Future Generation Computer Systems 118 (2021) 219–229

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Early detection of cyberbullying on social media networks

Manuel F. López-Vizcaíno, Francisco J. Nóvoa, Victor Carneiro, Fidel Cacheda [*]

CITIC Research Center, Computer Science and Information Technologies Department, University of A Coruña, Campus de Elviña, A Coruña, 15071, Spain

ARTICLE INFO

Article history:
Received 15 June 2020
Received in revised form 9 November 2020
Accepted 8 January 2021
Available online 13 January 2021

Keywords:
Cyberbullying
Social networks
Early detection
Machine learning

ABSTRACT

Cyberbullying is an important issue for our society and has a major negative effect on the victims, that can be highly damaging due to the frequency and high propagation provided by Information Technologies. Therefore, the early detection of cyberbullying in social networks becomes crucial to mitigate the impact on the victims. In this article, we aim to explore different approaches that take into account the time in the detection of cyberbullying in social networks. We follow a supervised learning method with two different specific early detection models, named threshold and dual. The former follows a more simple approach, while the latter requires two machine learning models. To the best of our knowledge, this is the first attempt to investigate the early detection of cyberbullying. We propose two groups of features and two early detection methods, specifically designed for this problem. We conduct an extensive evaluation using a real world dataset, following a time-aware evaluation that penalizes late detections. Our results show how we can improve baseline detection models up to 42%.

1. Introduction

Bullying can be defined as an aggressive, intentional act or behaviour that is carried out by a group or an individual against a victim who cannot easily defend him or herself repeatedly and over time [1]. When the aggression occurs using Information Technologies (IT), such as the Internet, we talk of cyberbullying [2].

Cyberbullying has been identified as a major issue [3] and has been documented as a national health problem [4] due to the continuous growth of online communication and the social media [5]. The percentages of individuals who have experienced cyberbullying at some point during their lifetime has doubled, with 18% in 2007 and 36% in 2019 according to [6], and this is only expected to continue raising taking the high use of IT, social networks and mobile devices by children and teenagers [7] into account.

The negative effects of cyberbullying share many similarities with traditional bullying [8] but, at the same time, it could be more damaging due to the frequency and high propagation allowed by technology [9]. Studies have linked cyberbullying with negative effects on psychological and physical health, academic performance [10,11], depression [12] and a higher risk of suicidal ideation [13,14].

Consequently, an early detection of cyberbullying in social networks is paramount to mitigate and reduce its negative effects on the victims. Moreover, the repetitive nature of cyberbullying makes it extremely important to detect and terminate, as soon as possible, the cyberagression to, on one side, identify the aggressors and, on the other side, support the victims.

In this work we aim to explore different approaches that do not only take into account, the appropriate detection of cyberbullying in social networks, but also the time required for the detection. In the literature there are multiple and diverse works that explore different approaches to detect cyberbullying in social networks but, to the best of our knowledge, this is the first attempt to investigate techniques specifically designed for an early detection of cyberbullying. We follow a supervised learning approach, as the majority of previous works, but focusing on two different specific early detection methods, named threshold and dual. The former follows a more simple approach, while the latter requires two machine learning models. We also propose two new groups of features, specifically designed for this problem. The first group is intended to capture textual similarities among comments using a Bag-of-Words (BoW) model, while the second group will capture repetitive time aspects on the comments. We conduct an extensive evaluation using a real world dataset and follow a time-aware approach that penalizes late detections. Our results show how the threshold model is able to improve baseline detection models by 26% and the dual model up to 42%.

The main contributions of this work could be summarized as:

- We define and characterize the cyberbullying early detection problem.

[*] Corresponding author.
E-mail addresses: manuel.fernandezl@udc.es (M.F. López-Vizcaíno), francisco.javier.novoa@udc.es (F.J. Nóvoa), victor.carneiro@udc.es (V. Carneiro), fidel.cacheda@udc.es (F. Cacheda).

https://doi.org/10.1016/j.future.2021.01.006

- We present two specific machine learning models (i.e. threshold and dual) for the cyberbullying early detection problem and we show the impact of two sets of features (i.e. BoW and time features) that contribute to improve the performance in the cyberbullying early detection problem.
- We carry out extensive experiments using a real world dataset and following a time-aware evaluation that proves the performance improvement over baselines.

The article is organized as follows: in Section 2 we present state-of-the-art on early phenomena and cyberbullying detection, with a special focus on social networks. Section 3 describes all the details of the cyberbullying early detection problem and Section 4 presents experimental evaluation and performance improvements over the baselines. Finally, Section 5 includes our conclusions and future work on this line of research.

2. Related works

2.1. Early phenomena detection on social networks

From a generic perspective, this work is related to the early detection of different phenomena or anomalies on social networks.

For example, over the last years, there has been a rising interest in the detection of fake news [15–20], rumours [21–24], misinformation [25–27] or fake profile detection [28–30] using the information published on social networks, but without considering the time required in the detection. In fact, the works that explore the early detection perspective are limited. For example, [31] explores the prediction of fake news before it has been propagated on social media. With this purpose, they propose a theory-driven model that represents the news content at four language levels (lexicon-level, syntax-level, semantic-level and discourse-level) achieving 88% accuracy and outperforming all baselines considered.

Qin et al. aim at improving early detection of rumours on social networks by using novelty based features that consider the increase presence of unconfirmed information in a message with respect to trusted sources of information [32]. Their experiments using data collected from Sina Weibo, a Chinese social media service, show that their proposed method performs significantly better in terms of effectiveness than other real-time and early detection baselines. Also, [33] presents a rumour detection approach by clustering tweets by their likelihood of actually containing a disputed factual claim. The authors include in the evaluation the earliness of detection by measuring how soon a method is able to detect a rumour assuming a batch processing of one hour.

Also recently, the workshop on early risk prediction on the Internet (eRisk) at the Conference and Labs for the Evaluation Forum (CLEF) has provided different challenges oriented to the problems of detecting depression, anorexia and self-harm on the Reddit social network [34]. One of the best performing methods for the early detection of depression employed linguistic metainformation extracted from the subjects' writings and developed a classifier using recurrent neural networks [35,36]. Alternatively, the model proposed in [37,38] uses a word-based approach that estimates risk based on different word statistics (within-class frequency, within-class significance and inter-class term significance), which obtained good results in the detection of depression and self-harm. We have also explored the early detection of depression on social media by developing specific learning models (e.g. singleton and dual), significantly improving previous works performance [39,40].

2.2. Cyberbullying detection

Cyberbullying detection has been explored quite extensively, starting with user studies from social sciences and psychology fields, and more recently, moving to computer science, aiming at developing models for its automatic detection.

Al-Garadi et al. in [5] presented an extensive analysis of cyberbullying prediction models on social media and point at some open challenges such as the prediction of cyberbullying severity, human data characteristics or language dynamics. There are multiple studies that explore different machine learning alternatives in the detection of cyberbullying. In [41], the authors explore the use of Support Vector Machines (SVM) and a lexical syntactical feature to predict offencive language achieving high precision values. Dadvar et al. [42] use labelled text instances to train a SVM model for creating a gender specific cyberbullying detection system that improves the discrimination capacity. In [43], the authors also use SVM to predict cyberbullying in Ask.fm social network. They present a new scheme annotation to describe the severity of cyberbullying and conclude that the detection of fine-grained categories (e.g. threats) is more challenging due to data sparsity.

In [44] the authors work on the detection of cyberbullying on a multimodal social media environment, identifying several features, non only textual, but also audio and visual. Their results suggest that audio–visual features can help improve the performance of purely textual cyberbullying detectors.

Van et al. also investigate the automatic detection of cyberbullying-related posts on social media [45], both in English and Dutch. Using six natural language processing features groups (word n-grams, subjectivity lexicons, character n-grams, term lists and topic models), the proposed model outperform the baseline considered and identify false positives on implicit cyberbullying or offences through irony.

The authors of [46,47] study the detection of cyberbullying in the Vine social network using several machine learning models, achieving the best results with AdaBoost, closely followed by Random Forest. Hosseinmardi et al. [48] work on the detection of cyberbullying incidents on the Instagram social network. They use Naïve Bayes and SVM classifiers, with the latter obtaining the best performance by incorporating multi-modal text and images features as well as media session data. Some other works focus on the features considered to detect cyberbullying, for example, by analysing the social network structure among users [49,50], combining text and images analysis techniques [51], profanity features [47,52–54], sentiment analysis [43,55–57] or location features [58], among others.

An extensive review of published research on automatic detection of cyberbullying can be found in [59] and [60]. From a global point of view, most works employ textual features, followed by sentiment attributes. User features (e.g. age, gender) and social features (e.g. number of friends or followers) are also commonly considered. Interestingly, few works incorporate temporal features into their models, such as [61,62], in order to capture the temporal and repetitive aspects of cyberbullying. Cheng et al. [61] incorporate a time interval prediction into their prediction model outperforming state-of-the-art models in terms of F1 and Area Under the Curve (AUC). In [62], the authors model the temporal dynamics of cyberbullying in online sessions and show how the inclusion of these temporal features increases the performance of cyberbullying detection.

However, all previous works measure their performance regarding how successfully the model can distinguish between cyberbullying and non-cyberbullying cases using standard evaluation metrics, such as, accuracy, precision, recall, F-measure or area under the curve [5,59,60], without taking the time required to produce the prediction into account. In this sense, to the best of our knowledge, our work is the first attempt to measure the cyberbullying detection performance taking into account, not only the accuracy of the system, but also the time required for the prediction.

3. Cyberbullying early detection

The problem of cyberbullying early detection on social networks can be considered different to cyberbullying prediction. In this case, there is a set of social media sessions, that we denote as S, where some may correspond to cyberbully aggressions. We define the social media sessions set as:

$$S = \{s_1, s_2, \ldots, s_{|S|}\}$$

where $|S|$ denotes the number of sessions and s_i refers to session i.

Each social media session, $s \in S$, is formed by a sequence of posts, denoted as P_s, and a binary indicator, b_s, that specifies whether this specific session is considered cyberbullying ($b_s = true$) or not ($b_s = false$). The sequence of posts for a specific session will change throughout time and is given by:

$$P_s = (\langle P_1^s, t_1^s \rangle, \langle P_2^s, t_2^s \rangle, \ldots, \langle P_n^s, t_n^s \rangle)$$

where the tuple $\langle P_k^s, t_k^s \rangle, k \in [1, n]$ represents the kth post for social media session s and t_k^s is the timestamp when post P_k^s was published.

At the same time, each post P_k^s is specified by a vector of features:

$$P_k^s = \left[f_{k_1}^s, f_{k_2}^s, \ldots, f_{k_m}^s \right], k \in [1, n]$$

Given a social media session s, the objective is to detect if the session corresponds to cyberbullying but processing as few posts from P_s as possible. Therefore, our target is to learn a function $f(b_{s_i} | s_i, \langle P_1^{s_i}, t_1^{s_i} \rangle, \ldots, \langle P_k^{s_k}, t_k^{s_k} \rangle)$ to predict whether a session is cyberbullying or not.

In this sense, the function will receive as input, posts from 1 to k, and it will return three possible values: $\{0, 1, 2\}$, following the methodology proposed at [63]. Where 0 corresponds to a session s that is considered normal (i.e. non-cyberbullying), 1 if it is considered cyberbullying and 2 if no definitive decision can be emitted for session s after processing k posts and more posts must be read (i.e. *delay*).

3.1. Dataset

To study the cyberbullying early detection problem we will use a public dataset collected from the Vine social network [46,47]. The dataset was collected using a snowball sampling method, where an initial user u is selected as a seed and then the collection continues with the users following $'u'$. The authors provide a detailed study to ensure the representativeness of the social network.

For each user, a standard information was collected (i.e. user name, full name, profile description, number of followers, number of videos posted, number of followings) and all videos posted along with their comments, number of likes and number of reposts. A social media session is composed of a posted video along with all the likes and comments associated. Sessions with less than 15 comments have been removed from the dataset by the authors in order to have a sufficient number of comments [47]. A total of 961 sessions were labelled as cyberbullying or normal using crowdsourcing and, following [47], we required at least 60% confidence from the labellers to be considered cyberbullying.

Table 1 shows the main statistics for the dataset. In our case, each comment is considered a post for a social media session and, instead of aggregating all comments for a session [46,47], we will work with each comment individually and determine, processing as few comments from a session as possible, if the session can be considered cyberbullying or normal. Note that operating with individual comments will allow us to easily aggregate comments up to a certain point, i.e. k.

Table 1
Dataset statistics.

	Cyberbullying	Normal	Total
Media sessions	190	771	961
Comments	16,332	61,129	77,461
Comments/session	85.96	79.29	80.61
Likes	261,009	1,667,943	1,928,952
Likes/session	1373.73	2163.35	2007.23
Average followers/user	132,299	188,413	176,611
Average following/user	4759	1971	2557
Average time span (s)	210	240	234

3.2. Features

For our experiments we start with the features that provided the best results in [47]. These features are grouped by:

- Profile owner features: that capture the characteristics of the user who posted the initial video. These features include numbers of followers and following, polarity and subjectivity of the user's profile description.
- Media session features: number of likes, comments and sharing and polarity and subjectivity of media caption.
- Comment features: that are intended to determine the negativity associated with the comment. These features include percentage of negative comments, profane words in the comment, average polarity and subjectivity for the comments, differentiating between owner and other comments.
- Video features: intended to capture the nature of the video, these features validate the emotions and content in the video.
- LDA features: top ten topics extracted using Latent Dirichlet Allocation from all comments.

We also further extend the features with two new group of characteristics that we consider may be relevant for the cyberbullying early detection problem: Bag-of-Words (BoW) similarity and time aspects.

Following previous works, such as [4,41,43], we consider BoW similarity. In our case, we aim at computing these features without supervision. For this purpose, the training dataset is divided into two disjunctive sets: cyberbullying and non-cyberbullying sessions. The main goal of these features is to estimate the likeliness between a given comment versus cyberbullying and normal comments, without considering a set of predefined terms (e.g. profane words).

For each comment, we calculate the average, standard deviation, minimum, maximum and median of the TF–IDF (Term Frequency–Inverse Document Frequency) similarity obtained comparing this comment to every other cyberbullying comment. Then, the same process is repeated for the similarities with non-cyberbullying comments. In both cases, the active comment is removed from the corresponding sample.

Since cyberbullying implies a certain repetition over time, we consider relevant to include some features to capture different time aspects of the comments. This is pointed from the dataset statistics on Table 1, where we can observe a shorter time span for cyberbullying sessions, which is confirmed by a Welch two sample t-test with a p-value close to zero. Fig. 1 represents the

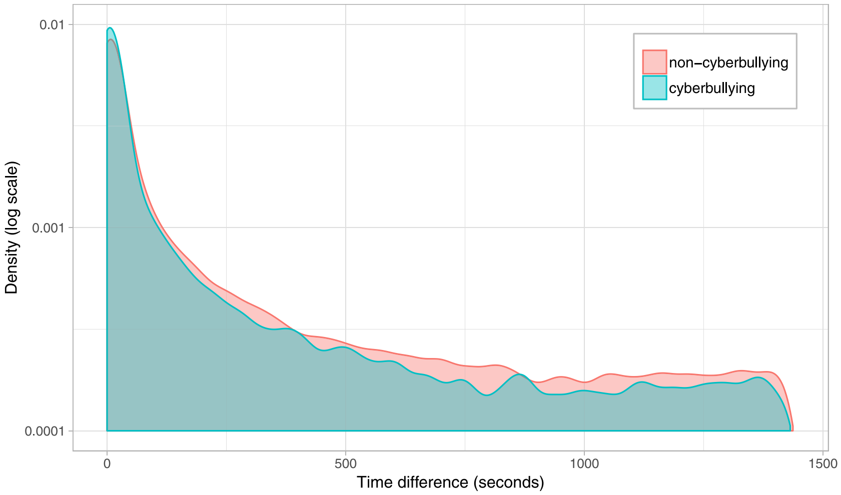


Fig. 1. Density plot for time difference between two consecutive comments for cyberbullying and non-cyberbulling for the Vine dataset. For each comment from the dataset, the time difference with the previous comment from the same media session was computed and the ground truth label was used for splitting between cyberbullying and non-cyberbullying. For the first comment of each media session, the difference is calculated with respect to the time when the video was posted.

time between two consecutive comments, both for cyberbullying and non-cyberbulling comments. From the figure we observe that there is a higher number of comments produced in a very short time span (a few seconds) for cyberbullying, and then, in the long tail, both types behave uniformly.

For this purpose, we include features that measure the time difference between two consecutive comments. We differentiate two cases: the time difference with the last comment and the time difference considering all previous comments. In both cases, we aggregate them by calculating the average, median, maximum and minimum values.

3.3. Baselines

As baselines we will consider the models reported on [47] that achieved the best performance results: Random Forest (*RF*), AdaBoost (*AB*), Extra Tree (*ET*), linear Support Vector Classification (*SVC*) and Logistic Regression (*LR*).

Since standard classification models provide a binary output, in order to predict if a session is considered cyberbullying or not (i.e. no delay result can be generated), we use a simple adaptation of the baselines considering a fixed number of input comments, where a delay is produced until this fixed number is reached. For example, if the number of input comments is fixed at 5, a delay will be produced for comments 1 to 4 and then a final decision will be emitted. For each baseline model we consider four pre-established input comments: 1, 5, 10 and 15. Therefore, for instance, the Random Forest model will be presented with 1, 5, 10 and 15 comments, and, in each case, a final decision will be generated, creating four early detection variants for the RF model, namely, RF_1, RF_5, RF_{10} and RF_{15}.

3.4. Early detection models

We also consider specific early detection models, that we adapt for the cyberbullying early detection problem. In particular, we adapt two early detection models that have reported good results [39,40]: threshold and dual.

Taking into account that this is not a classical binary classification problem because a non-final decision can be emitted, the threshold model is based on one learning model, which is trained using the features described in Section 3.2, and it integrates a decision function based on the class probabilities to determine if enough evidence is available to proceed with a firm decision. Initially, the decision function for the threshold model is defined as:

$$\delta_1(m, th_+(), th_-())$$

where m denotes a machine learning model used in (e.g. one of the baseline models), $th_+()$ is a threshold function used to set the limit of the class probability for a positive final decision and $th_-()$ is the threshold function to set the limit for a negative decision. The threshold model is an adaptation of the singleton model from [40] where the training and test sets are divided into ten homogeneous groups of posts (named batch or chunk processing), while in this case, each post is processed individually (named stream processing), following a more realistic evaluation approach.

We have explored different threshold functions, ranging from independent functions for positive and negative decisions to several decreasing functions depending on the number of posts processed. Contrary to [40], the best and most stable performance was obtained with a constant function of the form, $th() = \ell$, for both cases, positive and negative. This may be motivated because we provide a fine grain time evaluation, validating one comment at a time, opposed to the batch evaluation performed in [40]. In the experimental evaluation we analyse the performance for different values of ℓ.

We have also tested the order in which the threshold functions are applied to the class probabilities obtained by each model (i.e. first $th_+()$ and then $th_-()$, or vice versa), obtaining the same results for both cases. In our experiments, we will execute first the positive threshold and then the negative.

On the other hand, the objective of the dual model is to predict independently each option (i.e. positive and negative). In this sense, and inspired by multiclass classifiers one-versus-all, the dual model consists of two independent learning models, each one trained with an independent set of features. One model is trained to detect positive cases (denoted as m_+), while the other is trained to detect negative cases (denoted as m_-). Again, we adapt the proposal from [40] by defining a decision function of the form:

$$\delta_2(m_+, m_-, th_+(), th_-())$$

where, m_+ is the learning model responsible for positive predictions and m_- is the model in charge of negative predictions. As in the previous case, $th_+()$ and $th_-()$ are the threshold functions for positive and negative cases, which, in this case, are associated with their respective models, m_+ and m_-. Following the lead of the former model, we considered different constant functions for both thresholds.

Based on the baseline models, we have defined different threshold and dual model implementations that are expected to capture the special characteristics of the cyberbullying early detection problem in a better way than the standard baselines, since they have been specifically designed for the early detection problem.

3.5. Evaluation metrics

As evaluation metrics we will consider two metrics specifically designed for the early detection problem, although applied in a different environment: Early Risk Detection Error (*ERDE*) [63] and latency-weighted F1 [64]. In both cases, they were used to measure performance on the early detection of depression on individuals based on their posts on social networks.

The *ERDE* metric is measured at a specific time point, o (provided as a parameter), and for a session s after processing k comments (typically, because the model has required k comments to produce a final prediction) is defined as follows:

$$ERDE_o(s_i, k) = \begin{cases} \frac{\sum_{s_i \in S \wedge b_{s_i} = true} 1}{|S|} & \textit{if False Positive} \\ 1 & \textit{if False Negative} \\ 1 - \frac{1}{1+e^{k-o}} & \textit{if True Positive} \\ 0 & \textit{if True Negative} \end{cases}$$

In the case of wrong predictions (false positive or negative) the error will increase, in the former proportionally to the number of positive cases and in the latter by 1. A true negative will not increase the error, but a true positive may impact negatively if the number of posts required to make the prediction surpasses the measuring point o.

The latency-weighted F1, or in short $F_{latency}$ or *F1-latency*, is proposed by Sadeque et al. as an alternative to the ERDE metric, combining both latency and accuracy [64]. The F-latency metric is defined as:

$$F_{latency}(s_i, k) = F1 \cdot \left(1 - \underset{s_i \in S \wedge b_{s_i} = true}{median} \left(-1 + \frac{2}{1 + e^{-p(k-1)}}\right)\right)$$

where, p is a parameter that determines how quickly the penalty should increase, which is set to achieve 50% of latency penalty at the median number of items and $F1$ is the standard *F-measure* that is calculated as the harmonic mean of precision and recall.

Note how *ERDE* is an error measure and, therefore, values closer to 0 are better, while for $F_{latency}$ values closer to 1 are representative of good results.

Some limitations have been reported for the *ERDE* metric [65, 66] and so, we will rely more on $F_{latency}$. By default, we report results for *ERDE* at a low time point, $o = 5$, and p is set to 0.02288 for $F_{latency}$. Also, in the first experiments we will provide results for precision and recall, as complementary values.

4. Experimental evaluation

For evaluation purposes, we use 80% of the dataset for training and the remaining is used for testing. Note that the dataset has been divided by social media sessions and each session includes all its posts. The posts are presented to the models chronological and sequentially in order to make a prediction. When reported, confidence intervals are calculated at the 95% confidence level.

Table 2
Results for baseline models after processing 5 comments using individual groups of features. The best results for each group of features are highlighted for both *ERDE* and $F_{latency}$ and underlined for precision and recall.

	RF_5	AB_5	ET_5	SVC_5	LR_5
Profile owner features					
ERDE	0.1757	**0.1685**	0.1617	0.2054	0.1712
$F_{latency}$	0.1414	0.1363	0.2672	**0.2759**	0.1272
Precision	0.1905	0.3333	$\underline{0.4118}$	0.1805	0.2500
Recall	0.1212	0.0909	0.2121	$\underline{0.7273}$	0.0909
Media session features					
ERDE	0.1609	0.1746	**0.1599**	0.1967	0.1710
$F_{latency}$	**0.2881**	0.0465	0.2783	0.2196	0.0000
Precision	0.4000	0.1250	$\underline{0.4667}$	0.1625	0.0000
Recall	0.2424	0.0303	0.2121	$\underline{0.3939}$	0.0000
Comment features					
ERDE	0.1642	0.1694	**0.1556**	0.1627	0.1575
$F_{latency}$	0.2121	0.1332	**0.3368**	0.2776	0.3348
Precision	0.4167	0.3000	$\underline{0.5000}$	0.3636	0.4167
Recall	0.1515	0.0909	0.2727	0.2424	$\underline{0.3030}$
LDA features					
ERDE	**0.1633**	0.1738	0.1668	0.1686	0.1705
$F_{latency}$	0.1909	0.1193	0.1735	0.1660	**0.2045**
Precision	$\underline{0.5714}$	0.2000	0.3636	0.3077	0.2609
Recall	0.1212	0.0909	0.1212	0.1212	$\underline{0.1818}$
Video features					
ERDE	0.1728	**0.1719**	0.1728	**0.1719**	**0.1719**
$F_{latency}$	**0.0489**	0.0000	**0.0489**	0.0000	0.0000
Precision	$\underline{0.1667}$	0.0000	$\underline{0.1667}$	0.0000	0.0000
Recall	$\underline{0.0303}$	0.0000	$\underline{0.0303}$	0.0000	0.0000
BoW features					
ERDE	**0.1659**	0.1719	0.1685	0.1710	0.1710
$F_{latency}$	**0.1468**	0.0502	0.1363	0.0000	0.0000
Precision	$\underline{0.5000}$	0.2000	0.3333	0.0000	0.0000
Recall	$\underline{0.0909}$	0.0303	$\underline{0.0909}$	0.0000	0.0000
Time features					
ERDE	0.1730	0.1745	**0.1581**	0.1740	0.1710
$F_{latency}$	0.1497	0.0000	**0.2726**	0.1909	0.0000
Precision	0.2222	0.0000	$\underline{0.6667}$	0.2222	0.0000
Recall	0.1212	0.0000	$\underline{0.1818}$	$\underline{0.1818}$	0.0000

4.1. Baselines

In the first set of experiments, we validate the performance of the baseline models. We start by presenting on Tables 2 and 3 the results for baseline models after processing 5 comments.

We start our analysis by examining the behaviour of the features considered in this work in order to determine which can provide a better performance in the early detection of cyberbullying. Table 2 presents the results for the individual features in terms of *ERDE* and $F_{latency}$ as early detection metrics, and also precision and recall are presented for completeness.

The first five group of features correspond to those reported on [47] and we can observe that the best performance is obtained by comment features, both in terms of $F_{latency}$ (0.3368 ± 0.0048), and *ERDE* (0.1556±0.0037). This is motivated because comments concentrate the information that changes as the session advances, while the other features are shared for the whole session. Also note how the second-best performance is obtained with media session features (0.2881 ± 0.0046 and 0.1599 ± 0.0037, for $F_{latency}$ and *ERDE*, respectively), significantly below the comments features performance.

Regarding the features proposed, we observe that BoW features achieve modest results in terms of $F_{latency}$ and *ERDE*, somehow expected, since their capacity to identify similarities would be limited to textual equivalence, while time features reach reasonable results in terms of early detection, with no significant

Table 3
Results for baseline models after processing 5 comments using combinations of features. The best results for each combination are highlighted for both *ERDE* and $F_{latency}$ and underlined for precision and recall.

	RF_5	AB_5	ET_5	SVC_5	LR_5
Profile owner, media session, comment, LDA, video features					
ERDE	0.1650	0.1663	**0.1625**	0.1665	0.1711
$F_{latency}$	0.1507	0.2849	0.2219	**0.3141**	0.0516
Precision	0.6000	0.2941	0.5000	0.2826	0.2500
Recall	0.0909	0.3030	0.1515	0.3939	0.0303
+ BoW features					
ERDE	0.1641	0.1663	**0.1608**	0.1815	0.1711
$F_{latency}$	0.1547	**0.2849**	0.2545	0.2386	0.0516
Precision	0.7500	0.2941	0.5000	0.2000	0.2500
Recall	0.0909	0.3030	0.1818	0.3333	0.0303
+ Time features					
ERDE	0.1624	**0.1593**	0.1599	0.1980	0.1719
$F_{latency}$	0.1957	**0.3332**	0.2603	0.2688	0.0502
Precision	0.6667	0.3667	0.5455	0.1835	0.2000
Recall	0.1212	0.3333	0.1818	0.6061	0.0303
+ BoW + Time features					
ERDE	0.1615	0.1593	**0.1573**	0.1908	0.1719
$F_{latency}$	0.2009	**0.3332**	0.2969	0.2726	0.0502
Precision	0.8000	0.3667	0.5833	0.1935	0.2000
Recall	0.1212	0.3333	0.2121	0.5455	0.0303

difference in terms of *ERDE* with the comment features performance.

However, it is interesting to note that BoW and time features obtain good precision scores: the latter obtains the best score and the former is second-best, tied with comment features. In the case of BoW features, we consider that it is due to the fact that there are terms that are repeated on multiple cyberbullying comments and these features will identify them, although a high number of false negatives is generated (note the low recall value) because the same terms will appear on normal comments. Regarding the time features, the high precision confirms our intuition from Section 3.2 suggesting that time difference between cyberbullying comments tends to be shorter but, at the same time, some normal comments will present the same characteristic, producing low recall values. In summary, the features proposed do not outperform baseline features in terms of early detection, but they are expected to be a good complement, as we will discuss later.

Table 3 presents the results obtained when combining features. We start by combining all features from [47], that constitutes our starting point. The best performing model is linear SVC, improving precision and recall over individual features, but not outperforming $F_{latency}$ nor *ERDE* from individual comment features (0.3141 ± 0.0047 and 0.1665 ± 0.0038, respectively).

When incorporating, BoW and time features individually, the performance increases for RF, AB and ET with respect to the baseline, while it decreases for SVC and LR, suggesting that models based on the data space underperform in this problem. Analysing each feature individually, we observe that the bigger improvement is obtained by time features, pointing towards reiteration as an important characteristic in the early detection of cyberbullying.

When adding both features, $F_{latency}$ and *ERDE* further improve for RF and ET with respect to individual features. However, the best $F_{latency}$ score remains in AB_5 and, although being close, it does not improve the comment feature performance on Table 2, despite there is no significant difference among them (confidence interval (0.3284, 0.3380) for AB_5).

If we focus on the models, Random Forest and Extra Trees obtain consistently the best values in terms of $F_{latency}$ for individual features (Table 2), with ET_5 being the best performing method just using comment features. On the other side, AdaBoost and Linear Regression are the methods with a lower performance. These results contrast with [47], where AdaBoost was the best performing model, which may be due to its sensitivity to the noise and outliers [67,68] that may arise when dealing with little information. When combining features (Table 3), Extra Tree and AdaBoost are the best performing models for the early detection metrics. We consider that the motivation for AB performance improvement may be due to the incorporation of more information by the combination of features.

To further study the time impact in the models performance, we analyse the performance of the different models with respect to the number of posts processed.

Fig. 2 presents one graph for each feature group from Tables 2 and 3 with $F_{latency}$ score for all models. For the sake of presentation we have removed video feature graph from the figure as it provided the lowest performance and did not contribute to the discussion. The graph labelled "BS" represents the combination of owner profile, media session, comments, LDA and video features from [47] that constitutes our baseline features.

Interestingly, for the owner profile and media session features the negative impact in the performance as the time progresses is clear, with the performance for all models decreasing as more posts are processed, since these features do not change as new comments are processed. On the other hand, comment, LDA, BoW and time features, and their combinations, present a more heterogeneous behaviour as time progresses, since there is a dispute between the improvement obtained by providing more information to the models (i.e. more posts) versus the negative impact in the performance since the prediction is being delayed.

From the models perspective, Logistic Regression is providing the lower performance independently of features considered and, on the other side, Extra Tree, SVC and Random Forest tend to be on the upper section of the different graphs for all feature combinations. In fact, the best score is obtained combining all features (i.e. BS+BoW+Time) by ET_{15} with 0.3657 ± 0.0049, performing significantly better than the same model with only comment features (Table 2).

Regarding the *ERDE* metric, we skip the graphical results since they basically mimic the previous figure from an error metric perspective. *ERDE* values are more concentrated and differences are more difficult to spot but, again, Extra Tree model is performing consistently better than other models for most features, closely followed by Random Forest in many cases. Again, the best performance is achieved by ET_{15}, with a score of 0.1477 ± 0.0036.

To provide a better understanding of the models performance and to compare both metrics employed, we present a box-plot for $F_{latency}$ and *ERDE* on Fig. 3. From the figure we confirm that Extra Tree is, for both metrics, the best performing model. Regarding $F_{latency}$ (Fig. 3(a)), SVC is second-best, with no significant difference with ET, however it is the worst performing on the *ERDE* metric (Fig. 3(b)). Also note how the differences for *ERDE* scores are small, which makes $F_{latency}$ a better option in terms of sensitivity.

Therefore, in the remaining experiments we will focus on the best performing model, that is, Extra Tree, and we will report results only on $F_{latency}$ for sake of simplicity. We have run all experiments including all models, but the other models kept providing a lower performance. As baseline and value to beat, we consider the best score achieved in these experiments corresponding to ET_{15} with 0.3657 ± 0.0049.

4.2. Early detection models

In these experiments we test the performance of the early detection models. We start with the threshold model for Extra

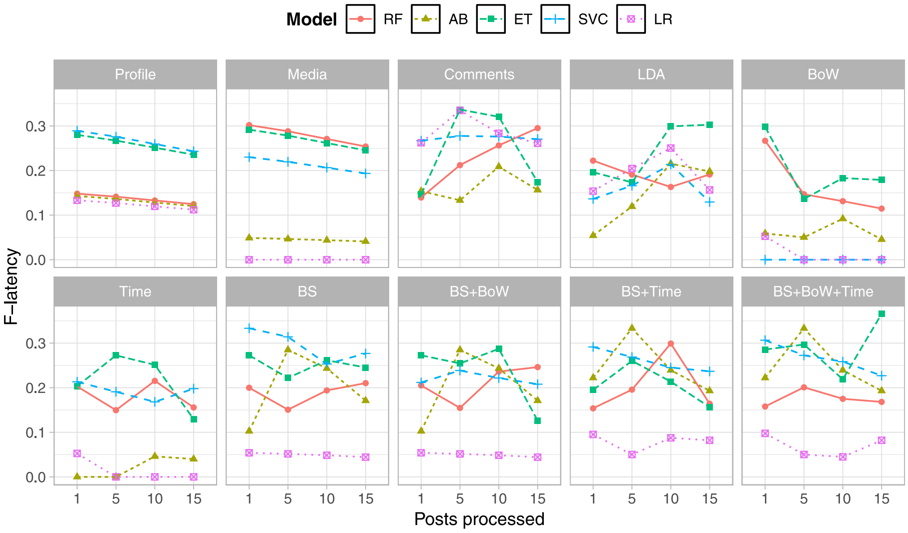
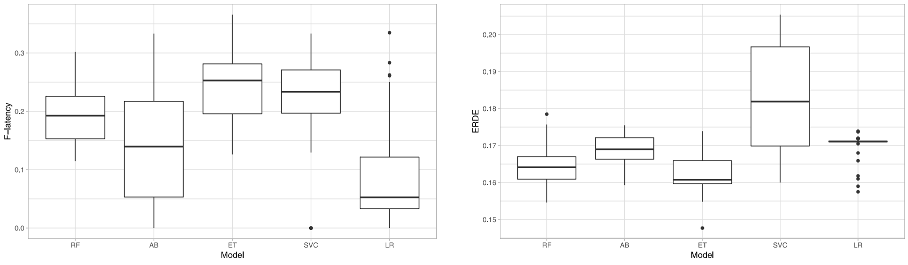


Fig. 2. $F_{latency}$ for all models. One graph is included for each feature and their combinations. BS graph refers to the combination of Profile, Media, Comments, LDA and Video features.



(a) $F_{latency}$

(b) $ERDE$

Fig. 3. Box-plots of the performance according to the model ((a) corresponds to $F_{latency}$ and (b) to *ERDE*). Performance is computed for all features and all number of posts processed. Lower and upper box boundaries are first and third quartile, respectively. Outliers correspond to data falling outside 1.5 times the interquartile range (IQR).

Tree, setting $th_{+}() = \ell_{+}$ and $th_{-}() = \ell_{-}$ and we test different values for the ℓ_{+} and ℓ_{-}, ranging from 0.9 to 0.5. Regarding the features, we start with the baseline features (i.e. Profile owner, media session, comment, LDA and video features), and we complete the results incorporating the features proposed (i.e. BoW and Time) both individually and combined. Fig. 4 presents the results for the threshold model.

In this case, for each social media session, the number of posts required to produce a final result (i.e. positive or negative) will vary because the class probability must be higher than the threshold and, therefore, the posts processed will vary for each session.

Focusing on the threshold values, we observe that, consistently, the best values are achieved when $th_{+}() = 0.5$ and, as the negative threshold (i.e. $th_{-}()$) increases, so does the performance up to 0.8 (decreasing at 0.9). In fact, the top score, 0.4615 ± 0.0051, is obtained by ET with $th_{+}() = 0.5$ and $th_{-}() = 0.8$ including all features (BS+BoW+Time), improving the baseline performance from the previous experiment (Fig. 2) by 26%. Focusing on the features, we observe that the group including all features is consistently providing the best scores (i.e. purple line steadily on top), which confirms the importance of the features proposed (i.e. BoW and Time features), along with the basic features.

In the final set of experiments, we study the performance of the dual model. For this purpose, we set initially $th_{+}() = 0.5$ and $th_{-}() = 0.8$, and present the results on Table 4 for Extra Tree. Since the dual variant requires two independent models we present a grid, where rows correspond to the features used by the negative model (best value for each row is highlighted) while columns represent the features employed by the positive model (best value for each column is underlined).

Firstly, we observe that best results are obtained when using all features for the positive model (i.e. last column), confirming the same tendency from the previous experiment. In general terms, when the positive model uses the baseline features in combination with our proposed features (i.e. last four columns on Table 4) the highest concentration of $F_{latency}$ values are obtained for the dual model based on Extra Tree. This confirms the importance of considering significant features for the positive model in order to discriminate cyberbullying sessions correctly.

It is also interesting to observe how the use of simple features on the negative model (e.g. Profile owner), leads to the best score (0.472 ± 0.0051), something that was already observed in previous works [40]. This may be motivated by the fact that multiple characteristics are required to determine if a social media session corresponds to cyberbullying, while a non-cyberbullying session can be decided in a much simpler way, using less information, since it has already been discarded as cyberbullying.

Finally, we have further explored the performance of the dual model by analysing combinations of best performing positive model features (i.e. BS, BS+BoW, BS+Time, BS+BoW+Time) with all negative model features, for different values of $th_{+}()$ and $th_{-}()$. On Fig. 5 we show $F_{latency}$ scores only for Extra Tree.

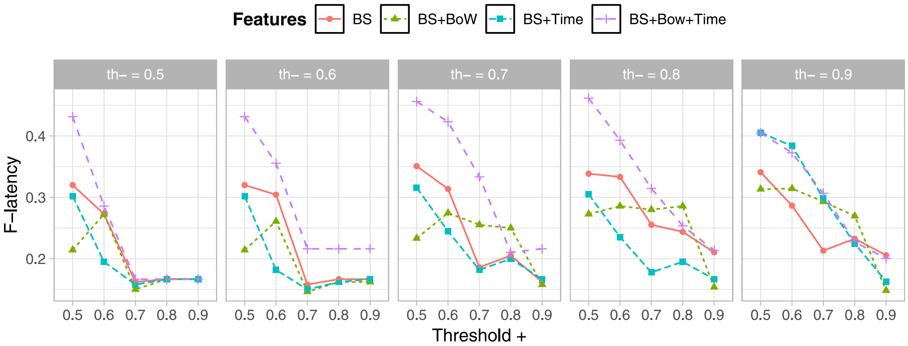


Fig. 4. $F_{latency}$ for threshold model using and Extra Tree. One graph is included for each value of $th_{-}() = \ell_{-}$. The X axis represents the values of $th_{+}() = \ell_{+}$. BS refers to the combination of Profile, Media, Comments, LDA and Video features.

Table 4
Results for $F_{latency}$ for dual models based on Extra Tree, with $th_{+}() = 0.5$ and $th_{-}() = 0.8$. The best value for each row (features negative model) is highlighted. The best value for each column (features positive model) is underlined. PO: Profile owner, MS: Media session, C: Comment, V: Video, BS: Profile owner + Media session + Comments + LDA + Video features, All: BS + BoW + Time.

	PO	MS	C	LDA	V	BoW	Time	BS	+BoW	+Time	All
PO	0.364	0.302	0.392	0.362	0.049	0.342	0.336	0.357	0.400	0.437	**0.472**
MS	0.364	0.302	0.379	0.360	0.049	0.359	0.336	0.437	0.372	**0.442**	0.434
C	0.364	0.302	0.200	0.200	0.051	0.257	0.187	0.302	0.241	0.345	**0.436**
LDA	0.364	0.302	0.211	0.219	0.051	0.250	0.208	0.286	0.273	0.316	**0.386**
V	0.364	0.302	0.374	0.293	0.049	0.358	0.282	0.333	0.327	**0.404**	**0.404**
BoW	0.364	0.302	0.310	0.222	0.051	0.243	0.200	0.339	0.262	0.333	**0.400**
Time	0.364	0.302	0.262	0.240	0.051	0.274	0.208	0.286	0.254	0.316	**0.393**
BS	0.364	0.302	0.282	0.260	0.051	0.279	0.227	0.339	0.316	0.343	**0.348**
+BoW	0.364	0.302	0.329	0.302	0.051	0.306	0.276	0.353	0.273	0.382	**0.412**
+Time	0.364	0.302	0.222	0.231	0.051	0.260	0.212	0.313	0.215	0.305	**0.388**
All	0.364	0.302	0.358	0.329	0.051	0.329	0.276	0.448	0.371	0.394	**0.462**

From the figure we observe that lower values of th_{+} keep concentrating the highest scores. In fact, the best performance is obtained with $th_{+}() = 0.5$ and $th_{-}() = 0.6$ using all features for the positive model and profile owner features for the negative model. This configuration achieves $F_{latency} = 0.5217$ (confidence interval $(0.5166, 0.5268)$), significantly improving baseline performance from Fig. 2 by 42% and best threshold model by 13%.

Interestingly, the top five configurations use all features for the positive model and profile owner features for the negative model, with different variations of threshold configurations (note the upper right graph from Fig. 5). This corroborates findings from previous experiments (Table 4), and confirms that an independent feature of the social media session, such as the owner characteristics, is relevant for identification of non-cyberbullying sessions, while the classification as cyberbullying requires specific session characteristics (e.g. comment features, BoW or time).

Focusing on the threshold values, best performing configurations set $th_{+}()$ always on the low side (i.e. values 0.5 and 0.6), while $th_{-}()$ takes values from the whole range. This, in combination with the use of all features for the positive model, suggest the importance of defining a positive model highly capable of properly detecting cyberbullying cases, with low class probabilities to reduce detection time and, hence, requiring a low threshold. Meanwhile, the negative model relies mainly on the use of simple features to accurately detect negative cases, once they have been discarded as positive.

5. Conclusions

In this paper, we introduced the cyberbullying early detection problem and we proposed two feature groups, specifically designed for this problem: text similarities and time features. Moreover, we have also adapted two specific machine learning models, threshold and dual, and verified their behaviour in our evaluation.

The experimental evaluation was based on a real world dataset from the Vine social network and we used specific time-aware metrics (i.e. *ERDE* and $F_{latency}$). Our results show how the threshold model is able to significantly improve the baseline detection models by 26% and the dual model is able to further increase this improvement up to 42%, in both cases using the Extra Tree model as basis. Moreover, the combination of proposed features along with baseline features (i.e. profile owner, media session, comment, LDA and video features) lead to the best performance for both, threshold and dual models.

As a main conclusion, the dual model consistently provides the best performance for the early detection of cyberbullying, based on the use of all features for the identification of positive cases along with low thresholds to produce early detections, and simpler features (i.e. profile owner characteristics) for the negative model.

In the near future, we expect to extend this research in several ways. First, we would like to explore heterogeneous combinations of different machine learning models on the dual model. For example, Extra Tree for the positive model, while Random Forest for the negative model. Second, we plan to further extend the features regarding comments, as these concentrate most of the information for the early detection. Third, we would like to investigate an evaluation based on time, instead of number of posts, since it may be relevant for the early detection of cyberbullying. Finally, we intend to experiment with other datasets from some other social media platforms to validate our approach and generalize the results.

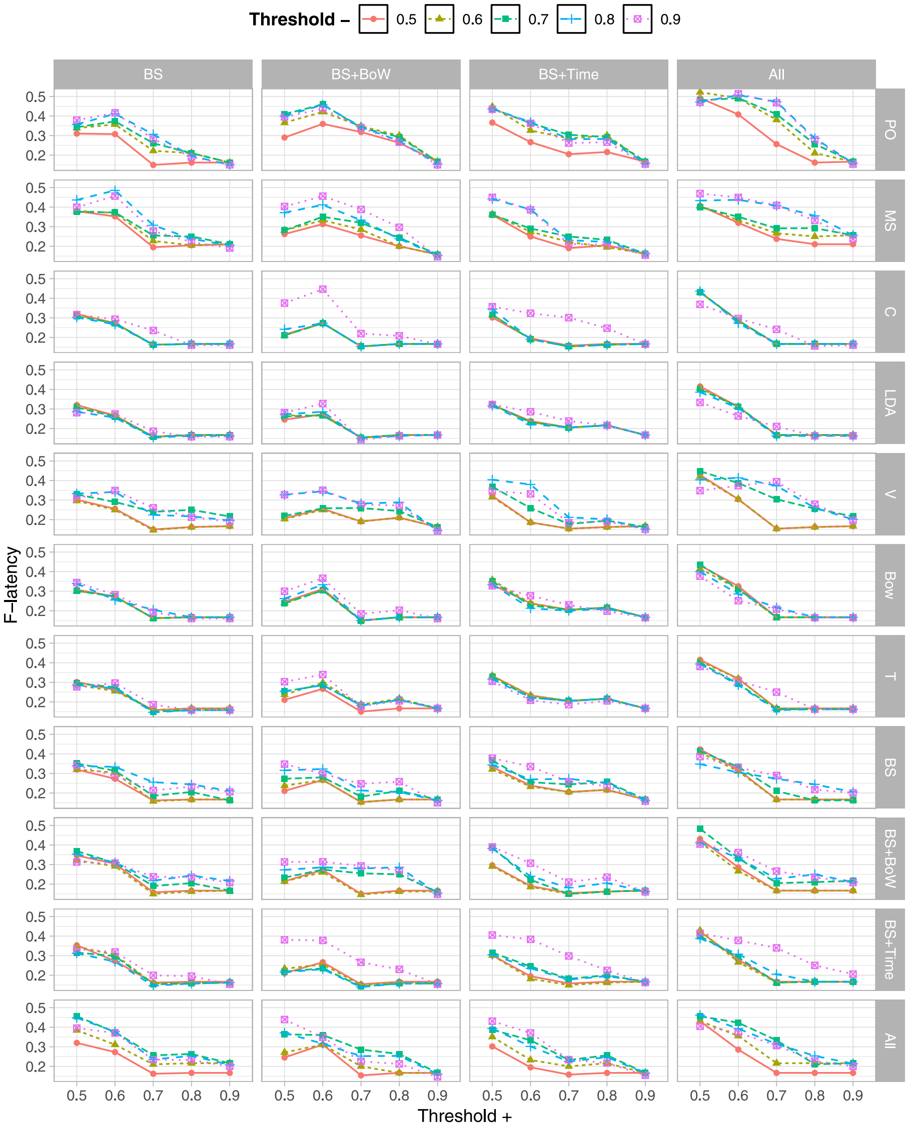


Fig. 5. $F_{latency}$ for dual model using Extra Tree. Columns represent positive model features (i.e. BS, BS+BoW, BS+Time, BS+BoW+Time), while rows represent negative model features. The X axis represents the values of $th_{+}() = \ell_{+}$, while one line is included for each value of $th_{-}() = \ell_{-}$. BS refers to the combination of Profile, Media, Comments, LDA and Video features.

CRediT authorship contribution statement

Manuel F. López-Vizcaíno: Data-curation, Methodology, Software, Writing- reviewing and editing. **Francisco J. Nóvoa:** Data-curation, Investigation, Writing - reviewing and editing. **Victor Carneiro:** Project administration, Writing - reviewing and editing. **Fidel Cacheda:** Conceptualization, Investigation, Writing - original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the Ministry of Economy and Competitiveness of Spain and FEDER funds of the European Union (Project PID2019-111388GB-I00) and by the Centro de Investigación de Galicia "CITIC", funded by Xunta de Galicia (Galicia, Spain) and the European Union (European Regional Development Fund — Galicia 2014–2020 Program), by grant ED431G 2019/01.

References

[1] D. Olweus, Bullying at school, in: Aggressive Behavior, Springer, 1994, pp. 97–130.

[2] R. Slonje, P.K. Smith, Cyberbullying: Another main type of bullying? Scand. J. Psychol. 49 (2) (2008) 147–154.

[3] G.S. O'Keeffe, K. Clarke-Pearson, et al., The impact of social media on children, adolescents, and families, Pediatrics 127 (4) (2011) 800–804.

[4] J.-M. Xu, K.-S. Jun, X. Zhu, A. Bellmore, 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, Association for Computational Linguistics, 2012, pp. 656–666.

[5] M.A. Al-Garadi, M.R. Hussain, N. Khan, G. Murtaza, H.F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H.A. Khattak, A. Gani, Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges, IEEE Access 7 (2019) 70701–70718.

[6] J.W. Patchin, Summary of our cyberbullying research, 2019, accessed March 10, 2020. URL https://cyberbullying.org/summary-of-our-cyberbullying-research.

[7] S. Hinduja, J.W. Patchin, Cyberbullying Fact Sheet: Identification, Prevention, and Response, Cyberbullying Research Center, 2020, pp. 1–9, accessed March 10.

[8] R.S. Tokunaga, Following you home from school: A critical review and synthesis of research on cyberbullying victimization, Comput. Hum. Behav. 26 (3) (2010) 277–287.

[9] I. Aoyama, T.F. Saxon, D.D. Fearon, Internalizing problems among cyberbullying victims and moderator effects of friendship quality, Multicult. Educ. Technol. J. 5 (2) (2011) 92–105.

[10] R.M. Kowalski, S.P. Limber, Psychological, physical, and academic correlates of cyberbullying and traditional bullying, J. Adolesc. Health 53 (1) (2013) S13–S20.

[11] A.T. Khine, Y.M. Saw, Z.Y. Htut, C.T. Khaing, H.Z. Soe, K.K. Swe, T. Thike, H. Htet, T.N. Saw, S.M. Cho, et al., Assessing risk factors and impact of cyberbullying victimization among university students in myanmar: A cross-sectional study, PLoS One 15 (1) (2020) e0227051.

[12] S. Rathore, P.K. Sharma, V. Loia, Y.-S. Jeong, J.H. Park, Social network security: Issues, challenges, threats, and solutions, Inf. Sci. 421 (2017) 43–69.

[13] H. Sampasa-Kanyinga, P. Roumeliotis, H. Xu, Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among canadian schoolchildren, PLoS One 9 (7) (2014) e102145.

[14] S. Hinduja, J.W. Patchin, Bullying, cyberbullying, and suicide, Arch. Suicide Res. 14 (3) (2010) 206–221.

[15] S. Kumar, N. Shah, False information on web and social media: A survey, arXiv preprint arXiv:1804.08559.

[16] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explor. Newsl. 19 (1) (2017) 22–36.

[17] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, Y. Liu, Combating fake news: A survey on identification and mitigation techniques, ACM Trans. Intell. Syst. Technol. (TIST) 10 (3) (2019) 1–42.

[18] C. Janze, M. Risius, Automatic detection of fake news on social media platforms, in: PACIS, 2017, p. 261.

[19] C. Buntain, J. Golbeck, Automatically identifying fake news in popular twitter threads, in: 2017 IEEE International Conference on Smart Cloud (SmartCloud), IEEE, 2017, pp. 208–215.

[20] M. Aldwairi, A. Alwahedi, Detecting fake news in social media networks, Procedia Comput. Sci. 141 (2018) 215–222.

[21] C. Andrews, E. Fichet, Y. Ding, E.S. Spiro, K. Starbird, Keeping up with the tweet-dashians: The impact of 'official' accounts on online rumoring, in: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, 2016, pp. 452–465.

[22] A. Arif, K. Shanahan, F.-J. Chou, Y. Dosouto, K. Starbird, E.S. Spiro, How information snowballs: Exploring the role of exposure in online rumor propagation, in: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, 2016, pp. 466–477.

[23] J. Ma, W. Gao, K.-F. Wong, Rumor Detection on Twitter with Tree-Structured Recursive Neural Networks, Association for Computational Linguistics, 2018.

[24] S.A. Alkhodair, S.H. Ding, B.C. Fung, J. Liu, Detecting breaking news rumors of emerging topics in social media, Inf. Process. Manage. 57 (2) (2020) 102018.

[25] V. Qazvinian, E. Rosengren, D.R. Radev, Q. Mei, Rumor has it: Identifying misinformation in microblogs, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 1589–1599.

[26] S.D. Bhattacharjee, W.J. Tolone, V.S. Paranjape, Identifying malicious social media contents using multi-view context-aware active learning, Future Gener. Comput. Syst. 100 (2019) 365–379.

[27] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the 20th International Conference on World Wide Web, 2011, pp. 675–684.

[28] C. Cai, L. Li, D. Zeng, Detecting social bots by jointly modeling deep behavior and content information, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 1995–1998.

[29] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, J. Leskovec, Anyone can become a troll: Causes of trolling behavior in online discussions, in: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 2017, pp. 1217–1230.

[30] Z. Chu, S. Gianvecchio, H. Wang, S. Jajodia, Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE Trans. Dependable Secure Comput. 9 (6) (2012) 811–824.

[31] X. Zhou, A. Jain, V.V. Phoha, R. Zafarani, Fake news early detection: A theory-driven model, arXiv preprint arXiv:1904.11679.

[32] Y. Qin, D. Wurzer, V. Lavrenko, C. Tang, Spotting rumors via novelty detection, arXiv preprint arXiv:1611.06322.

[33] Z. Zhao, P. Resnick, Q. Mei, Enquiring minds: Early detection of rumors in social media from enquiry posts, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 1395–1405.

[34] D.E. Losada, F. Crestani, J. Parapar, erisk 2020: Self-harm and depression challenges, in: European Conference on Information Retrieval, Springer, 2020, pp. 557–563.

[35] M. Trotzek, S. Koitka, C.M. Friedrich, Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression, in: CLEF (Working Notes), 2017.

[36] M. Trotzek, S. Koitka, C.M. Friedrich, Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences, IEEE Trans. Knowl. Data Eng..

[37] M.P. Villegas, D.G. Funez, M.J.G. Ucelay, L.C. Cagnina, M.L. Errecalde, Lidic-unsl's participation at erisk 2017: Pilot task on early detection of depression, in: CLEF (Working Notes), 2017.

[38] S.G. Burdisso, M. Errecalde, M. Montes y Gómez, Unsl at erisk 2019: a unified approach for anorexia, self-harm and depression detection in social media, in: Working Notes of the Conference and Labs of the Evaluation Forum-CEUR Workshop Proceedings, Vol. 2380, 2019.

[39] F. Cacheda, D.F. Iglesias, F.J. Nóvoa, V. Carneiro, Analysis and experiments on early detection of depression, CLEF (Work. Notes) 2125 (2018) 1–11.

[40] F. Cacheda, D. Fernandez, F.J. Novoa, V. Carneiro, Early detection of depression: Social network analysis and random forest techniques, J. Med. Internet Res. 21 (6) (2019) e12554.

[41] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, IEEE, 2012, pp. 71–80.

[42] M. Dadvar, F.d. Jong, R. Ordelman, D. Trieschnigg, Improved cyberbullying detection using gender information, in: Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), University of Ghent, 2012.

[43] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, V. Hoste, Detection and fine-grained classification of cyberbullying events, in: International Conference Recent Advances in Natural Language Processing (RANLP), 2015, pp. 672–680.

[44] D. Soni, V.K. Singh, See no evil, hear no evil: Audio-visual-textual cyberbullying detection, Proc. ACM Hum.-Comput. Interact. 2 (CSCW) (2018) 1–26.

[45] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, V. Hoste, Automatic detection of cyberbullying in social media text, PLoS One 13 (10).

[46] R.I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, S.A. Mattson, Careful what you share in six seconds: Detecting cyberbullying instances in vine, in: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2015, pp. 617–622.

[47] R.I. Rafiq, H. Hosseinmardi, S.A. Mattson, R. Han, Q. Lv, S. Mishra, Analysis and detection of labeled cyberbullying instances in vine, a video-based social network, Soc. Netw. Anal. Min. 6 (1) (2016) 88.

[48] H. Hosseinmardi, S.A. Mattson, R.I. Rafiq, R. Han, Q. Lv, S. Mishra, Detection of cyberbullying incidents on the instagram social network, 2015, arXiv preprint arXiv:1503.03909 1503.03909.

[49] Q. Huang, V.K. Singh, P.K. Atrey, Cyber bullying detection using social and textual analysis, in: Proceedings of the 3rd International Workshop on Socially-Aware Multimedia, 2014, pp. 3–6.

[50] A. Squicciarini, S. Rajtmajer, Y. Liu, C. Griffin, 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, 2015, pp. 280–285.

[51] K.B. Kansara, N.M. Shekokar, A framework for cyberbullying detection in social network, Int. J. Curr. Eng. Technol. 5 (1) (2015) 494–498.

[52] K. Dinakar, R. Reichart, H. Lieberman, Modeling the detection of textual cyberbullying, in: Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[53] K. Reynolds, A. Kontostathis, L. Edwards, Using machine learning to detect cyberbullying, in: 2011 10th International Conference on Machine Learning and Applications and Workshops, Vol. 2, IEEE, 2011, pp. 241–244.

[54] V. Nahar, X. Li, C. Pang, An effective approach for cyberbullying detection, Commun. Inf. Sci. Manage. Eng. 3 (5) (2013) 238.

[55] H. Sanchez, S. Kumar, Twitter bullying detection, Ser. NSDI 12 (2011) (2011) 15.

[56] A. Kontostathis, K. Reynolds, A. Garron, L. Edwards, Detecting cyberbullying: query terms and techniques, in: Proceedings of the 5th Annual ACM Web Science Conference, 2013, pp. 195–204.

[57] H. Dani, J. Li, H. Liu, Sentiment informed cyberbullying detection in social media, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2017, pp. 52–67.

[58] P. Galán-García, J.G.d.l. Puerta, C.L. Gómez, I. Santos, P.G. Bringas, Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying, Log. J. IGPL 24 (1) (2016) 42–53.

[59] S. Salawu, Y. He, J. Lumsden, Approaches to automated detection of cyberbullying: A survey, IEEE Trans. Affect. Comput..

[60] H. Rosa, N. Pereira, R. Ribeiro, P.C. Ferreira, J.P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A.V. Simão, I. Trancoso, Automatic cyberbullying detection: A systematic review, Comput. Hum. Behav. 93 (2019) 333–345.

[61] L. Cheng, R. Guo, Y. Silva, D. Hall, H. Liu, Hierarchical attention networks for cyberbullying detection on the instagram social network, in: Proceedings of the 2019 SIAM International Conference on Data Mining, SIAM, 2019, pp. 235–243.

[62] D. Soni, V. Singh, Time reveals all wounds: Modeling temporal characteristics of cyberbullying, in: Twelfth International AAAI Conference on Web and Social Media, 2018.

[63] D.E. Losada, F. Crestani, A test collection for research on depression and language use, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2016, pp. 28–39.

[64] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, ACM, 2018, pp. 495–503.

[65] D.E. Losada, F. Crestani, J. Parapar, Overview of erisk at clef 2019 early risk prediction on the internet (extended overview), in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019.

[66] M.F. Lopez-Vizcaino, F.J. Novoa, D. Fernandez, V. Carneiro, F. Cacheda, Early intrusion detection for os scan attacks, in: 2019 IEEE 18th International Symposium on Network Computing and Applications (NCA), IEEE, 2019, pp. 209–213.

[67] H.-W. Liao, D.-L. Zhou, Review of adaboost and its improvement, Jisuanji Xitong Yingyong- Comput. Syst. Appl. 21 (5) (2012) 240–244.

[68] H. Allende-Cid, R. Salas, H. Allende, R. Nanculef, Robust alternating adaboost, in: Iberoamerican Congress on Pattern Recognition, Springer, 2007, pp. 427–436.

Manuel F. López-Vizcaíno was born in Lugo, Spain, in 1990. He received the B.S. degree in computer science from the University of A Coruña, Spain, in 2015. He is currently developing his Ph.D. studies in the same University, where he also works as teaching assistant at the same time. His research focuses on the evaluation and application of early detection methods to anomalies in cybersecurity, although he is also interested in other topics regarding Artificial Intelligence, evaluation metrics and network security.

Francisco J. Nóvoa was born in Ourense, Spain in 1974. He received the M.S degree in computer science from the University of Deusto, Spain, in 1998. He obtained his Ph.D. degree in computer science at the University of A Coruña, Spain, in 2007. From 1998 to 2007, he developed his professional career in the business field of Information Technology, reaching multiple professional certifications such as CCNA, CCNP, MCP, among others. From 2007 to 2018, he was assistant professor at the Computer Science Department at the University of A Coruña. Since then, he was Associate Professor at the same department. He is author of 12 journal articles, 10 book chapters and more than 30 conference articles. His research interests include network security, intrusion detection, data flow analysis, IoT, medical informatics, biomedical imaging, artificial intelligence and neural networks.

Víctor Carneiro received his Ph.D. and B.S. degree in Computer Science from University of A Coruña, A Coruña, Spain, in 1998 and 1993, respectively. He has been an associate professor of the Department of Information and Communication Technologies, University of A Coruña, Spain, since 1995.

He has participated in a lot of research projects and professional experiences related with network management, distributed systems, information retrieval over Internet and recommender systems based in collaborative filtering techniques. Nowadays he is working in technologies based on collective intelligence applied to the detection of anomalies and attacks in TCP/IP networks and IoT protocols.

Fidel Cacheda was born in Poissy, France in 1973. He received the B.S. degree in computer science from the University of A Coruña, Spain, in 1994 and the M.S. degree in computer science from the same university in 1996. He obtained his Ph.D. degree in computer science at the University of A Coruña, Spain in 2002.

From 1998 to 2006, he was an Assistant Professor at the Computer Science Department at the University of A Coruña. Since then, he has been an Associate Professor at the same department. He is author of four books, nine book chapters, more than 20 journal articles and more than 60 conference articles. His research interests include information retrieval, recommender systems and early detection of anomalies applied to cybersecurity.