces.

Timeliness Analysis. To compare the best performing variance of *CONcISE*, namely, *CONcISE–Noswearing*, against baselines with respect to *timeliness*, I introduce a measure "# of saved comments", defined as $i - j$ where i and j, respectively, denotes the number of comments that *CONcISE* and a given baseline have to inspect before raising an alert. My definition of timeliness along with the above mentioned scenarios are illustrated in Figure 3. The result has shown that *CONcISE–Noswearing* can detect cyberbullying media sessions faster than the best performing offline approach (i.e. RDFS), for all thresholds considered (in addition to being more accurate). As comparing to the best performing online approach (i.e. RF–comment), the result suggests that RF–comment achieves slightly faster detection than *CONcISE–Noswearing*. However, with less than 70% precision (see Table 1), RF–comment tends to classify non–aggressive comments as aggressive. In turn, RF–comment reaches the threshold faster than *CONcISE–Noswearing* in true cyberbullying sessions. The downside to this outcome is a higher number of false positives in sessions that are not in reality cyberbullying sessions.

Scalability Analysis. I evaluate the scalability of *CONcISE* in both classifying comments and in raising a session–level alert. In my non optimized implementation, *CONcISE-10*, *CONcISE-Profane* and *CONcISE-Noswearing* classify on average 45,140 comments, 33,964 comments, and 30,849 comments per second, accordingly. With respect to runtime for session–level decisions, I found baselines RDFS, RF, and TM to be the fastest among all methods, for an average of 0.0021, 0.0011, and 0.002 seconds respectively in making a session level decision. However, such "speed" is meaningless if considered in isolation to timeliness. Both RDFS, RF, and TM require all comments of a media session to be available for classification, making them offline for all practical purposes. The rest of the methods achieve similar runtime (s), ranging from 0.0077 to 0.0153 second per media session, with *CONcISE–10* and RF–comment being the fastest and the slowest, respectively. Even though the difference may seem negligible for this dataset, considering a real–world

Table 1: Performance comparison of *CONcISE* with state–of–the–art cyberbullying detection methods.

Method	Accuracy	Recall	Precision	AUC	Avg. # of features (std.)	Avg. # of comments (std.)
RDFS [6]	0.751	0.449	**0.858**	**0.877**	10 (0)	83
RF [2]	0.791	0.703	0.749	0.862	13 (0)	83
RDFS–comment [6]	0.733	0.463	0.760	0.684	10 (0)	56.15 (36.64)
RF–comment [2]	0.776	0.751	0.699	0.724	13 (0)	33.07 (22.59)
DLR [8]	0.497	0.651	0.616	0.521	4 (0)	35.90 (45.43)
TM [3]	0.749	0.783	0.649	0.816	115(0)	83(0)
CONcISE–10	0.783	**0.794**	0.695	0.864	**2.97 (1.17)**	**26.68 (16.92)**
CONcISE–Profane	**0.806**	0.769	0.745	0.860	3.76 (3.19)	30.62 (21.82)
CONcISE–Noswearing	**0.806**	0.776	0.742	0.862	3.92 (2.16)	29.75 (21.09)

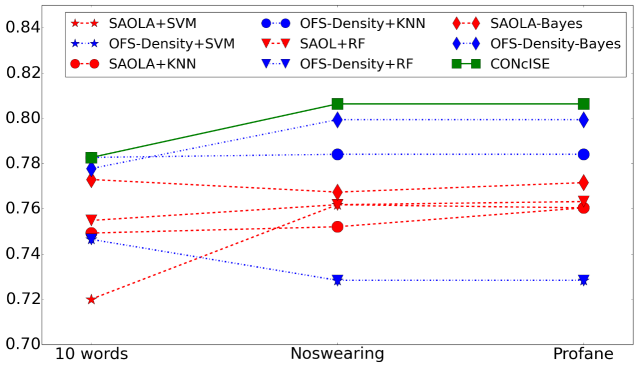

Figure 4: Accuracy comparison of *CONcISE*, and SAOLA and OFS-Density.

scenario where a million media sessions are to be evaluated in real–time. By extrapolating the results of my analysis to such scenario, we can expect *CONcISE–10* to classify all sessions within $\sim 7,600$ seconds (i.e., ~ 2 hours) less than RF–comment.

Finally, note that even though such numbers refer to a sequential implementation of *CONcISE*, my approach can be trivially parallelized by classifying each comment/session in parallel. This empirical result, in addition to the fact that the runtime of *CONcISE* is linearly proportional to the number of features it uses to reach a classification decision, provide strong evidence about the real–world applicability of *CONcISE*.

6 ACKNOWLEDGMENTS

I would like to thank my advisor, Dr.Charalampos Chelmis, for all his help and guidance that he has provided me over the past two years.

7 CONCLUSION AND FUTURE WORK

The main goal of my PhD study is to introduce socio–technical advances that enhance Web platforms. Specifically, I strive to achieve timely and accurate detection of cyberbullying instances in online social networks while being scalable to the staggering rates at which content is generated in online communities. The main research problems of my Ph.D. are (1) Timely and Scalable detection (2) Accurate detection with minimal false negative (3) Unsupervised learning method that achieves the same goals as in (1) and (2).

My preliminary experiments showed the feasibility of the proposed framework, with up to 62.17% of improvement in accuracy and 67.86% reduction in terms of timeliness comparing to baseline approaches.

In future work, I plan to explore features such as user– and network–information and the sequence of conversations to further improve classification accuracy. I am also planning to evaluate the performance of my approach on additional datasets from diverse platforms including Ask.fm and Twitter, which are reported to be key social networking venues where users frequently become victims of cyberbullying. Given the broad definition of cyberbullying, I may also design detection strategies grounded in a more nuanced, multi–dimensional representation of repetitive harassment instead of striving for a global and/or simplified indicator of cyberbullying. The most interesting part of the work will be in finding out how to make use of unsupervised methods, since both cost function and the learning process can be inherently different from the proposed sequential framework which relies on training labels.

REFERENCES

[1] D. P. Bertsekas. 2005. *Dynamic Programming and Optimal Control*. Vol. 1. Athena Scientific.
[2] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 13–22.
[3] Vivek Singh Devin Soni. [n. d.]. Time Reveals AllWounds: Modeling Temporal Dynamics of Cyberbullying Sessions. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*.
[4] AllSlang family. [n. d.]. Internet Slang Swear Word List & Curse Filter. https://www.noswearing.com/dictionary.
[5] Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. 2018. Online Learning: A Comprehensive Survey. *arXiv preprint arXiv:1802.02871* (2018).
[6] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Prediction of cyberbullying incidents in a media-based social network. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 186–192.
[7] Haiguang Li, Xindong Wu, Zhao Li, and Wei Ding. 2013. Group feature selection with streaming features. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 1109–1114.
[8] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, and Shivakant Mishra. 2018. Scalable and Timely Detection of Cyberbullying in Online Social Networks. (2018).
[9] Albert N Shiryaev. 2007. *Optimal Stopping Rules*. Vol. 8. Springer Science & Business Media.
[10] Luis von Ahn. [n. d.]. Offensive/Profane Word List. https://www.cs.cmu.edu/~biglou/resources/bad-words.txt.
[11] Xindong Wu, Kui Yu, Hao Wang, and Wei Ding. 2010. Online streaming feature selection. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. Citeseer, 1159–1166.
[12] Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. 2019. Cyberbullying Ends Here: Towards Robust Detection of Cyberbullying in Social Media. In *2019 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.
[13] Kui Yu, Xindong Wu, Wei Ding, and Jian Pei. 2016. Scalable and accurate online feature selection for big data. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11, 2 (2016), 16.

[14] Rui Zhao and Kezhi Mao. 2017. Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder. *IEEE Transactions on Affective Computing* 8, 3 (2017), 328–339.

[15] Peng Zhou, Xuegang Hu, Peipei Li, and Xindong Wu. 2019. OFS-Density: A novel online streaming feature selection method. *Pattern Recognition* 86 (2019), 48–61.