Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying

PATXI GALÁN-GARCÍA*, JOSÉ GAVIRIA DE LA PUERTA**, CARLOS LAORDEN GÓMEZ†, IGOR SANTOS‡ and PABLO GARCÍA BRINGAS§, *DeustoTech Computing, University of Deusto.*

Abstract

The use of new technologies along with the popularity of social networks has given the power of anonymity to the users. The ability to create an alter-ego with no relation to the actual user, creates a situation in which no one can certify the match between a profile and a real person. This problem generates situations, repeated daily, in which users with fake accounts, or at least not related to their real identity, publish news, reviews or multimedia material trying to discredit or attack other people who may or may not be aware of the attack. These acts can have great impact on the affected victims' environment generating situations in which virtual attacks escalate into fatal consequences in real life. In this article, we present a methodology to detect and associate fake profiles on Twitter social network which are employed for defamatory activities to a real profile within the same network by analysing the content of the comments generated by both profiles. Accompanying this approach we also present a successful real life use case in which this methodology was applied to detect and stop a cyberbullying situation in a real elementary school.

Keywords: Online social networks, trolling, information retrieval, cyberbullying.

1 Introduction

Online Social Networks (OSNs) are one of the most frequently used Internet services. There is not a generic definition of these platforms, although Boyd *et al*. [3] defined them as web services that allow an individual to do three things: (i) generate a public or semi-public profile in a specific system, (ii) create a list of users to interact with and browse through the list of contacts and (iii) see what was done by others within the system.

The massive presence that users have in this platforms and the relative ease to hide a user's real identity implies that false profiles and 'troll users'[1] are spreading, becoming a nuisance to legitimate users of these services.

In some cases, these 'malicious' users, use OSN platforms to commit crimes like identity impersonation, defamation, opinion polarization or cyberbullying. Despite all of them present real and worrying dangers [19, 27] the later, cyberbullying, has actually become one of the most hideous

*E-mail: patxigg@deusto.es
**E-mail: jgaviria@deusto.es
†E-mail: claorden@deusto.es
‡E-mail: isantos@deusto.es
§E-mail: pablo.garcia.bringas@deusto.es

[1]Users posting inflammatory extraneous or off-topic messages in an online community.

Vol. 24 No. 1, © The Author 2015. Published by Oxford University Press. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
doi:10.1093/jigpal/jzv048 Advance Access published 31 October 2015

Downloaded from https://academic.oup.com/jigpal/article/24/1/42/2893010 by National Institute of Technology Karnataka user on 04 January 2022

problems in our society, generating even more side effects than 'real life bullying' [31] due to the impact and the permanent nature of the comments flooding OSN platforms. As a hard to forget, but not isolated, example recently one teenager, Amanda Todd, committed suicide due to harassment in the form of blackmailing, bullying and physical assaults[2] using OSNs as the channel of abuse.

Another sample of this type of harassment, was the case of Ghyslain Raza. He recorded himself in the well-known '*Star Wars Kid*'[3] video, emulating the movements of '*Darth Maul*' in Start Wars Episode 1. This video attracted negative attention in the form of critics and bullying messages to Ghyslain. Some of them even suggested him to commit suicide. Finally, and fortunately, Ghyslain decided to help bring to light the type of bullying and negative attention that children might receive in similar incidents due to or supported by the rise of social media.

This hurtful event, which raised the discussion on criminalizing cyberbullying in several areas, was kind of aided by the aforementioned possibility of easily creating false profiles in social network platforms. Thanks to these anonymous and not-linked-to-real-life profiles some twisted minded individuals are able to torment their victims without even being brought to justice. In a similar vein, it was recently brought to our attention an incident in an Spanish elementary school where an alleged student was writing defamatory comments in Twitter using a fake profile, causing anxiety attacks and depression episodes among the affected students.

Some social platforms are trying to manually identify the real person behind their profiles but, until the job is done, being able to correlate or link a false profile to a real person within the network is the only option to fight the problem.

In light of this background, we present a methodology to associate a false profile's tweets with one real individual, provided he/she has another profile created with real information. The assumption that the trolling user will have another 'real' profile is not fortuitous, it relies on the fact that these kind of users like to interact with the fake identity and stay updated and participate in parallel conversations. Moreover, we apply the presented methodology to a real life cyberbullying situation inside an elementary school.

2 Background

2.1 OSNs

OSNs are platforms that provide users with some useful tools to interact with other users through connections. The connections start by creating a profile in an OSN, which consists of a user's representation and can be private, semi-public or public.

The importance of the connections users make within these social platforms resides on the amount of information, usually private, that is published in these usually public profiles. Therefore, privacy in OSNs is a serious issue. Despite the importance of this fact it is commonly ignored by the users of social platforms, resulting in the publication of too much personal information, information which can be (and in some cases is) used by sexual predators, criminals, large corporations and governmental bodies to generate personal and behavioural profiles of the users.

But not always this information is published by the users. Sometimes, these data are leaked by: (i) browsers wich give more information than we perceive (Operating System that we use, our IP address, browser version, etc.); (ii) Instant Messaging (IM) services, due to their online and offline status; or (iii) even photo tagging, which OSN are automating based on facial recognition systems. The last example, automatic photo tagging, allows identifying a person on a large amount

[2]http://en.wikipedia.org/wiki/Suicide_of_Amanda_Todd
[3]https://en.wikipedia.org/wiki/Star_Wars_Kid

Downloaded from https://academic.oup.com/jigpal/article/24/1/42/2893010 by National Institute of Technology Karnataka user on 04 January 2022

of photographs, including the embarrassing or inappropriate ones. These algorithms, combined with other technologies, allow finding singular persons on OSN with an acceptable accuracy [19], and many associated side effects.

In fact, one of the currently most worrying problems in OSNs is cyberbullying. This problem is a relatively new and widespread situation and most of the times implies an emotional trauma for the victim. Teenagers use these platforms to express their feelings and life experiences because they are not able to do it in their real lives, while other users abuse and torment others with impunity [3]. This common misuse of OSNs, the emotional abuse, is commonly referred to as trolling [2] and can happen in many different ways like defacement of deceased person pages, name calling, controversial comments with the intention to cause anger and cause arguments. To solve this type of problems, some OSNs have chosen the approach of age verification systems with real ID cards, but not always with successful results [22].

2.2 Cyberbullying

Cyberbullying, according to [33], refers to any harassment that occurs via the internet, cell phones or other devices. This type of bullying uses communication technologies to intentionally harm others through hostile behaviour such as sending text messages and posting ugly comments on the Internet. But, usually, the definition of this phenomenon starts using the traditional definition of bullying.

In the literature of cyberbullying detection, the main focus has been directed towards the content of the conversations. For example using text mining paradigms to identify online sexual predators [17, 18], vandalism detection [30], spam detection [32] and detection of internet abuse and cyberterrorism [28]. This type of approaches are very promising, but not always applicable to every aspect of cyberbullying detection because some attacks can only be detected by analysing user's contexts.

In a recent study on cyberbullying detection, Dinakar *et al*. [8] applied a range of binary and multiclass classifiers on a manually labelled corpus of YouTube comments. The results showed that a binary individual topic-sensitive classifiers approach can outperform the detection of textual cyberbullying when compared to multiclass classifiers. They showed the application of common sense knowledge in the design of social network software for detecting cyberbullying. The authors treated each comment on its own and did not consider other aspects to the problem as such the pragmatics of dialogue and conversation and the social networking graph. They concluded that taking into account such features would be more useful on OSN websites and which could do a better modelling of the problem.

Another approaches focus on the characteristics of the participants involved in the conversation. For example in [7], Maral *et al*. analysed *profane words*, *cyberbullying words*, *pronouns* and each user *comments history* to improve the active detecction of harassers in Youtube. In another interesting approach [6], the same team improved cyberbullying detection, taking into account the most utilized words depending on the gender (i.e. male or female) using that input as a prefilter to later apply speciffic harrasment models associated to that gender.

2.3 Information retrieval

Information Retrieval (IR) is a sub-field of Computer Science about searching relevant information in digital documents, digital documentary collections or Internet information resources. Searches can be based on metadata information, full-text indexing information or relational databases systems.

Downloaded from https://academic.oup.com/jigpal/article/24/1/42/2893010 by National Institute of Technology Karnataka user on 04 January 2022

The goal of this search is to obtain documents, photos, sounds or any information related to the desired query.

In 1945 Vannevar Bush published a ground breaking article titled 'As We May Think' [5] that gave birth to the idea of automatic access to large amounts of stored knowledge. In the 1950s, this idea materialized into more concrete descriptions of how archives of text could be searched automatically.

In 1950s, several works emerged that elaborated upon the basic idea of searching text with a computer. This increase was due to the end of Second World War and the problems of the US army to index and retrieve scientific research documents captured from Germans on wartime. One of this works to solve the problem was Hans Peter Luhns' work. He began to work on a mechanized punch card-based system for searching chemical compounds [20] and, the most influential method that he proposed was to employ words as indexing units for documents and measuring word overlap as a criterion for retrieval [21].

With the Cold War, the US army promoted initiatives to improve mechanized literature searching systems [14] and the invention of citation indexing [10]. In 1950, Calvin Mooers used at first the term '*information retrieval*'[9]. In 1955, Allen Kent published a article in American Documentation describing the precision and recall measures. Also, in this article, Allen proposed one framework to evaluate IR systems using statistical methods for determining the number of relevant documents not retrieved [15].

The 1970s and 1980s saw many developments built on the advances of the 1960s. These new approaches were experimentally tested on small text collections because at that time, researches did not have access to large text collections. To solve it, in 1992 several US Government agencies, under the National Institute of Standards and Technology (NIST) supervision, created the Text Retrieval Conference (TREC-1) [12].

Since 1996, the most relevant approaches to IR are Web Search engines like Google, Bing or Yahoo!. These systems use IR to obtain references, documents and available cross linkage information on the web about the users' interations on the Internet.

3 Proposed approach

The main idea underlying our approach is that *every trolling profile is followed by the real profile of the user behind the trolling one*. This assumption is based on the fact that this kind of users want to stay updated on the activity that surrounds the fake profile. Besides, each individual writes in a characteristic way. Studying different features of the written text is possible to determine the authorship of, e.g. e-mails [34]. In this case, despite users behind fake profiles may try to write in a different way to avoid detection, Twitter provides other characteristics that, in conjunction with the text analysis, may be used to link a trolling account to a real user's profile.

Therefore, with these ideas in mind, we postulated the following hypothesis: *It is possible to link a trolling account to the corresponding real profile of the user behind the fake account, analysing different features present in the profile, connections' data and tweets' characteristics, including text, using machine learning algorithms*.

To prove the hypothesis, we first prepared the method to determine the authorship of twitter profiles based on their published tweets and then applied these techniques to a real cyberbullying situation in one elementary school.

The authorship identification step was performed studying a group of profiles which have some kind of relation among them (to replicate to the best possible extent the conditions of a classroom

Downloaded from https://academic.oup.com/jigpal/article/24/1/42/2893010 by National Institute of Technology Karnataka user on 04 January 2022

cyberbullying event). The methodology would follow the next steps: (i) select the profiles under study, (ii) collect all the information of the profile and its tweets, (iii) select the features to be extracted from the retrieved tweets and (iv) apply machine learning methods to build the models that will determine the authorship of the gathered tweets.

3.1 Selecting different profiles

The number of selected profiles for this study was 19. The idea was to gather several profiles with social relations both inside the social network and in real life. We wanted to avoid retrieving very different Twitter accounts, with respect to their content, which would ease the task of determining the authorship. In this way, the conditions found in cyberbullying situations, with respect to the participants and conditions, are sufficiently replicated.

In order to select the profiles we checked that it was not private, the number of own tweets (less than 50–100 samples would imply almost no activity), number of followers and following users (no connections would show no interaction with other users) and the relation between other selected profiles (we wanted the selected accounts to be connected).

Therefore, the first selected profile was of one of the authors, and then we continued analysing their followers and followings, keeping the desired ones. The final dataset was homogeneous, regarding writting style, but was also diverse, regarding the real person behind the Twitter account (e.g. different ages, genders, location, etc.).

3.2 Collecting profiles data and tweets

It is important to notice of limitation imposed by Twitters' API. The number of requests could not exceed 350 per hour, which limits considerably the retrieve a large amount of samples, so we had to use several accounts to gather them.

Our Java-based collecting method obtained, from the selected profiles, the users' ID and the timeline tweets, until having at least 100 *genuine tweets* with all abbreviations and slangs. A genuine tweet is the tweet that is generated by the user itself (i.e. written by itself) and is not one retweet of another user's tweet.

3.3 Features

The tweets, time of publication, language, geoposition and Twitter client. The first feature, the tweet, is the text published by the user, which gives us the possibility to determine a writing style, very differentiating for each individual. The time of publication helps determining the moments of the day in which the users interact in the social network. The language and geoposition also help filtering and determining the authorship because users have certain behaviours which can be extrapolated analysing these features. Finally, despite being possible that users have several devices from where they tweet (e.g. PC, smartphone or tablet), they usually choose to do it using their favourite Twitter client, which gives us another filtering mechanism.

An example of all these features composing the dataset, which were extracted from each of the profiles can be seen on Table 1.

Downloaded from https://academic.oup.com/jigpal/article/24/1/42/2893010 by National Institute of Technology Karnataka user on 04 January 2022

TABLE 1. Description of selected features

Label	Description	Value
Twitter text	Content of the message sent to Twitter	*you are like a bitch*
Time of publication	Date and time when the message was published	*Mon Mar 05 18:31:56 2013*
Language	The language of users' profile	*en*
Geoposition	Coordinates or location name the message was sent from	*San Francisco, CA*

3.4 Supervised learning

Once we have the profile data, user's tweets and the chosen features, the next step is to generate an ARFF [13] file (i.e. Attribute Relation File Format) to classify the profiles according to the writing style of the tweets using WEKA (i.e. the Waikato Enviroment for Knowledge Analysis) [11].

In this experiment, we have chosen to compare the performance of different classification algorithms included in WEKA:

- ***Random Forest***. It is an aggregation sorter developed by Leo Breiman [4] and comprising a range of *decision trees* so as to improve the classification accuracy since in the construction of each individual classifier introduces a stochastic component, either in the partition of the space (the actual construction of the trees) or in the training sample.
- **J48**. J48 is an open-source implementation of the C4.5 algorithm [24] for Weka. C4.5 creates decision trees given a number of training information using the concept of information entropy [26]. Training data is a set $S = s_1, s_2, ..., s_n$ of $examples = s_1 x_1, x_2, ..., x_m$ and classified under that $x_1, x_2, ..., x_m$ represent the attributes or characteristics of the sample. At each node of the decision tree, the algorithm chooses an attribute of the data more effectively divide the sample set into subsets enriched in one kind or another selection criterion using the entropy or difference of the aforementioned *normalized information gain*.
- ***K-Nearest Neighbor* (KNN)**. The KNN algorithm is one of the simplest sorting algorithms in machine learning. This is a classification method based on analysis of the k nearest neighbours in space analysis ($\forall k \in \mathbb{N}$). In this case and given the simplicity of the algorithm, the values of k used to check the effectiveness thereof have been k = 1,2,3,4,5 to determine if taking account of more neighbours significantly improves the results.
- ***Sequential Minimal Optimization* (SMO)**. SMO, invented by John Platt [23], is an iterative algorithm for solving optimization problems that arise with the training of the *Support Vector Machines* or support vector machines. SMO part the problem into a series of smaller subproblems that are solved analytically later.

To optimize the results, before training the classifiers, we filtered the *tweets* text with stopwords [35] in Spanish.[4] To validate the suitability of the results, we employed K-fold cross-validation [16], a technique which consists on dividing the dataset into K folds, using the instances corresponding to $K-1$ folds for training the model, and the instances in the remaining fold for testing. K training rounds are performed using a different fold for testing each time, and thus, training and testing the model with every possible instance in the dataset.

[4]http://paginaspersonales.deusto.es/claorden/resources/SpanishStopWords.txt

Downloaded from https://academic.oup.com/jigpal/article/24/1/42/2893010 by National Institute of Technology Karnataka user on 04 January 2022

At last, to evaluate the results, we used *True Positive Ratio* (*TPR*), *False Positive Ratio* (*FPR*) and *Area Under ROC Curve* (*AUC*):

- *True Positive Ratio* (*TPR*), which is calculated by dividing the number of tweets correctly classified (*TP*) between the total samples taken ($TP+FN$). $TPR=TP/(TP+FN)$.
- *False Positive Ratio* (*FPR*), which is calculated by dividing the number of tweets whose classification is wrong (*FP*) by the total number of samples ($FP+TN$). $FPR=FP/(FP+TN)$.
- *Area Under ROC Curve* (*AURC*) [29], which establishes the relationship between the the false negatives and the false positives. The ROC curve is often used to generate performance statistics representing or effectiveness, in a broader sense—the classifier.

4 Experiments

To evaluate the capabilities of the method to assign the correct authorship to the Twitter profiles, we used a dataset comprising 1,900 tweets corresponding to 19 different twitter accounts (100 tweets per profile).

For the experiments, we modelled the tweets using the Vector Space Model (VSM) [25]. VSM is an algebraic approach for Information Filtering (IF), Information Retrieval (IR), indexing and ranking. This model represents natural language documents mathematically by vectors in a multidimensional space where the axes are terms within messages. We used the *Term Frequency – Inverse Document Frequency* (TF–IDF) [25] weighting schema, where the weight of the i^{th} term in the j^th document, denoted by $weight(i,j)$, is defined by $weight(i,j)=tf_{i,j}\cdot idf_i$ where *term frequency* $tf_{i,j}$ is defined as $tf_{i,j}=n_{i,j}/\sum_k n_{k,j}$ where $n_{i,j}$ is the number of times the term $t_{i,j}$ appears in a document d, and $\sum_k n_{k,j}$ is the total number of terms in the document d. The inverse term frequency idf_i is defined as $idf_i=|\mathcal{D}||\mathcal{D}:t_i\in d|$ where $|\mathcal{D}|$ is the total number of documents and $|\mathcal{D}:t_i\in d|$ is the number of documents containing the term t_i.

Moreover, VSM requires a pre-processing step in which messages are divided into tokens by separator characters (e.g. space, tab, colon, semicolon or comma). The *tokenisation*, the process of breaking the stream of text into the minimal units of features (i.e. the tokens) [25], was based in an n-gram selection with sizes of $n=1$, $n=2$ and $n=3$. This process is performed to construct the VSM representation of the messages and it is required for the learning and testing of classifiers [1].

Table 2 shows the results obtained after applying the selected algorithms to our dataset, measured in Accuracy, *FPR*, *TPR* and *AUC*.

The results show that SMO and Decision Trees are the most appropriate algorithms. More precisely, the best results are obtained using a PolyKernel with 68.47% accuracy and 0.96 of *AUC*. In second and third position, very close, we have J48 with 65.81% accuracy and 0.94 and NormalizedPolyKernel with 65.29% accuracy and 0.94 of *AUC*. Random Forest, in fourth position, obtains 66.48% accuracy and 0.93 *AUC*. Finally, *KNN* and Naive Bayes algorithms do not have remarkable results, with values from 59.39% to 61.06% of accuracy for *KNN* and of 33.91 for Naive Bayes in terms of accuracy and from 0.89 to 0.92 and 0.90 respectively in terms of *AUC*.

5 Real case study

The proposed methodology, has been tested in a real situation. In one school in the city of Bilbao (Spain), some students were implicated in a cyberbullying situation. The staff of this school proposed to us whether it was possible to find which of the students had been the author/s behind the trolling profile or not.

Downloaded from https://academic.oup.com/jigpal/article/24/1/42/2893010 by National Institute of Technology Karnataka user on 04 January 2022

TABLE 2. Obtained results for the selected machine learning algorithms.

Algorithm	Accuracy	*FPR*	*TPR*	*AUC*
SMO-PolyKernel	68.47	0.02	0.68	0.96
J48	65.81	0.02	0.66	0.94
SMO-NormalizedPolyKernel	65.29	0.02	0.65	0.94
RandomForest	66.48	0.02	0.66	0.93
KNN $k=10$	59.79	0.02	0.60	0.92
KNN $k=3$	59.7	0.02	0.60	0.90
KNN $k=5$	59.39	0.02	0.59	0.90
NaiveBayes	33.91	0.04	0.34	0.90
KNN $k=2$	61.06	0.02	0.61	0.89

Note: It must be noted that we applied the default configurations under WEKA for each of the algorithms.

In this case the profile was named 'Gossip', in a clear reference to the popular TV Show *Gossip Girl*, and, for two weeks, the student using this profile commented personal indiscretions about his/her classmates. Initially, it was only the publication of not hurtful tweets. These comments included events or facts such as one student not doing the assigned homework.

Two weeks later, the profile started publishing things a little more private but not very important. At that moment, all the classmates were following that profile and in the school hallways the students theorized about who could be the responsible but never had the certainty to prove it.

Then, the teachers at the school started to fear the relevance of what at first seemed as a childish game, but that had evolved into a serious problem. They did not know what they could do with it and decided it was time to ask for help.

Once we were introduced to the situation, we first analyzed the trolling profile, the published comments and the interaction with other profiles. We noticed a repeated behaviour, most of the contents were referred to a particular girl and had a lot of personal and school-related information. Those facts revealed that the author behind the fake profile had to be a member of the same school or even the same class. Moreover, we took the assumption that the real person behind the 'Gossip Girl' was following the fake profile, which is consistent with the theory that most of these users want to keep track of the activities and parallel conversations surrounding the trolling profile.

With these considerations in mind, we retrieved all the tweets from the 'gossip profile' and their followers and followings profiles. The idea was to train our classifiers with the tweets published by all the users interacting with the trolling profile and then try to identify the authorship of Gossip's tweets using the acquired knowledge.

As a result, we obtained 17,536 tweets corresponding to the 92 users who were followers and/or followings of Gossip Girl, and 43 tweets from the trolling profile, Gossip Girl.

Table 3 offers the results of the authorship identification carried out by the best four classifiers analysed in Section 4: SMO-PolyKernel, J48, SMO-NormalizedPolyKernel and RandomForest.

The table shows the level of authorship attributed to each user from the whole collection of messages (43 tweets) published by Gossip Girl. It can be appreciated that three of them appear among the top 4 in the fourth classifiers (highlighted cells). These results made us realize that those three subjects had a great probability of being the responsible ones behind the trolling profile. It is interesting to add, that what seemed to be a one-person misbehaviour had turned, apparently, into a group abuse.

Downloaded from https://academic.oup.com/jigpal/article/24/1/42/2893010 by National Institute of Technology Karnataka user on 04 January 2022

TABLE 3. Results of the authorship identification for the 92 users followers and/or followings of Gossip Girl's trolling profile.

SMO PolyKernel PolyKernel		**J48 NormalizePolykernel**		**SMO NormalizePolykernel Forest**		**Random**	
Sub.#	**Authorship**	**Sub.#**	**Authorship**	**Sub.#**	**Authorship**	**Sub.#**	**Authorship**
34	**23%**	34	**21%**	34	**21%**	42	**21%**
66	**19%**	42	**16%**	42	**14%**	34	**16%**
42	**14%**	87	**14%**	66	**14%**	87	**14%**
87	**14%**	46	**12%**	87	**14%**	46	**12%**
8	**12%**	8	**7%**	30	**12%**	31	**7%**
31	**5%**	50	**7%**	31	**7%**	50	**7%**
63	**2%**	31	**5%**	33	**5%**	8	**2%**
20	**2%**	83	**5%**	50	**2%**	14	**2%**
29	**2%**	14	**2%**	8	**2%**	39	**2%**
53	**2%**	39	**2%**	20	**2%**	53	**2%**
83	**2%**	53	**2%**	29	**2%**	83	**2%**
91	**2%**	29	**2%**	53	**2%**	20	**2%**
33	**0%**	30	**2%**	63	**2%**	29	**2%**
49	**0%**	91	**2%**	24	**0%**	44	**2%**
52	**0%**	33	**0%**	28	**0%**	48	**2%**
1	**0%**	48	**0%**	49	**0%**	63	**2%**
2	**0%**	66	**0%**	52	**0%**	6	**0%**
3	**0%**	6	**0%**	1	**0%**	49	**0%**
4	**0%**	57	**0%**	2	**0%**	56	**0%**
5	**0%**	63	**0%**	3	**0%**	66	**0%**

Notes: The profile name of the account has been replaced with a subject number due to anonymity issues. The authorship percentage corresponds to the number of tweets published as Gossip Girl, out of the total 43 tweets from the trolling profile, that have been related to the subject. Note that only 20 of the users are presented in this table, as most of them have absolutely no indication of being behind the trolling profile.

After the analysis, we reported our findings to the school's staff. With the knowledge of the three names of the alleged abusers the managing office in the school summoned all the students, warning them about the consequences of this misconduct, trying to reduce the impact of their acts, should they confess before it was too late. At the end, the anonymous attackers voluntarily revealed their identity frightened by the possible consequences. These identities had been previously reported by our system.

Finally the staff of the school did not reported the event but required them to publicly apologize.

6 Conclusions

More and more children are nowadays connected to the Internet. Although this communication channel provides a lot of important advantages, in many cases, because of the anonymity, different kind of abuses may arise, being one of them the cyberbullying. A rapid identification of this type of

Downloaded from https://academic.oup.com/jigpal/article/24/1/42/2893010 by National Institute of Technology Karnataka user on 04 January 2022

users on the Internet is crucial, giving a lot of importance to the systems and/or tools able to identify these threats, in order to protect this population segment on the Internet.

Therefore, we consider that the hereby proposed methodology offers a safe way to identify the real user or users behind a trolling account given some previous conditions: (i) the real user/s behind the fake profile has/have a 'real' and active account in the social network, (ii) the real account of the user/s behind the fake profile is/are somehow connected to the fake profile. These conditions are in theory easy to fulfil due to the assumption that a real person behind a trolling profile wants to keep track of the activities and parallel conversations surrounding the trolling profile.

However, the proposed mechanisms have several limitations. First, despite we assume the previous conditions will be fulfilled, there could be the case in which a user behind a trolling account has no relation with the fake profile to avoid rising suspicions. In this case, it would be necessary to enlarge the circle of users to be analysed or even find a more specific circle based on specific characteristics of the trolling profile. Second, expert abusive users can intentionally change their writing style and/or behaviour to avoid detection. Being the behaviour (e.g. device, time, location) the most difficult to change due to the unconsciousness nature of most human acts, it would also be a really effective way of avoiding detection with no clear solution. On the other hand, the change on writing style could be tackled by analysing the language in more depth, finding for example the use of synonyms or word alterations.

Therefore, as future work, it would be interesting to expand this work in three main directions. First, we would like to analyse different language characteristics and semantics present in the tweets. That analysis could/should include more NLP techniques such as language phenomena study (e.g. synonymity, metonymy or homography), Word Sense Disambiguation (WSD) or Opinion Mining, among others. Besides, given the nature of social networks, it is 'easy' to hide among the vast number of users populating these platforms. A possible approach would be to create a kind of *writing style/behaviour signature* able to identify twitter users by the published content. In case of detecting an abuse, that information could be used to reduce the number of users to be further analysed. Finally, we would like to adopt this work to other social networks, chat rooms and similar environments.

References

[1] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 1999.

[2] J. Bishop. Scope and limitations in the government of wales act 2006 for tackling internet abuses in the form of 'flame trolling'. *Statute Law Review*, **33**, 207–216, 2012.

[3] D. Boyd and N. Ellison. Social network sites: definition, history, and scholarship. *Journal of Computer-Mediated Communication*, **13**, 210–230, 2007.

[4] L. Breiman. Random forests. *Machine Learning*, **45**, 5–32, 2001.

[5] V. Bush. As we may think, 1945.

[6] M. Dadvar, F. de Jong, R. Ordelman and R. Trieschnigg. Improved Cyberbullying Detection using Gender Information, 2012.

[7] M. Dadvar, D. Trieschnigg, R. Ordelman and F. de Jong. Improving cyberbullying detection with user context. In *Advances in Information Retrieval*, Berlin Heidelberg, pp. 693–696. Springer, 2013.

[8] K. Dinakar, R. Reichart and H. Lieberman. Modeling the detection of textual cyberbullying. In *International Conference on Weblog and Social Media-Social Mobile Web Workshop*, 2011.

[9] R. Fairthorne and C. Mooers. *Towards information retrieval*. Archon Books, 1968.

Downloaded from https://academic.oup.com/jigpal/article/24/1/42/2893010 by National Institute of Technology Karnataka user on 04 January 2022

[10] E. Garfield and R. Merton. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*, Vol. 8. Wiley New York, 1979.
[11] S. R. Garner, *et al*. Weka: The waikato environment for knowledge analysis. In *Proceedings of the New Zealand Computer Science Research Students Conference*, pp. 57–64. Citeseer, 1995.
[12] D. Harman. Overview of the first trec conference. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 36–47. ACM, 1993.
[13] G. Holmes, A. Donkin and I. H. Witten. Weka: A machine learning workbench. In *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pp. 357–361. IEEE, 1994.
[14] A. Kent. *Textbook on mechanized information retrieval*. Interscience, 1966.
[15] A. Kent, M. Berry, F. Luehrs and J. Perry. Machine literature searching viii. operational criteria for designing information retrieval systems. *American documentation*, **6**, 93–101, 2007.
[16] R. Kohavi, *et al*. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, Vol. 14, pp. 1137–1145. Lawrence Erlbaum Associates Ltd, 1995.
[17] A. Kontostathis. Chatcoder: toward the tracking and categorization of internet predators. In *Text Mining Workshop 2009 held in Conjunction with the 9th Siam International Conference on Data Mining (SDM 2009)*. Sparks, NV. May 2009.
[18] C. Laorden, P. Galán-García, I. Santos, B. Sanz, J. M. G. Hidalgo and P. G. Bringas. Negobot: a conversational agent based on game theory for the detection of paedophile behaviour. In *International Joint Conference CISIS'12-ICEUTE'* 12-SOCO' 12 Special Sessions, pp. 261–270. Springer, 2013.
[19] C. Laorden, B. Sanz, G. Alvarez and P. G. Bringas. A threat model approach to threats and vulnerabilities in online social networks. In *Computational Intelligence in Security for Information Systems 2010*, Vol. 85 of *Advances in Intelligent and Soft Computing*, pp. 135–142, 2010.
[20] H. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, **1**, 309–317, 1957.
[21] H. Luhn. *Auto-Encoding of Documents for Information Retrieval Systems*. IBM Research Center, 1958.
[22] J. Palfrey, D. Sacco, D. Boyd, L. DeBonis and J. Tatlock. Enhancing child safety & online technologies. *Accessed Online: cyber. law. harvard. edu/sites/cyber. law. harvard. edu/files/ISTTF_Final_Report. pdf*, 2008.
[23] J. Platt, *et al*. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods—support vector learning 3*, 1999.
[24] J. Quinlan. *C4. 5: programs for machine learning*. Morgan kaufmann, 1993.
[25] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill New York, 1983.
[26] S. L. Salzberg. C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, **16**, 235–240, 1994.
[27] B. Sanz, C. Laorden, G. Alvarez and P. G. Bringas. A threat model approach to attacks and countermeasures in online social networks. In *In Proceedings of the 11th Reunion Española de Criptograf a y Seguridad de la Informacion (RECSI), 7-10th September, Tarragona (Spain).*, pp. 343–348, 2010.

Downloaded from https://academic.oup.com/jigpal/article/24/1/42/2893010 by National Institute of Technology Karnataka user on 04 January 2022

[28] D. A. Simanjuntak, H. P. Ipung, C. Lim and A. S. Nugroho. Text classification techniques used to faciliate cyber terrorism investigation. In *Advances in Computing, Control and Telecommunication Technologies (ACT), 2010 Second International Conference on*, pp. 198–200. IEEE, 2010.

[29] Y. Singh, A. Kaur and R. Malhotra. Comparative analysis of regression and machine learning methods for predicting fault proneness models. *International Journal of Computer Applications in Technology*, **35**, 183–193, 2009.

[30] K. Smets, B. Goethals and B. Verdonk. Automatic vandalism detection in wikipedia: towards a machine learning approach. In *AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pp. 43–48, 2008.

[31] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell and N. Tippett. Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, **49**, 376–385, 2008.

[32] P.-N. Tan, F. Chen and A. Jain. Information assurance: detection of web spam attacks in social media. In *Proceedings of Army Science Conference, Orland, Florida*, 2010.

[33] H. Vandebosch and K. Van Cleemput. Defining cyberbullying: a qualitative research into the perceptions of youngsters. *CyberPsychology & Behavior*, **11**, 499–503, 2008.

[34] O. De Vel, A. Anderson, M. Corney and G. Mohay. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, **30**, 55–64, 2001.

[35] W. J. Wilbur and K. Sirotkin. The automatic identification of stop words. *Journal of information science*, **18**, 45–55, 1992.

Received 10 July 2014

Downloaded from https://academic.oup.com/jigpal/article/24/1/42/2893010 by National Institute of Technology Karnataka user on 04 January 2022