Minority Report: Cyberbullying Prediction on Instagram

Charalampos Chelmis
cchelmis@albany.edu
Department of Computer Science
University at Albany, State University of New York
Albany, New York, United States

Mengfan Yao
myao@albany.edu
Department of Computer Science
University at Albany, State University of New York
Albany, New York, United States

ABSTRACT

Introduction. Cyberbullying, as a form of abusive online behavior, although not well–defined, is a repetitive process, i.e., a sequence of harassing messages sent from a bully to a victim over a period of time with the intent to harm the victim. Numerous automated, data–driven approaches have been developed for the automatic classification of cyberbullying instances, with emphasis on classification accuracy. While the importance of highly accurate classifiers is undoubted, a key pitfall of existing cyberbullying detection methods is that (i) they disregard the repetitive nature of the harassing process, and (ii) they work retrospectively (i.e., after a cyberbullying incident has occurred), making it difficult to intervene before an interaction escalates. Motivated by the scarcity of methods to anticipate cyberbullying, we focus on cyberbullying prediction with the goal of reducing the time from detection to intervention.

Methods. We formulate the prediction of the number of harassing comments a media session will receive over a period of time as a regularized multi–task regression problem. In our formulation, we consider two settings where (i) the progression of cyberbullying behavior from some time point in the near future to subsequent time points further into the future is modeled given limited knowledge of the recent past, and (ii) increasingly more historical data is accumulated to improve prediction accuracy. To validate our approach, we conduct an extensive experimental evaluation on a real–world dataset from Instagram, the online social media platform with the highest percentage of users reporting experiencing cyberbullying.

Results. Intuitively, the larger the number of observed comments in the recent past of a media session, the better the predictive power of our approach. The downside to using more historical data is that decisions must be postponed until more comments are collected. Therefore, the trade–off between accuracy and decision speed is examined. In general, our approach outperforms competing approaches by up to 31.4% and 46.2% in Recall and Mathew correlation coefficient respectively.

Discussion. Our approach can be used to effectively prioritize media sessions for increased monitoring as time goes by or for immediate intervention before a conversation escalates. In future work, we plan to incorporate additional features and investigate the generalizability of our approach on other key social networking venues where users frequently become victims of cyberbullying. Beyond cyberbullying prediction, our work is, to the best of our knowledge, the first to provide insights on the forecasting performance of multi–task regression as a function of the prediction horizon and the length of available historical data. We thus believe that our work can serve as a reference point on the forecasting performance of multi–task regression both for researchers and practitioners.

WebSci '19, June 30-July 3, 2019, Boston, MA, USA

ACM ISBN 978-1-4503-6202-3/19/06…$15.00
https://doi.org/10.1145/3292522.3326024

CCS CONCEPTS

• **Information systems** → *World Wide Web*; *Social networks*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Web and society; cyberharassment; cybersafety; online well–being

ACM Reference Format:
Charalampos Chelmis and Mengfan Yao. 2019. Minority Report: Cyberbullying Prediction on Instagram. In *11th ACM Conference on Web Science (WebSci '19), June 30-July 3, 2019, Boston, MA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292522.3326024

1 INTRODUCTION

A growing number of online users abuse the Internet to harass other users, leading to a tide of cyberbullying incidents [12, 18]. 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 [24, 37]. Cyberbullying, a type of cyberharassment, can take many forms, 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 [24]. Cyberbullying is permanent (i.e., content remains accessible online unless removed) and potentially widespread (i.e., online social media provide a wide audience, and quick spread of online posts).

The potentially devastating real–world consequences to victims, which include but are not limited to psychological suffering and isolation, escalated physical confrontations, and suicide [18, 20], have led to the development of numerous methods for the automatic classification of cyberbullying instances [1, 35, 36] in a variety of online social networks and with a plethora of constraints [5, 8, 32, 33, 43, 44]. While highly accurate classifiers are of paramount importance to greatly reduce the burden on human moderators employed by online social media platforms, a key pitfall of existing cyberbullying detection methods is that they work retrospectively (i.e., after a cyberbullying incident has occurred), making it difficult to intervene before an interaction escalates. In contrast, approaches for cyberbullying prediction would be advantageous in

(i) identifying in advance vulnerable users that may fall victims of cyberbullying (i.e., before toxic comments, from which text–based features can be extracted, appear), using only limited data such as the image and caption provided by the creator of the media, and (ii) scaling detection methods to the staggering rates at which content is generated (e.g., 95 million photos and videos are shared on Instagram per day[1]) in online social media by targeting available computational resources on the subset of media sessions projected to experience cyberbullying, rather than blindly classifying all media sessions indiscriminately.

Present Work. Motivated by the scarcity of methods to anticipate cyberbullying [22, 26], and the fact that most existing approaches disregard the repetitive nature of the harassing process [44], we focus on cyberbullying prediction on Instagram, the online social media platform with the highest percentage of users reporting experiencing cyberbullying [18]. Instagram has more than 800 million registered users as of Sep. 2017, and over 40 billion uploaded photos as of Oct. 2015.

Instead of trying to detect all possible harassment, aggressive, antisocial, inflammatory, or toxic content in online social media, we focus on *harassing comments*, that are common to a number of types of unwanted behavior, including cyberharassment and cyberbullying. More importantly, we are interested in exploiting the *temporal dynamics* of the *repetitive* bullying behavior over time directly in our modeling. To this end, we formulate cyberbullying prediction as regularized multi–task regression [2, 4, 13], where the progression of the number of hateful comments Instagram content will receive over time is estimated from limited historical data. Our experimental results show that our proposed approach consistently outperforms competing methods. We also perform sensitivity analysis to examine the impact of the parameters on the performance of the proposed approach.

Our main contributions can be summarized as follows:

- **Novel Formulation:** We propose a novel formulation of cyberbullying prediction on Instagram as a regularized multi–task regression problem. In order to support different intervention strategies, as well as to assess the difficulty of variants of this problem, we consider two settings as follows. In our first formulation, we estimate the progression of harassment from some time point in the near future to subsequent time points further into the future based on limited knowledge of the recent past. In our second formulation, increasingly more historical data is accumulated to improve prediction accuracy. A model learned from the first formulation would result in projections at multiple times in the future, therefore providing a potential timeline for escalating discourses, whereas, the second formulation attempts to improve overall prediction accuracy by leveraging common knowledge shared across the forecasting tasks.
- **Experimental Evaluation:** We evaluate the forecasting accuracy of our approach as a function of the prediction horizon and the length of historical data on a real–world dataset of 10K Instagram comments. To ensure the **reproducibility** of our work, we make the source code of our approach available at https://github.com/IDIASLab/CyberBullyingPrediction.
- **Broader Applicability:** Intuitively, the predictive power of forecasting models improves with the accumulation of historical data and deteriorates further into the future predictions. To the best of our knowledge, this is the first work to examine the forecasting performance of multi–task regression as a function of the prediction horizon and the length of available historical data.

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

Outline. The rest of this paper is organized as follows. We first review prior and related work in Section 2. We formulate the problem of cyberbullying prediction in online social networks in Section 3. We describe our evaluation methodology and results on a real–world dataset in Section 4. We conclude with a discussion of our results, limitations, and possible future directions in Section 5.

2 RELATED WORK

Cyberbullying Detection. We argue that it is imperative to predict the potential of a media session to receive harassing comments in the future so as to facilitate timely interventions. However, with the exception of few recent attempts at cyberbullying prediction [22, 26], the majority of prior work, an overview of which can be found at [1, 35, 36], focuses on cyberbullying detection. Nevertheless, with the exception of [44], no prior work has studied cyberbullying as a repetitive process. Out of the two recent methods for cyberbullying prediction, [22] examined prediction feasibility given only the initial image–content and text caption of an Instagram post, whereas [26] focused on predicting harassment escalation in comments following the first hostile comment in a discussion. In contrast to these methods, the approach presented in this work can make predictions at any given time.

Cyberbullying Indicators. Despite the scarce research work on cyberbullying prediction, features that may be useful for predicting cyberbullying instances have been explored in the context of cyberbullying detection. Specifically, text has been a major factor in detecting cyberbullying in online social media. However, features ranging from gender information, user context, linguistic and non–verbal features, and graph properties have also been used [5, 6, 9, 46]. The use of profanity and hate speech has been well correlated with toxic comments on the Web [10, 11, 16, 17]. However, what may constitute hate speech and profanity is context dependent (i.e., relative to time and location) [3]. The sociological literature review in [26] motivated some of the features used in this work.

Hostility & Harassment Detection. The related problem of detecting harassing and hostile behavior on the Web has been very well studied [5, 7, 10, 16, 21, 23, 25, 28, 29, 34, 38, 42], with the majority of this body of work having primarily focused on text–based features, excluding critical information in the various modalities (e.g., image, video, user profile, time, and location) typically associated with content shared on online social media [8]. Similarly to state–of–the–art for cyberbulying detection, to the best of our knowledge, all existing approaches for detecting hostile content on the Web ignore the fact that as a process that unfolds with time, cyberbullying is repetitive in nature [44].

Multi–Task Learning. Multi–task learning utilizes commonalities among multiple related prediction problems to improve performance [2, 4, 13]. The key challenges in multi–task learning are to define and exploit such relatedness [13], while maintaining a small number of predictive features shared across all learned models [2]. Multi–task learning approaches have been applied in many domains, however, to the best of our knowledge, ours is the first work that applies multi–task learning for harassment intensity prediction. More importantly, no prior work has examined the forecasting power of the multi–task learning framework as a function of the prediction horizon and the length of historical data.

3 PROBLEM FORMULATION

Consider a large set $\mathcal{M}$ of N media sessions, where each media session $s \in \mathcal{M}$ belongs to 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 $\{(c_1, t_1), \ldots, (c_{N_s}, t_{N_s})\}$ from users in $\mathcal{U}$, where $c_i, i \leq N_s$ indicates the i–th comment with corresponding timestamp t_i, and N_s denotes the number of comments in s. For training and testing purposes, we additionally consider $\forall s \in \mathcal{M}$ set $\{y_1, \ldots, y_{N_s}\}$, where y_i denotes the cumulative number of harassing comments session s has received up to time t_i.

Our goal is to predict **harassment intensity** (i.e., the future number of harassing comments) at t time points in the future, for any given media session. Specifically, given time point γ, we wish to estimate harassment intensity up to timestamp $\gamma + h$, where h is the prediction horizon. In this context, *short–* and *long–term* prediction refer to the scenarios where $h = 1$ (e.g., the next time point) and $h > 1$, respectively.

From an intervention perspective, a greater prediction horizon provides more flexibility in taking preventative measures as the time between the final comment observed by the system and the time of escalation increases. However, the ability to accurately forecast the number of harassing comments up to some time in the future may depend on (i) the number of past comments on a media session (i.e., *length of historical data*), and (ii) how far ahead in the future a prediction is to be made (i.e., *prediction horizon*). Intuitively, long–term prediction is harder than short–term prediction, as many factors can potentially affect human behavior online [30].

Figure 1 shows the evolution of harassment intensity over time for a random sample of media sessions in our dataset (Section 4.3). As expected, not all media sessions experience the same level of harassment. More importantly, harassing behavior tends not to be evenly spread out in time. Both (i) bursts of harassing comments, which may be indicative of abusive behavior in which several people gang up on a victim [39], and (ii) incremental changes, which may reflect repetitive harassing comments from a single individual, can be observed.

To capture such dynamics, we formulate the problem of harassment intensity prediction as a **regression problem**. Specifically, in order to predict at time γ the harassment intensity, $y_s^{\gamma+h}$, of media session s at future time $\gamma + h$, we extract training features x_s^{γ} from the past *lag* comments (i.e., at times $\gamma - 1, \gamma - 2, \ldots, \gamma - lag$). For each media session, we construct a training input x_s^{γ} and output $y_s^{\gamma+h}$, and we wish to learn a function $y_s^{\gamma+h} = f(x_s^{\gamma}, h)$ for multiple

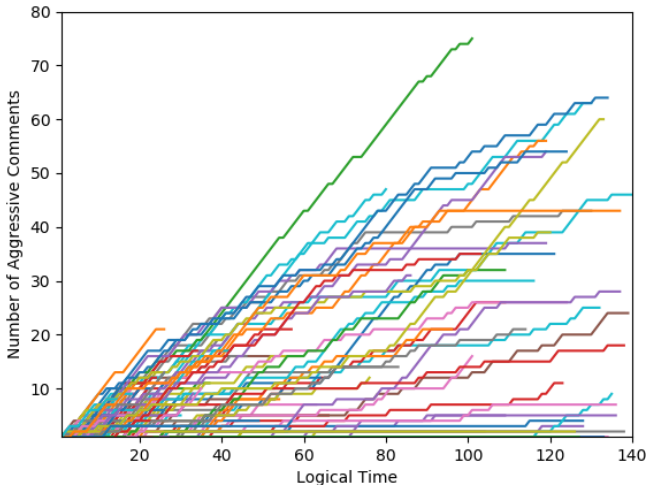

Figure 1: Temporal dynamics of harassment intensity in a random sample of Instagram media sessions in our dataset. Each curve corresponds to a media session, and shows the cumulative number of harassing comments the session attracts over time. The x–axis represents logical time, which increases for each media session when a new harassing comment arrives.

combinations of γ and h. By considering the prediction of harassment intensity at a single time point as a regression task, a simple approach to learn $f(x_s^{\gamma}, h)$ is to train one model for each combination of γ and h independently. However, different time points in the future may be represented as distinct tasks, or alternatively, tasks can be defined by the length of historical data used for prediction. Additionally, jointly training multiple regression problems for different combinations of γ and h may be advantageous due to the intrinsic temporal smoothness relationship among the regression problems (e.g., the difference in the number of harassing comments between two consecutive time points should in general be small).

The above reasoning motivates us to formulate harassment intensity prediction as a multi–task regression problem [2, 4, 13]. Specifically, we consider two formulations, namely:

- Fixed–Lag Varying–Horizon Prediction Model (*FLVH*), and
- Varying–Lag Fixed–Horizon Prediction Model (*VLFH*),

which are detailed in Sections 3.1 and 3.2 respectively. *FLVH* attempts to model the progression of cyberbullying behavior from some time point in the near future to subsequent time points further into the future, given a limited knowledge of the recent past (this knowledge of the recent past is common among tasks). Conversely, *VLFH* focuses on improving the prediction of harassment intensity at a given point in time (common to all tasks) by accumulating increasingly more historical data. Figure 2 provides a visual illustration of our proposed formulations.

3.1 Fixed–Lag Varying–Horizon Prediction Model

Consider a multi–task regression problem of t time points with n training samples of d features each. Let $X_i = \{X_{i,1}, ..., X_{i,n}\}, 1 \leq$

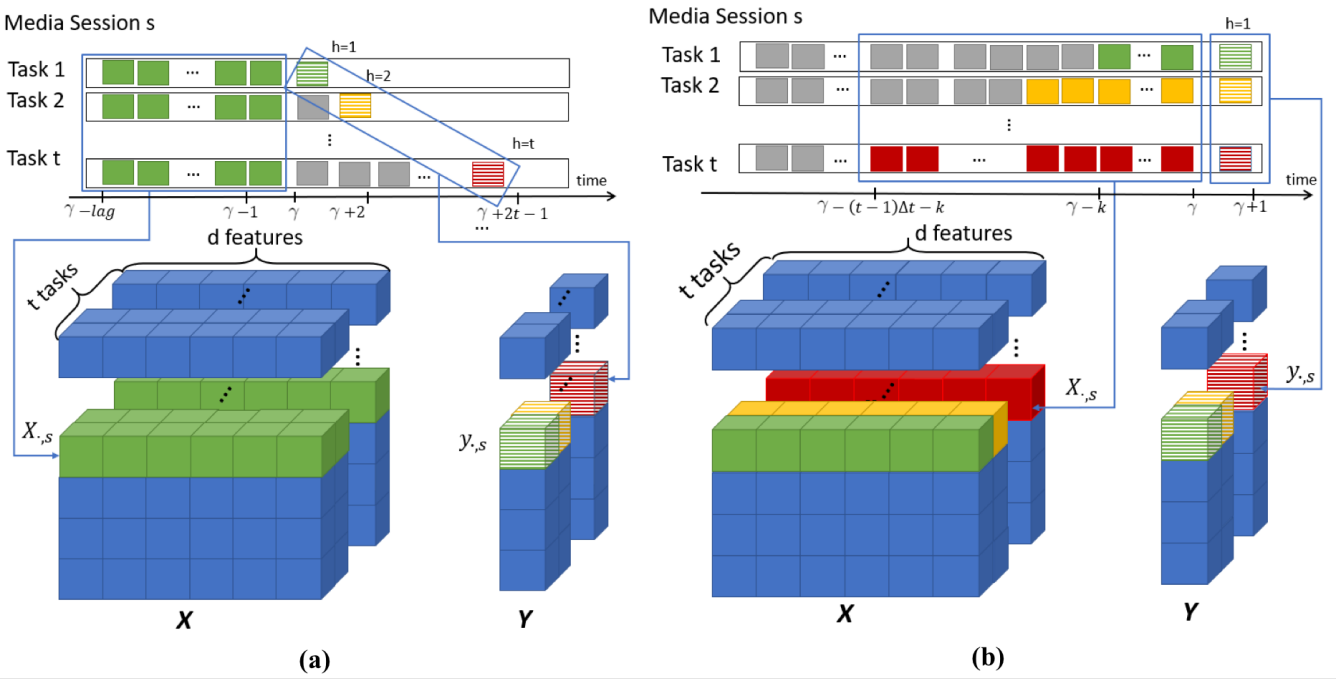

Figure 2: Illustration (better seen in color) of the proposed formulations for cyberbullying prediction on Instagram. (a) Fixed–Lag Varying–Horizon prediction model (Section 3.1). (b) Varying–Lag Fixed–Horizon prediction model (Section 3.2)

$i \leq t$ be the input data, and $Y_i = \{y_{i,1}, .., y_{i,n}\}$ be the targets, where $X_{i,s}$ is a vector of length d, each element of which is a feature extracted from the set of observed comments for task i, and $y_{i,s}$ is the predicted number of aggressive comments that media session s will receive up to future time $h_i = i\tau$. Parameter $\tau \in \mathbb{Z}^+$ is used to introduce prediction gaps (i.e., make the time points for which prediction is to be made non consecutive). The goal is to learn t models $f^i(X_i) = X_i^T W_i$, with weight matrix $\mathbf{W} = \{W_i | i = 1, ..., t\}$, where a linear model W_i is to be estimated for $\forall i$ so as to predict a harassment intensity score Y_i for each media session up to future time h_i, given X_i. Therefore, our objective is to estimate matrix $\mathbf{W}$. Figure 2(a) illustrates this formulation.

The rationale behind this formulation is that in practical applications, it can be used to predict the progression of harassment in a media session over time, from the near future (e.g., $h = 1$), to a time $h = t\tau$ further into the future, where t or τ can be arbitrarily large, given only limited knowledge (i.e., k comments) from the recent past. Note that long–term prediction is generally harder than short–term prediction considering the many factors that can potentially affect the discourse of social interactions in online social networks [30]. Typically two broad categories of methods exist for long–term prediction [41]: (i) training a model for each prediction horizon, and (ii) iteratively use previously predicted values as input to the next prediction task. We chose the first category when formulating *FLVH* to avoid the error accumulation problem of iterative methods [41].

3.2 Varying–Lag Fixed–Horizon Prediction Model

Similarly to our *FLVH* model, our goal in this formulation is to learn t models $\mathbf{W} = \{W_i | i = 1, ..., t\}$ to predict the number of aggressive comments a media session will receive by a future time h. However, unlike *FLVH* in which the number of observed comments across the t tasks is fixed, here, we vary the length of the historical data used in each task by considering varying lags, $lag_i = (i-1) \cdot \Delta t + k$, so as to incorporate progressively more historical data into the predictive model. Variable $k \geq 1$ controls the minimum number of comments considered. In other words, *VLFH* is designed to predict harassment intensity for each media session at a future time point h (which is common across tasks) by learning a model based on past comments, starting from the k most recent comments, and incorporating Δt more comments at a time. Input data X and targets Y are constructed in the same manner as in *FLVH*. Figure 2(b) illustrates this formulation.

The benefit of this formulation is that the impact (if any) of additional past information on the performance of harassment intensity prediction can be quantitatively evaluated.

3.3 Loss Function

Each of the t tasks in our two formulations above can be learned independently using conventional single–task learning. However, independently learning models to predict far into the future could result in inferior predictive power; intuitively the further the prediction into the future, the less indicative historical data becomes. Moreover, the majority of media sessions on Instagram has been

shown to receive $\leq$ 15 comments on average [22], leading to a sparsity problem (i.e., not enough training data).

To address these challenges, both of our proposed formulations learn all prediction tasks simultaneously, leveraging in this way commonalities among tasks (e.g., shared features extracted from the same historical data) to effectively address the training data sparsity problem, as well as the intrinsic temporal smoothness among different tasks (i.e., the number of harassing comments cannot vary significantly from one timestamp to another) to potentially improve overall prediction performance. The additional advantage of our proposed formulations is that they can both be used in an online setting; a prediction can be made at any point in the lifetime of a session once training has been performed.

Formally, each of our proposed formulations solves the following general optimization problem: $\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) + \lambda\Omega(\mathbf{W})$, where $\mathcal{L}(\mathbf{W})$ denotes the empirical loss function, $\Omega(\mathbf{W})$ encompasses regularization terms that encode task relatedness, and λ is a vector of tuning parameters used to balance the trade–off between the loss and regularization terms. Specifically, the goal for each task i is to learn matrix $\mathbf{W}$, such that the penalized empirical loss

$$\mathcal{L}(\mathbf{W}, \mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{t} \|X_{i,s} W_i - Y_i\|_F^2 + \lambda_1 \|\mathbf{W}\|_F^2 + \lambda_2 \sum_{i=1}^{t-1} \|W_i - W_{i+1}\|_F^2 + \lambda_3 \|\mathbf{W}\|_{2,1} \quad (1)$$

is minimized. In Eq. 1, the first term corresponds to the least–squares loss function. The first regularization term, $\lambda_1 \|\mathbf{W}\|_F^2$, controls the generalization error, λ_1 controls the sparsity of $\mathbf{W}$ (equivalently the complexity of the trained models), and $\|.\|_F^2$ is the square of Frobenius norm of a matrix. The second regularization term, $\lambda_2 \sum_{i=1}^{t-1} \|W_i - W_{i+1}\|_F^2$, controls the similarity between two neighboring tasks. When λ_2 is large, the difference between any two neighboring tasks is forced to be small (i.e., large prediction deviations at neighboring time points are penalized). The group Lasso regularization term, $\lambda_3 \|W\|_{2,1}$, based on the $\ell_{2,1}$–norm penalty for feature selection [45], ensures that all models at different time points share a common set of features. This is achieved by introducing row sparsity in $\mathbf{W}$ across all tasks using the $\ell_{2,1}$–norm.

Learning all models requires obtaining the optimal weight matrix $\mathbf{W}$ by computing: $\mathbf{W}^* = \arg\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \mathbf{X}, \mathbf{Y})$. As the objective function consists of smooth and non–smooth terms, methods such as subgradient descent can be used to tackle convex and non–differentiable functions, but have high complexity $O(\frac{1}{\sqrt{K}})$, where K denotes the number of iterations. For faster convergence rate, we use the accelerated gradient descent method, an iterative algorithm with $O(\frac{1}{K^2})$ complexity [27]. In accelerated gradient descent, the solution $W_i + 1$ on each step is computed as a gradient of the current search point W_i. The key operation in this iterative process is the computation of the proximal operator $\mathbf{W}^* = \arg\min_{\mathbf{W}} \frac{\eta}{2} \left\|\mathbf{W} - (W_i - \frac{1}{\eta}\nabla\mathcal{L}(W_i)\right\|^2 + \Omega(\mathbf{W})$, where η is the step size, to find the next search point W^* based on the current search point W_i. Accelerated gradient descent differs from the subgradient method in the sense that the current search point W_i is the affine combination of the previous two points with parameter α, instead of only using the latest one. Specifically, W_i is updated as $W_{i+1} = W_i - \alpha(W_i - W_{i-1})$, where W_i is initialized by $X_i \times Y_i$, and the stopping criterion is set to 10^{-5}. Parameters α and η are initialized to −1 and 1, respectively.

4 EXPERIMENTAL EVALUATION

In this section, we provide a thorough experimental evaluation of our proposed formulations. Specifically, we begin by evaluating the effectiveness and efficiency of our approach on real data in comparison to baselines. We continue by studying the parameter sensitivity of our formulations. In our experiments, we considered $t = 10$ tasks, and set $k = 10$ and $\tau = 1$ for both *FLVH* and *VLFH*, and $\Delta t = 2$ for *VLFH*. All experiments were conducted on a 64–bit machine with a dual–core Intel processor @2.7GHz and 16GB memory.

4.1 Baselines

To the best of our knowledge, only two methods have thus far been proposed for cyberbullying prediction, both on Instagram [22, 26]. We included both of these methods in our experimental evaluation for performance comparison.

- **LRFS [22]:** Logistic Regression classifier with forward Feature Selection for cyberbullying prediction on Instagram media sessions based on a cohort of features extracted from the first 15 comments and caption, post time, user properties, and image content. Our performance comparison in Section 4.5 demonstrates the superiority of our approach over LRFS, while using only a fraction of the features used by LRFS.
- **LRLR [26]:** Logistic Regression classifier with L2 Regularization that, given all comments up to and including the first hostile comment, predicts whether the total number of hostile comments on a media session will be greater than or equal to a predetermined threshold N after some future time h. LRLR has been shown to achieve its best performance for a threshold of 10 (i.e., $N = 10$) and a lead time of 3 hours (i.e., $h = 3$) [26]. We reached the same conclusion in our experiments, and thus set parameters $N = 10$ and $h = 3$ (this corresponds to ~ 1 timestep, on average, in our formulation) for a fair comparison in Section 4.5.

In addition to state–of–the–art for cyberbullying prediction, we are also interested in evaluating the benefit of learning multiple tasks simultaneously as opposed to learning t tasks independently. Therefore, we consider a final baseline, termed **STL** for Single Task Learning, in which the penalized empirical loss function in Eq. 1 is used, but contrary to our methods, all models are trained independently by setting the term $\lambda_2 \sum_{i=1}^{t-1} \|W_i - W_{i+1}\|_F^2 = 0$ in Eq. 1.

4.2 Evaluation Metrics

We use coefficient of determination (i.e., R^2 score) as our main evaluation criterion to measure how well observed outcomes are replicated, based on the proportion of total variation of outcomes explained by the model [40]. A negative R^2 score indicates bad fit, whereas $R^2 = 1$ indicates perfect fit. The regularization terms in our loss function used to learn a more generalizable model (i.e., reduce overfitting) are expected to result in a relatively low R^2 score due to

larger mean squared error. Additionally, the frequency of harassing comments could be arbitrarily large, leading to unbounded errors.

We further consider the most prevalent performance metrics in empirically evaluating the classification performance of cyberbullying detection methods. These include accuracy, precision, recall, and F–measure. As such metrics however can result in misleading conclusions in highly imbalanced datasets [19], we additionally consider the Matthews correlation coefficient (MCC), which is less sensitive to data skewness as it considers mutual accuracies of both classes and all four values of the confusion matrix [31]. We compute these metrics for the cyberbullying class by treating the harassment intensity prediction problem as a binary classification task. Specifically, we consider a naive classification rule motivated by the definition in [22], where a media session is classified as an instance of cybebullying if the predicted frequency of aggressive comments is ≥ 2. This simplified classification problem additionally enables a direct and fair comparison with the baselines.

4.3 Dataset

We use comments spanning 22.1% of all media sessions containing $\geq$ 40% profanities from the Instagram dataset available by [22]. Comments have been manually annotated by 10 experts. The original dataset has been 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 had been collected. Of all media sessions containing at least 40% profanities 47.5% had 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*" [22]. We performed 3–fold cross validation for model selection, where within each fold, 2/3 of the data was used for training, and the rest for testing.

4.4 Feature Engineering

Although feature engineering may be as important as modeling in a prediction problem, our main objective in this work is to improve upon existing cyberbullying prediction methods. For a fair comparison, we consider the following features: #. of mentions, #. of words, density of uppercase, density of punctuation, #. of hashtags, #. of urls, density of bad words from a dictionary [14], compound Vader sentiment [15], and ten unigrams selected by [22] and [26]. Although [26] uses additional features (i.e., unigrams, word2vec, and lexicons), incorporating all such features into our multi–task regression formulation could be problematic in terms of efficiently solving the optimization problem to find $\mathbf{W}$. The use of a much smaller subset of features as compared to [26] could put our approach at a major disadvantage in terms of data representation. Nevertheless, our results show improvements over *LRLR* even with such a significantly smaller feature set.

Table 1 shows the specific features, ranked from highest to lowest, selected by our proposed formulations across different tasks. The features selected are very relevant to cyberbullying detection, as well as consistent across tasks.

Approach	Features
FLVH	"fuck", "bitch", #. of mentions, "hate", #. of uppercase letters, "beauty", text length
VLFH	"fuck", "bitch", #. of mentions, "hate", #. of uppercase letters, "ugly", # of hashtags, "shut", "gay"

Table 1: Top features (ranked by coefficients) selected by *FLVH* (top) and *VLFH* (bottom).

	Accuracy	Recall	Precision	F–measure	MCC
FLVH (h = 1)	0.7087	0.8807	0.6085	0.7173	0.4743
FLVH (h = 2)	0.7152	0.8841	0.6290	0.7330	0.4807
FLVH (h = 3)	0.7138	0.8984	0.6365	0.7431	0.4792
FLVH (h = 4)	0.7198	0.9079	0.6497	0.7558	0.4868
FLVH (h = 5)	0.7237	0.9164	0.6592	0.7653	0.4914
FLVH (h = 6)	0.7261	0.9305	0.6656	0.7747	0.4945
FLVH (h = 7)	0.7317	0.9350	0.6758	0.7834	0.4992
FLVH (h = 8)	0.7345	0.9390	0.6829	0.7897	0.4990
FLVH (h = 9)	0.7412	0.9405	0.6939	0.7975	0.5049
FLVH (h = 10)	0.7431	0.9471	0.6982	0.8026	0.5055
LRFS	0.7489	0.7206	0.7217	0.7211	0.4768

Table 2: *FLVH* performance comparison with *LRFS* [22].

4.5 Performance Comparisons

1) Fixed–Lag Varying–Horizon Prediction: LRFS [22] resembles our *FLVH* model in that a fixed number of comments is used to classify media sessions. However, unlike *FLVH*, in which predictions are made in predetermined points h in the future from any given time γ, a "prediction" by *LRFS* is made only after the last comment for a given media session becomes available. We also report the aggregate performance of *FLVH* across all tasks for comparison with *LRFS*. Note that this comparison is unfair to our model since the problem we are trying to solve does not concern the "eventual" status of media sessions (i.e., binary classification after the last comment for a given media session becomes available). Instead, *FLVH* focuses on predicting the number of harassment comments a session is expected to receive in the near future (i.e., regression problem).

Table 2 summarizes this performance comparison. The results indicate that our approach significantly outperforms *LRFS* with respect to recall, F-measure and MCC both for individual tasks (i.e., different prediction horizons) and on average.

2) Varying–Lag Fixed–Horizon Prediction: LRLR [26] uses all available comments in a media session to predict the presence of cyberbullying 3 hours after the most recent hostile comment. We found the average number of comments received within 3 consecutive hours in our dataset to be 0.042. Since *LRLR* uses a fixed prediction horizon for all media sessions, it is more similar to our proposed *VLFH* method when the prediction horizon is set to 1. Unlike *VLFH* where in task i, only $k + \Delta i$ lag comments are used, *LRLR* uses all past comments up to the first hostile comment to predict harassment intensity for all media sessions. Nevertheless, to verify the extent to which the amount of past knowledge affects (if at all) prediction performance, we compare all *VLFH* tasks to *LRLR*.

Table 3 shows the results. Notably, the performance of *VLFH* across all metrics improves as more past comments are considered.

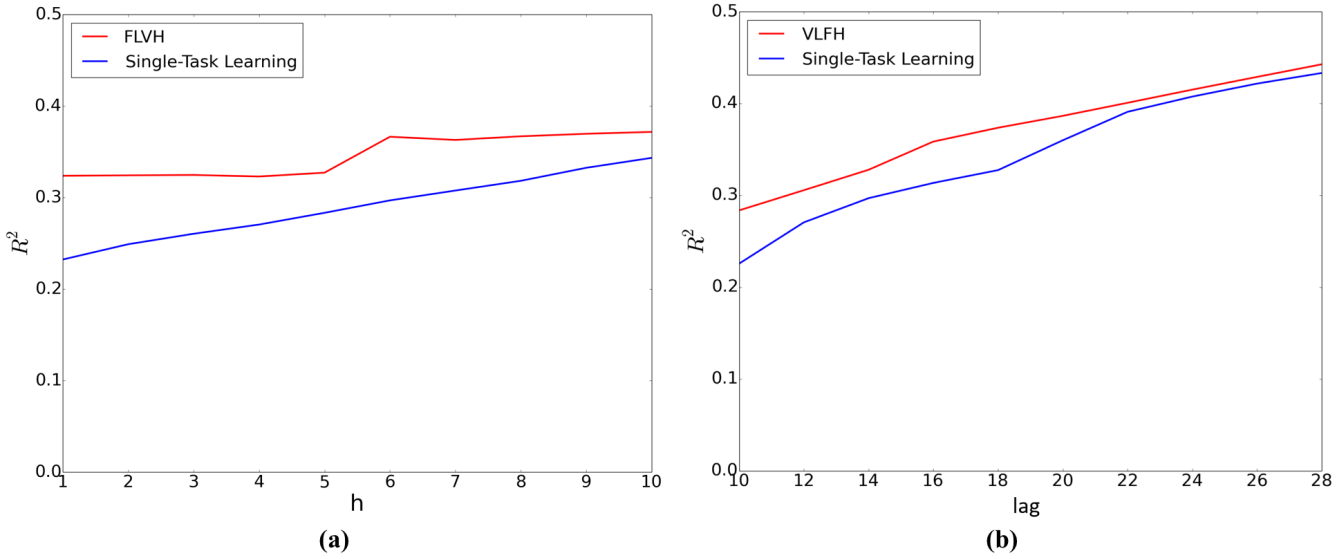

Figure 3: Comparison between (a) *FLVH* and (b) *VLFH* with single–task learning in terms of R^2 score.

	Accuracy	Recall	Precision	F–measure	MCC
VLFH (lag = 10)	0.7131	0.8759	0.6237	0.7286	0.4712
VLFH (lag = 12)	0.7272	0.9179	0.6504	0.7613	0.5020
VLFH (lag = 14)	0.7395	0.9370	0.6728	0.7832	0.5204
VLFH (lag = 16)	0.7560	0.9517	0.6970	0.8047	0.5441
VLFH (lag = 18)	0.7740	0.9651	0.7196	0.8244	0.5719
VLFH (lag = 20)	0.7883	0.9698	0.7392	0.8389	0.5902
VLFH (lag = 22)	0.8002	0.9731	0.7557	0.8508	0.6052
VLFH (lag = 24)	0.8087	0.9761	0.7677	0.8594	0.6149
VLFH (lag = 26)	0.8151	0.9784	0.7772	0.8662	0.6198
VLFH (lag = 28)	0.8157	0.9803	0.7800	0.8688	0.6161
LRLR	0.5918	0.4759	0.3193	0.3704	0.1017

Table 3: *VLFH* performance comparison with *LRLR* [26].

The results also show that *VLFH* consistently outperforms *LRLR* in terms of accuracy, recall, precision, F–measure and MCC.

3) Advantage of Multi–Task Regression: Figure 3 reveals the advantage of *FLVH* and *VLFH* in predicting the future by training tasks jointly given the same amount of access to past information. Both of our models take advantage of the commonality between tasks to address the sparsity of training samples and improve the forecasting performance for all future time points, as opposed to *STL* (i.e., single–task learning model) in terms of R^2 score. Specifically, the results demonstrate that learning all tasks jointly is advantageous to single–task learning under the assumption that the further in the future a prediction is to be made, the less indicative the present (similarly, the recent past) becomes.

4.6 Parameter Sensitivity Analysis

An important issue in the practical application of *FLVH* and *VLFH* is the selection of regularization parameters λ_1, λ_2, and λ_3. Ideally, λ_1 and λ_3 should each be set to a large value so that only the most powerful features will be selected, and the time complexity of the learned model becomes small. Similarly, a large value for λ_2 should be used to penalize large prediction deviations at neighboring time points. Setting $\lambda_2 = 1000$, while varying parameter λ_1 and λ_3 accordingly from 0 to 1000 with a step size of 200, we found the maximum difference between accuracy scores across all tasks to be less than 0.9%. We obtained similar results for λ_2 when keeping λ_1 fixed. We obtained these results for both *FLVH* and *VLFH*. The "insensitivity" to parameter λ_3 can be potentially explained by the large variance in the discriminating capability of features for the particular problem of harassment intensity prediction. In other words, it is possible that even without regularization (i.e. $\lambda_3 = 0$), the coefficients of "less important" features are already set close to zero.

Next, we focus on the sensitivity (if any) of *FLVH* and *VLFH* on the number of observed comments used for prediction (i.e., *lag*), and prediction horizon *h*. Figure 4 shows the sensitivity results of varying the value of *h* from 5 to 20 for *FLVH* (i.e., a task refers to a time point progressively further away in the future). In general, for all tasks, accuracy increases as *h* increases. Intuitively, the larger the number of observed comments in the recent past of a media session, the better the predictive power of the model, even for 10 comments into the future. Thus, setting *h* to larger values can yield better accuracy. The downside to setting *h* to larger values is that decisions must be postponed until more comments are collected. Therefore, the trade–off between accuracy and speed of decision must be considered when deciding which value of *h* to use.

Figure 5 shows the sensitivity results of *VLFH* (i.e., the number of observed comments varies across tasks) on the prediction horizon, in the range $\{3, 5, 10, 15, 20\}$. Note that in our dataset, the maximum number of comments in a media session is 147. Thus, for $lag > 28$, no media sessions exist with enough comments to train a model with > 6 tasks. Accuracy is relatively stable for all tasks, with a slight increase as *h* increases for $10 \leq lag \leq 16$ (i.e., for up to 16 past comments). Accuracy is quite stable even when *h* is large,

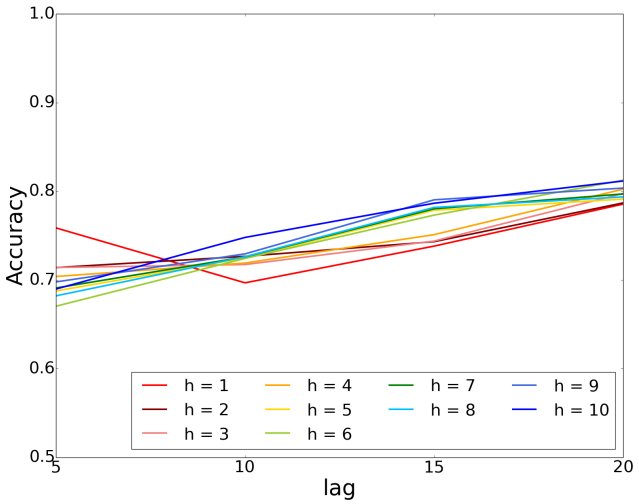

Figure 4: Sensitivity analysis of the proposed *FLVH* model on the number of observed comments (i.e., *lag*).

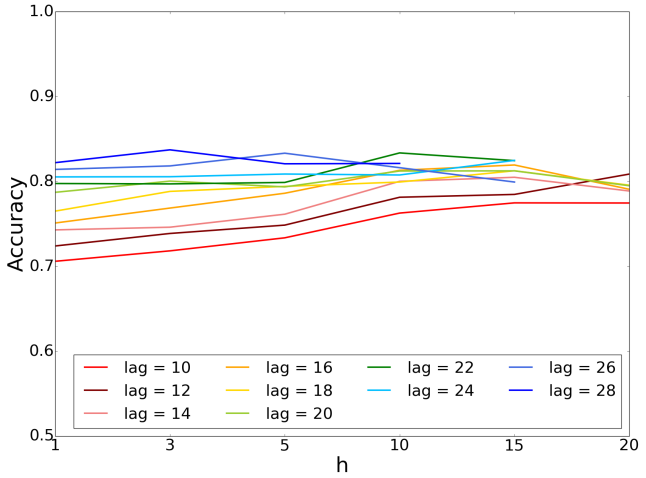

Figure 5: Sensitivity analysis of the proposed *VLFH* model on prediction horizon.

demonstrating that a good prediction can be made even with a small number of observed comments.

5 CONCLUSION

Contributions. In this paper, we presented two novel multi–task regression formulations to the problem of harassment intensity prediction on Instagram given a limited amount of historical data. By distinguishing between media sessions that are more likely to receive many harassing comments in the future, and those that are expected to receive few or none, our proposed approach can be used to effectively prioritize media sessions either for increased monitoring as time goes by or for immediate intervention before a conversation escalates, rather than investigating an event after its occurrence.

Our work considered the estimation of predictive models at different time points in the future with both fixed and varying lengths of historical data as a multi–task regression problem. Our extensive experimental evaluation results demonstrate the benefit of leveraging shared information between prediction tasks, which effectively increases the sample size, and incorporating into the training process the intrinsic temporal smoothness relationship between tasks to improve forecasting accuracy. Our results additionally showed that our approach can effectively predict harassment intensity on Instagram media sessions, outperforming competing methods by up to 31.4% and 46.2% in recall and Mathew correlation coefficient respectively.

Future Directions. In our ongoing work, we focus on features that have been shown to be informative for cyberbullying classification and more recently for harassment prediction. Given that Instagram is primarily a photo-sharing site, in future work, we plan to investigate the predictive power of non–text features extracted from image classification algorithms. Attributes engineered from user profiles and activity history as well as network structure information may provide additional context for forecasting. Finally, it may be possible to extract from these data insights into human behavior. For example, it may be possible to identify responses (if any) that may diffuse, as opposed to escalate, harassment in online social media. 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.

Broader Applicability. Beyond cyberbullying, our work is, to the best of our knowledge, the first to examine the forecasting power of the multi–task regression framework as a function of the prediction horizon and the length of historical data. Our findings based on our experimental results reveal the advantage of multi–task regression over traditional single–task learning methods for forecasting, particularly so as the prediction horizon increases. We thus believe that our work can serve as a reference point on the forecasting performance of multi–task learning both for researchers and practitioners.

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] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. 2006. Multi-task feature learning. In *Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS'06)*. MIT Press, Cambridge, MA, USA, 41–48.

[3] Amy Bellmore, Angela J Calvin, Jun-Ming Xu, and Xiaojin Zhu. 2015. The Five w's of "Bullying" on Twitter: Who, What, Why, Where, and When. *Computers in Human Behavior* 44 (2015), 305–314.

[4] Rich Caruana. 1997. Multitask learning. *Machine learning* 28, 1 (1997), 41–75.

[5] 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.

[6] Charalampos Chelmis, Daphney-Stavroula Zois, and Mengfan Yao. 2017. Mining patterns of cyberbullying on twitter. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*. IEEE, 126–133.

[7] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 71–80.

[8] Lu Cheng, Jundong Li, Yasin N Silva, Deborah L Hall, and Huan Liu. 2019. XBully: Cyberbullying Detection within a Multi-Modal Context. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 339–347.

[9] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*. Springer, 693–696.

[10] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International AAAI Conference on Web and Social Media*.

[11] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. *The Social Mobile Web* 11, 02 (2011), 11–17.

[12] Maeve Duggan. 2017. Online harassment 2017. http://www.pewinternet.org/2014/10/22/online-harassment/.

[13] Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*. ACM, New York, NY, USA, 109–117.

[14] FrontGate Media. 2014. A list of 723 bad words to blacklist & how to use Facebook's moderation tool. https://www.frontgatemedia.com/a-list-of-723-bad-words-to-blacklist-and-how-to-use-facebooks-moderation-tool/.

[15] CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media*.

[16] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering* 10, 4 (2015), 215–230.

[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. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 229–233.

[18] Leam Hackett. 2017. The Annual Bullying Survey 2017. https://www.ditchthelabel.org/wp-content/uploads/2017/07/The-Annual-Bullying-Survey-2017-1.pdf.

[19] Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.* 21, 9 (Sept. 2009), 1263–1284.

[20] Sameer Hinduja and Justin W Patchin. 2010. Bullying, cyberbullying, and suicide. *Archives of suicide research* 14, 3 (2010), 206–221.

[21] Homa Hosseinmardi, Amir Ghasemianlangroodi, Richard Han, Qin Lv, and Shivakant Mishra. 2014. Towards understanding cyberbullying behavior in a semi-anonymous social network. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 244–252.

[22] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Prediction of cyberbullying incidents in a media-based social network. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 186–192.

[23] April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. 2013. Detecting cyberbullying: Query terms and techniques. In *Proceedings of the 5th Annual ACM Web Science Conference (WebSci '13)*. ACM, New York, NY, USA, 195–204.

[24] Robin M Kowalski, Susan P Limber, Sue Limber, and Patricia W Agatston. 2012. *Cyberbullying: Bullying in the digital age*. John Wiley & Sons.

[25] Srijan Kumar, Justin Cheng, and Jure Leskovec. 2017. Antisocial behavior on the Web: Characterization and detection. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 947–950.

[26] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the presence and intensity of hostility on Instagram using linguistic and social features. In *Twelfth International AAAI Conference on Web and Social Media*.

[27] Yurii Nesterov. 2013. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media.

[28] 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.

[29] Alexandra Olteanu, Kartik Talamadupula, and Kush R Varshney. 2017. The limits of abstract evaluation metrics: The case of hate speech detection. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 405–406.

[30] Zizi Papacharissi. 2009. The virtual geographies of social networks: a comparative analysis of Facebook, LinkedIn and A SmallWorld. *New media & society* 11, 1-2 (2009), 199–220.

[31] David Martin Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (2011).

[32] 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 (SAC '18)*. ACM, New York, NY, USA, 1738–1747.

[33] Elaheh Raisi and Bert Huang. 2018. Weakly supervised cyberbullying detection with participant-vocabulary consistency. *Social Network Analysis and Mining* 8, 1 (2018), 38.

[34] Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L Shalin, and Amit Sheth. 2018. A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, 33–36.

[35] H Rosa, N Pereira, R Ribeiro, PC Ferreira, JP Carvalho, S Oliveira, L Coheur, P Paulino, AM Veiga Simão, and I Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior* 93 (2019), 333–345.

[36] Semiu Salawu, Yulan He, and Joanna Lumsden. 2017. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing* 1 (2017), 1–1.

[37] Robert Slonje and Peter K Smith. 2008. Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology* 49, 2 (2008), 147–154.

[38] Sara Owsley Sood, Elizabeth F Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology* 63, 2 (2012), 270–285.

[39] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin. 2015. Identification and Characterization of Cyberbullying Dynamics in an Online Social Network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15)*. ACM, New York, NY, USA, 280–285.

[40] Robert George Douglas Steel and James Hiram Torrie. 1960. *Principles and procedures of statistics: with special reference to the biological sciences*. McGraw-Hill.

[41] Andreas S Weigend. 2018. *Time series prediction: forecasting the future and understanding the past*. Routledge.

[42] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 656–666.

[43] 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.

[44] Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. 2019. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *Proceedings of the International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

[45] Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68, 1 (2006), 49–67.

[46] Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. 2016. Content-driven detection of cyberbullying on the Instagram social network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 3952–3958.