Cyberbullying Ends Here: Towards Robust Detection of Cyberbullying in Social Media

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

Charalampos Chelmis
University at Albany, SUNY
Department of Computer Science
cchelmis@albany.edu

Daphney–Stavroula Zois
University at Albany, SUNY
Electrical and Computer Engineering Department
dzois@albany.edu

ABSTRACT

The potentially detrimental effects of cyberbullying have led to the development of numerous automated, data–driven approaches, with 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 harassment (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, two key practical challenges remain unaddressed: (i) detection timeliness, which is necessary to support victims as early as possible, and (ii) scalability to the staggering rates at which content is generated in online social networks.

In this work, we introduce *CONcISE*, a novel approach for timely and accurate Cyberbullying detectiON on Instagram media SEssions. We propose a sequential hypothesis testing formulation that seeks to drastically reduce the number of features used in classifying each comment while maintaining high classification accuracy. *CONcISE* raises an alert only after a certain number of detections have been made. Extensive experiments on a real–world Instagram dataset with $\sim 4M$ users and $\sim 10M$ comments demonstrate the effectiveness, scalability, and timeliness of our approach and its benefits over existing methods.

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

ACM Reference Format:
Mengfan Yao, Charalampos Chelmis, and Daphney–Stavroula Zois. 2019. Cyberbullying Ends Here: Towards Robust Detection of Cyberbullying in Social Media. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313462

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

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313462

1 INTRODUCTION

Cyberbullying, a type of cyberharassment, can take many forms [34], typically however, refers to **repetitive hostile behavior** using digital media (e.g., hurtful comments, videos and images) in an effort to intentionally and repeatedly harass or harm individuals [1, 20]. Cyberbullying is permanent (i.e., content remains accessible online unless removed) and potentially widespread[1] (i.e., online social media provide a wide audience, and quick spread of online posts) [10].

The potentially devastating real–world consequences to victims (including learning difficulties, psychological suffering and isolation, escalated physical confrontations, suicide) [12, 13] have resulted in numerous cyberharassmet classification methods with a focus on detection accuracy [5, 16, 27–29, 33, 43, 44]. 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[2]) has largely remained unaddressed.

Present Work. We propose a two–stage **online** approach designed to reduce the time to raise a cyberbullying alert by (i) sequentially examining comments as they become available over time, and (ii) minimizing the number of feature evaluations necessary for a decision to be made for each comment. We formalize the problem as a sequential hypothesis testing problem, and propose a novel algorithm that satisfies four **key properties**: *accuracy*, *repetitiveness*, *timeliness*, and *efficiency*. We focus on hateful comments, captions and hashtags on Instagram, the online social media platform with the highest percentage of users reporting experiencing cyberbullying [10]. Instagram has more than 800 million registered users as of Sep. 2017, and over 40 billion uploaded photos as of Oct. 2015. To ensure the **reproducibility** of our work, we make the source code of our approach available at https://github.com/IDIASLab/CONcISE.

Outline. The rest of this paper is organized as follows. We first review prior and related work in Section 2. We then formulate the problem of cyberbullying detection in online social networks, define

[1] stopbullying.gov Facts About Bullying: https://www.stopbullying.gov/media/facts/index.html#stats

[2] 33 Mind–Boggling Instagram Stats & Facts for 2018: https://www.wordstream.com/blog/ws/2017/04/20/instagram-statistics

our optimization function, and derive the optimal stopping rule and classification strategy in Section 3. We describe our evaluation methodology and results on a real–world Instagram dataset in Section 4. We conclude with a discussion of our results, limitations, and possible future directions in Section 5.

2 RELATED WORK

In this section, we briefly summarize prior and related work, which mainly constitutes 3 main bodies of research: (i) state–of–the–art cyberbullying detection, (ii) online streaming feature selection (OSFS) methods, and (iii) online learning algorithms for classification.

Cyberbullying Detection. An excellent review of existing cyberbullying detection methods is provided in [1]. More recent approaches include [5, 24, 27, 29, 44]. Meanwhile, only a few approaches have been proposed for cyberbullying detection on Instagram [7, 16, 27, 29, 30, 39, 45]. However, all existing works focus on classification accuracy, neglecting the equally important aspects of scalability and timeliness. With the exception of an incremental cyberbullying detection method to improve detection responsiveness [27], and our proposed approach, 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 however, deteriorates with time, whereas their retraining may be prohibitive if data is large and/or evolving. At the same time, online learning algorithms [4, 9, 19, 36, 41, 42] that examine data points one at a time, updating their model parameters as new samples arrive, are facing scalability constraints [14]. Nevertheless, most existing methods assume high data quality, a property that is often violated in streaming data, in which noise and missing values are the norm [14]. Although related, our 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. Feature selection methods are often employed to identify a small subset of informative features that can globally describe the data well so as to reduce training time, memory requirements and ultimately improve classification accuracy [11, 17, 22, 23, 25, 31]. In contrast to offline feature selection methods, which require all features to be available upfront, online streaming feature selection (OSFS) methods [21, 38, 40, 46] fix the size of the training data, and strive to maintain a feature subset that is sequentially updated with the arrival of each new feature based on relevancy and redundancy [38, 46] and group membership [21, 40]. Unlike OSFS methods, online feature selection methods (e.g., [15, 18, 26, 37]) may update the subset of features with the arrival of new data points (as opposed to new features) while assuming that the full feature space remains unchanged.

To the best of our knowledge, all existing feature selection methods, both off– and on–line, and batch or streaming, output a global subset of discriminatory features to be used for classification (i.e. the same set of features is used after selection for classifying all data points). This is in stark contrast to our approach that conducts online streaming feature selection and classification simultaneously for each new data point, resulting in a potentially different number of features used to classify each data point. Figure 1 illustrates this difference between existing classification methods (Figure 1.a) and our framework (Figure 1.b).

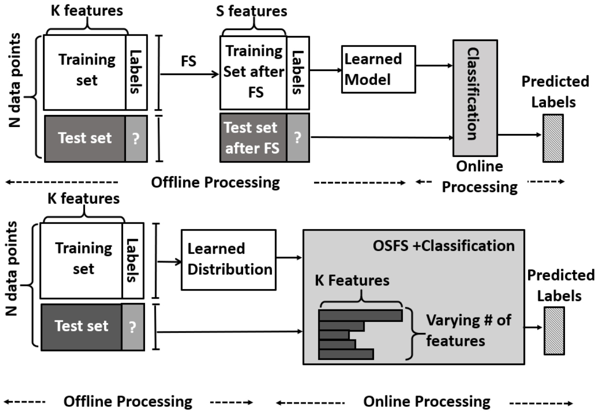


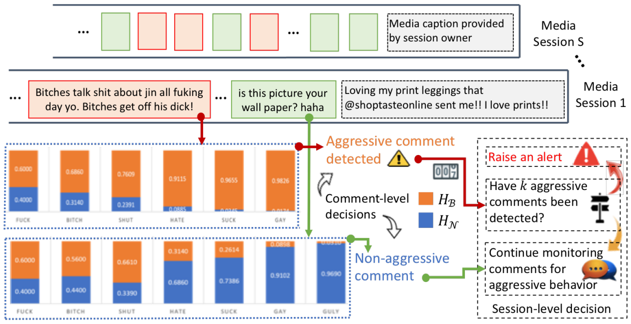
Figure 1: Given a $N \times K$ data matrix and column vector Y of labels, feature selection (FS) is applied to select $S << K$ features before training and classification in (a) existing classification methods. Compare this with (b) our framework, where a model is trained offline once on all features, and online streaming feature selection (OSFS) and classification are performed simultaneously, resulting in a variable number of features used for each data point.

3 PROPOSED METHOD

We 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. Our main idea is to formulate cyberbullying detection on Instagram media sessions as a two–stage, online framework with the following desired properties:

- **High Accuracy:** Following [16, 27], we differentiate between **cyberbullying** (i.e., an act of aggression that is repeated over time) and **cyberaggression** (i.e., harassment manifested as one–off profanity). Each media session s is only considered as cyberbullying if more than a given number of aggressive comments have been posted, so that the number of false positives can be reduced.
- **Timeliness:** All comments in a media session are ordered in time. As a result, earlier detection using a **partial** set of the media session (i.e., a small number of comments) allows for timely alerts to be raised. In our approach, we encode the belief of a media session to be indicative of cyberbullying as the number of identified aggressive comments. We measure timeliness as the number of comments "saved" in comparison to a baseline. Figure 4 illustrates this metric.
- **Efficiency:** Given the soaring amount of daily Instagram sessions and comments, classification must be scalable.

To achieve the aforementioned desired properties, we propose a two–stage approach for cyberbullying detection on Instagram as


Bitches talk shit about jin all fuking day yo. Bitches get off his dick!

is this picture your wall paper? haha

Loving my print leggings that @shoptasteonline sent me!! I love prints!!

Figure 2: Overview of the proposed approach. Given a set of media sessions $\mathcal{M}$, and an alert threshold, our approach examines comments as they become available over time and raises an alert only after the number of comment–level detections surpasses the threshold. The posterior probability evolution of an aggressive (upper) and non–aggressive (lower) 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.

shown in Figure 2. 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. In the second stage, before a new comment is examined, the number of comment–level detections is compared to a predefined threshold, and the media session is classified as cyberbullying when the threshold is exceeded.

Next, we introduce our proposed optimization framework that automatically computes the probability of a comment to be indicative of aggression with high accuracy, while accounting for the effort of the framework in improving its chances of reaching a highly accurate conclusion.

3.1 Problem Formulation

In our 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 $f(m) = \{y_1, y_2, \dots, y_K\}$, 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 $f(m)$ sequentially as illustrated in Figure 2. 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 $n = 1, \dots, K$ represents the value of time and effort spent evaluating the nth feature. We additionally consider misclassification costs $C_{ij} \geq 0, i = \mathcal{B}, \mathcal{N}, j = 1, \dots, L$, 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). We factor misclassification costs into our approach to quantify the relative importance of detection errors.

We now formally describe our proposed sequential evaluation process to minimize the number of features used to accurately classify each comment m. Specifically, our proposed sequential evaluation process comprises a pair (R, D_R) of random variables. Random variable R takes values in the set $\{0, \dots, K\}$, 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 [32]. Random variable D_R denotes the possibility to select among L possible choices. Assuming that the random variables y_n are *independent under each hypothesis* $H_i, i = \{\mathcal{B}, \mathcal{N}\}$, the conditional joint probability of $\{y_1, \dots, y_n\}$ can be computed as $P(y_1, \dots, y_n|H_i) = \prod_{k=1}^{n} p(y_k|H_i), i = \mathcal{B}, \mathcal{N}$. 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 $\{y_1, \dots, y_R\}$ by the stopping time R. Equivalently, features that may be examined in the future are not used.

3.2 Optimization Setup

Our goal is to use the least number of features for detecting aggression at the comment–level 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:

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

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. Our 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, and use the optimum classification rule D_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.

3.3 Classification Rule

The classification rule D_R depends only on the accumulated information $\{y_1, \dots, y_R\}$ by stopping time R. Such accumulated information increases as more features are evaluated, thus, it becomes infeasible to optimize the cost function in Eq. (1). Instead, the *a posteriori probability* $\pi_n \triangleq P(H_{\mathcal{B}}|y_1, \dots, y_R)$, which corresponds to a sufficient statistic of the accumulated information, can be used to reach the same solution in Eq. (1) without having to consider all $|\mathcal{Y}|^R$ possible combinations in the accumulated information. Specifically, the posterior probability, π_n, at stage n, where the nth feature is extracted and evaluated to generate outcome y_n, can be iteratively computed as:

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

where π_{n-1} denotes the posterior probability of a comment being an aggressive comment given the first $n-1$ features, and $\pi_0 = p$.

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

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

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:

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

From Eq. (4), we deduce 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:

$$\tilde{J}_R \triangleq J(R, \mathcal{D}_R^{optimal}) = \min_{D_R} J(R, D_R) = \mathbb{E}\left\{\sum_{n=1}^{R} c_n + g(\pi_R)\right\}, \tag{5}$$

which depends only on the stopping time R.

3.4 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 [32]. 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 [2] 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 rule, $R^{optimal}$, as follows:

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

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:

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

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

3.5 Practical Considerations

In this section, we describe *CONcISE,* a novel algorithm for cyberbullying detection on Instagram media sessions based on Eq. (4) and Eq. (6). For each media session $s \in \mathcal{M}$, we maintain the number of aggressive comments identified at any given time along with the time of that classification. We use this information to decide when to raise an alert. Specifically, we set a threshold, and raise an alert the first time the number of comment–level detections surpasses that threshold. This design provides the flexibility of trading timeliness for false positives. To select the threshold t, we performed a grid search over values $\{2, 3, \ldots, 10\}$ and computed Area Under Receiver Operating Characteristic curves (AUC) scores of *CONcISE* as a function of t so as to optimize the precision–recall trade–off [3]. We found that best precision and recall is achieved when $t = 5$ (c.f. Section 4).

Within a given session s, the posterior probability π_0 of a given comment m is initially set to the prior probability p of a comment being an instance of aggression, and the two terms in Eq. (7) are compared. If the first term is less than the second, *CONcISE* classifies the comment based on the optimal rule of Eq. (4). Otherwise, the first feature is extracted from the comment and evaluated. *CONcISE* repeats these steps until either it decides to classify the given comment using $< K$ features, or all K features are examined.

4 EXPERIMENTAL EVALUATION

In this section, we provide a thorough experimental analysis to evaluate our proposed approach. All methods are implemented in Python, and all experiments were performed on a 64–bit machine with a dual–core Intel processor @2.7GHz and 16GB memory.

4.1 Baseline Methods

State–of–the–art Cyberbullying Detection Methods. We evaluate the effectiveness of *CONcISE* against the state–of–the–art for cyberbullying detection on Instagram, i.e., **DRFS [16]**, the best performing approach for cyberbullying detection on Twitter, **RF [6]** (i.e., **R**andom **F**orest classifier), adapted for Instagram with the best performing features used in [16], **DLR [27]** (i.e., **D**ynamic **L**ogistic **R**egression), the method for incremental cyberbullying detection on Vine [27], and **TM [7]**, a method that uses the entire history of all comments in a media session to derive a **T**emporal **M**odel for classification. Note that with the exception of [27], no research exists in the literature that addresses the challenge of timely cyberbullying detection by examining cyberbullying as a repetitive process.

Online Streaming Feature Selection Methods. Since our proposed approach, *CONcISE*, sequentially examines features one–by–one, as opposed to indiscriminately using a common set of features for all comments, we compare *CONcISE* with **SAOLA [40]** and **OFS–Density [46]**, i.e., the two best performing online streaming feature selection methods in terms of accuracy and scalability.

4.2 Dataset

We use the Instagram dataset collected by Hosseinmardi et al. [16]. The dataset has been originally collected using snowball sampling starting from a random seed node. For each user, all media the user shared, users who commented on the media, and the comments posted on the media have been collected. In total, the dataset contains 3, 829, 756 users and 9, 828, 760 comments. Of all media sessions which contain at least 40% profanities, 47.5% have been manually labeled as positive if "*there are negative words and/or comments with intent to harm someone or other, and the posts include two or more repeated negativity against a victim*" [16].

We augmented this dataset with ∼ 10K comment–level labels obtained from 10 experts. The comments span in total 22.1% of the media sessions with 40% or more profanity. We 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 dataset from [16]) for testing.

4.3 Offensive Language Dictionaries

For a fair comparison with **DRFS**, we use the exact same features (i.e., the counts of keywords 'ugly', 'shut', 'suck', 'gay', 'beautiful', 'sick', 'bitch', 'work', 'hate', and 'fuck') after removing punctuation marks and stopwords. To directly compare our approach with **DLR** [27], we consider the frequencies of 1, 384 **profane** unigrams and bigrams [35]. We further consider a dictionary of 349 offensive unigrams and bigrams (**Noswearing**) [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.

4.4 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 our 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.

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. Even though RDFS achieves the highest precision among all approaches, *CONcISE* consistently achieves the highest (or closely follows the highest) performance, with the additive advantage of using only a small percent of all features. Specifically, *CONcISE* uses up to 99.7% less features on average than the total number of features available (i.e., for Noswearing) to make a comment–level decision. Similarly, *CONcISE* uses at least 20.5% less comments on average than the best performing baseline (i.e., RF–comment with 33.07 comments on average) to make a decision at the media session level.

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	0.751	0.449	**0.858**	**0.877**	10 (0)	83
RF	0.791	0.703	0.749	0.862	13 (0)	83
DLR	0.497	0.651	0.616	0.521	4 (0)	35.90 (45.43)
DLR–10	0.733	0.463	0.760	0.684	10 (0)	35.37 (36.54)
DLR–Profane	0.522	0.644	0.638	0.541	1,384 (0)	47.70 (53.71)
DLR–Noswearing	0.762	0.598	0.743	0.733	349 (0)	45.26 (32.49)
TM	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)

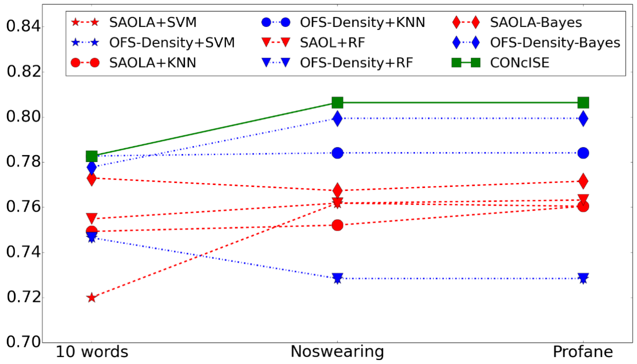


Figure 3: Accuracy comparison of *CONcISE*, SAOLA, and OFS–Density.

4.5 Comparison with State–of–the–art OSFS.

Next, we compare *CONcISE* to the baselines in terms of feature selection quality, 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, we 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 [40, 46]. We report the accuracy of *CONcISE*, as well as all combinations of baselines and classifiers. Figure 3 shows that *CONcISE* consistently outperforms OFS-Density and SAOLA over all 3 feature spaces.

4.6 Timeliness

To compare the best performing variance of *CONcISE*, namely, *CONcISE–Noswearing*, against baselines with respect to *timeliness*, we introduce a measure "# of saved comments" (as illustrated in Figure 4), 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.

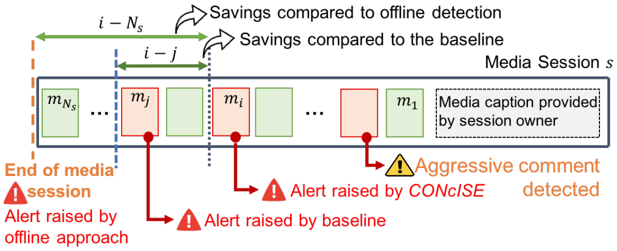

Savings compared to offline detection
$i - N_s$
Savings compared to the baseline
$i - j$
Media Session s
m_{N_s}
m_j
m_i
m_1
Media caption provided by session owner
End of media session
Alert raised by offline approach
Alert raised by baseline
Alert raised by *CONcISE*
Aggressive comment detected

Figure 4: Illustration of timeliness, i.e., the number of comments "saved".

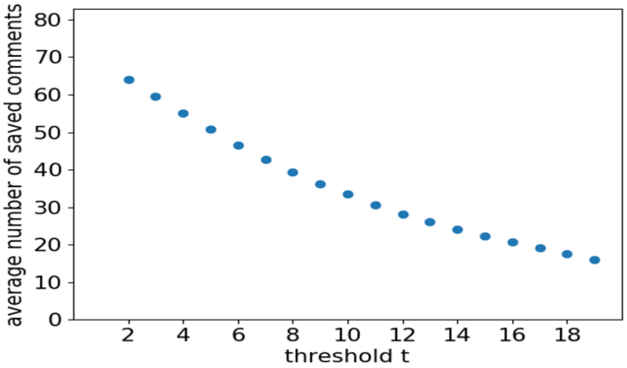


Figure 5: Average number of saved comments with respect to RDFS (a static method) as the threshold for comments detected as harassment content used to raise an alert increases.

By this definition, we respectively compare the timeliness of *CONcISE–Noswearing* to the best performing offline (i.e., RDFS) and online approaches (i.e., DLR–Noswearing), with respect to accuracy for varying threshold t. Figure 5 shows the absolute number of comments "saved" by CONcISE as compared to RDFS. As expected, the higher the threshold the more comments must be marked for harassment before an alert can be raised. Hence, the higher the threshold, the less the savings. Figure 6 shows the distribution of density of the number of saved comments as compared to DLR–Noswearing and provides strong evidence that *CONcISE–Noswearing* can detect cyberbullying media sessions faster than DLR–Noswearing (in addition to being more accurate) for all thresholds considered.

4.7 Scalability Analysis

We found baselines RDFS, RF, and TM to be the fastest among all methods with respect to runtime for session–level decisions, 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* being the fastest. Even though the difference may seem negligible for this dataset, considering a real–world scenario where a million media sessions are to be evaluated in real–time.

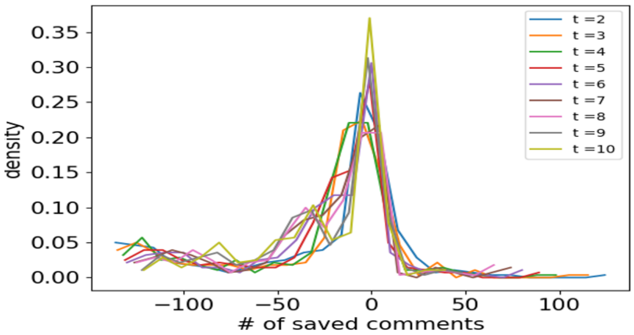


Figure 6: Distribution of average number of saved comments with respect to DLR–Noswearing for various thresholds.

5 CONCLUSION

Contributions. In this paper, we have proposed a novel approach for online cyberbullying detection on Instagram which can achieve *timely* and *accurate* detection while being *scalable* to the staggering rates at which content is generated. To achieve this goal, our two–stage approach begins by classifying comments in a media session as they become available, and an alert at the session level is raised when a threshold is exceeded. Our sequential hypothesis testing formulation drastically reduces the number of features used for classification, while maintaining high accuracy. Extensive evaluation experiments using a large–scale, real–world dataset from Instagram demonstrate the effectiveness of our proposed approach with respect to accuracy, timeliness, efficiency, and robustness, and show that it consistently outperforms the state–of–the–art, often by a considerable margin.

Limitations. Next, we would like to note the limitations of our present work that point to interesting future research directions. For one, our experimental evaluation is limited to a single dataset and therefore the performance of our approach should not be generalized to other platforms. Additionally, the inability to validate the labels in the dataset provided by [16] prevents us from obtaining a more granular understanding of the effect of false positives (similarly false negatives) on the accuracy of our approach. More importantly, comment–level labels, even though imperative for capturing the repetitive nature of cyberbullying as a process over a period of time, can be costly or time consuming to obtain.

Future Directions. There are several opportunities to build on this work. First, we plan to explore features such as user– and network–information and the sequence of conversations to further improve classification accuracy. We are also planning to evaluate the performance of our 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, we 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.

REFERENCES

[1] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. 2016. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior* 63 (2016), 433–443.

[2] D. P. Bertsekas. 2005. *Dynamic Programming and Optimal Control*. Vol. 1. Athena Scientific.

[3] Michael Buckland and Fredric Gey. 1994. The relationship between recall and precision. *Journal of the American society for information science* 45, 1 (1994), 12–19.

[4] Jiuwen Cao, Tao Chen, and Jiayuan Fan. 2014. Fast online learning algorithm for landmark recognition based on BoW framework. In *Industrial Electronics and Applications (ICIEA), 2014 IEEE 9th Conference on*. IEEE, 1163–1168.

[5] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Detecting Aggressors and Bullies on Twitter. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 767–768.

[6] 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.

[7] 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)*.

[8] AllSlang family. [n. d.]. Internet Slang Swear Word List & Curse Filter. https://www.noswearing.com/dictionary.

[9] Sujatha Das Gollapalli, Cornelia Caragea, Prasenjit Mitra, and C Lee Giles. 2013. Researcher homepage classification using unlabeled data. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 471–482.

[10] Leam Hackett. 2017. The Annual Bullying Survey 2017. https://www.ditchthelabel.org/wp-content/uploads/2017/07/The-Annual-Bullying-Survey-2017-1.pdf. (accessed on Aug. 30 2018).

[11] M. A. Hall. 1999. *Correlation-based feature selection for machine learning*. Ph.D. Dissertation. The University of Waikato.

[12] Sameer Hinduja and Justin W Patchin. 2007. Offline consequences of online victimization: School violence and delinquency. *Journal of school violence* 6, 3 (2007), 89–112.

[13] Dianne L Hoff and Sidney N Mitchell. 2009. Cyberbullying: Causes, effects, and remedies. *Journal of Educational Administration* 47, 5 (2009), 652–665.

[14] Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. 2018. Online Learning: A Comprehensive Survey. *arXiv preprint arXiv:1802.02871* (2018).

[15] Steven CH Hoi, Jialei Wang, Peilin Zhao, and Rong Jin. 2012. Online feature selection for mining big data. In *Proceedings of the 1st international workshop on big data, streams and heterogeneous source mining: Algorithms, systems, programming models and applications*. ACM, 93–100.

[16] 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.

[17] Guichun Hua, Min Zhang, Yiqun Liu, Shaoping Ma, and Liyun Ru. 2010. Hierarchical feature selection for ranking. In *Proceedings of the 19th international conference on world wide web*. ACM, 1113–1114.

[18] Hao Huang, Shinjae Yoo, and Shiva Prasad Kasiviswanathan. 2015. Unsupervised feature selection on data streams. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1031–1040.

[19] Jyrki Kivinen, Alexander J Smola, and Robert C Williamson. 2004. Online learning with kernels. *IEEE transactions on signal processing* 52, 8 (2004), 2165–2176.

[20] Robin M Kowalski and Susan P Limber. 2013. Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health* 53, 1 (2013), S13–S20.

[21] 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.

[22] Jiguang Liang, Xiaofei Zhou, Li Guo, and Shuo Bai. 2015. Feature selection for sentiment classification using matrix factorization. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 63–64.

[23] T. Marill and D. Green. 1963. On the effectiveness of receptors in recognition systems. *IEEE Transactions on Information Theory* 9, 1 (1963), 11–17.

[24] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 145–153.

[25] Guansong Pang, Longbing Cao, Ling Chen, and Huan Liu. 2016. Unsupervised feature selection for outlier detection by modelling hierarchical value-feature couplings. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 410–419.

[26] Simon Perkins and James Theiler. 2003. Online feature selection using grafting. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 592–599.

[27] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, and Shivakant Mishra. 2018. Scalable and timely detection of cyberbullying in online social networks. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. ACM, 1738–1747.

[28] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. 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, 617–622.

[29] Elaheh Raisi and Bert Huang. 2017. Cyberbullying detection with weakly supervised machine learning. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 409–416.

[30] Elaheh Raisi and Bert Huang. 2018. Weakly Supervised Cyberbullying Detection Using Co-Trained Ensembles of Embedding Models. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 479–486.

[31] Weixiang Shao, Lifang He, Chun-Ta Lu, Xiaokai Wei, and S Yu Philip. 2016. Online unsupervised multi-view feature selection. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 1203–1208.

[32] Albert N Shiryaev. 2007. *Optimal Stopping Rules*. Vol. 8. Springer Science & Business Media.

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

[34] Peter K Smith, Jess Mahdavi, Manuel Carvalho, and Neil Tippett. 2006. An investigation into cyberbullying, its forms, awareness and impact, and the relationship between age and gender in cyberbullying. *Research Brief No. RBX03-06. London: DfES* (2006).

[35] Luis von Ahn. [n. d.]. Offensive/Profane Word List. https://www.cs.cmu.edu/~biglou/resources/bad-words.txt.

[36] Jialei Wang, Peilin Zhao, and Steven CH Hoi. 2016. Soft confidence-weighted learning. *ACM Transactions on Intelligent Systems and Technology (TIST)* 8, 1 (2016), 15.

[37] Jialei Wang, Peilin Zhao, Steven CH Hoi, and Rong Jin. 2014. Online feature selection and its applications. *IEEE Transactions on Knowledge and Data Engineering* 26, 3 (2014), 698–710.

[38] 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.

[39] Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. 2018. Cyberbullying Detection on Instagram with Optimal Online Feature Selection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 401–408.

[40] 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.

[41] Aonan Zhang, Jun Zhu, and Bo Zhang. 2013. Sparse online topic models. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1489–1500.

[42] Liang Zhang, Jie Yang, and Belle Tseng. 2012. Online modeling of proactive moderation system for auction fraud detection. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 669–678.

[43] Xiang Zhang, Jonathan Tong, Nishant Vishwamitra, Elizabeth Whittaker, Joseph P Mazer, Robin Kowalski, Hongxin Hu, Feng Luo, Jamie Macbeth, and Edward Dillon. 2016. Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network. In *15th IEEE International Conference onMachine Learning and Applications (ICMLA)*. 740–745.

[44] 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.

[45] Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. In *IJCAI*. 3952–3958.

[46] 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.