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
patxigg@deusto.es, jgaviria@deusto.es, claorden@deusto.es, isantos@deusto.es, pablo.garcia.bringas@deusto.es

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 paper, 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 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: On-line Social Networks, Trolling, Information Retrieval, Identity Theft, Cyberbullying.

1 Introduction

On-line Social Networks (OSNs) are some of the most frequently used Internet services. There is not a generic definition of these platforms, although Boyd et al [1] 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 easiness 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.

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

In some cases, these "malicious" users, use OSN platforms to commit crimes like identity impersonation, defamation, opinion polarisation or cyberbullying. Despite all of them present real and worrying dangers [2,3] the later, cyberbullying, has actually become one of the most hideous problems in our society, generating even more side effects than "real life bullying" [4] 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.

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 On-line Social Networks

On-line Social Networks (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.

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

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 [1]. This common misuse of OSNs, the emotional abuse, is commonly referred to as trolling [5] 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 [6].

2.2 Cyberbullying

Cyberbullying, according to [7], 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 on-line sexual predators [8] [9], vandalism detection [10], spam detection [11] and detection of internet abuse and cyberterrorism [12]. 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. [13] 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 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 will be more useful on social networking websites and which could to a better modelling of the problem.

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, for example, e-mails [14]. 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 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 varied, we had men and women, different ages, different studies and different behaviours of their Twitter account.

3.2 Collecting profiles data and tweets

It is important to notice that there is one important limitation imposed by the Twitter API. The number of requests could not exceed 350 per hour, which limits considerably the possibility to 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*. 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

Our dataset contains the following features extracted from each of the profiles 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 of determine a writing style, very characteristic of 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.

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 [15] 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) [16].

In this experiment, we have chosen to compare the performance of different classification algorithms included in WEKA: i) Random Forest, ii) J48 (WEKA's C4.5 implementation), iii) K-Nearest Neighbor (KNN), iv) Sequential Minimal Optimization (SMO) and v) Bayes Theorem-based algorithms.

To optimize the results, before training the classifiers, we filtered the *tweets* text with stopwords [17] in Spanish[3].

To validate the suitability of the results, we employed K-fold cross-validation [18], 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.

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

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).

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

For the experiments, we modelled the tweets using the Vector Space Model (VSM) [19]. 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) [19] 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) [19], 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 [20].

Table 1 shows the results obtained after applying the selected algorithms to our dataset, measured in Accuracy, FPR, TPR and AUC.

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

Table 1. Obtained results for the selected machine learning algorithms. It must be noted that we applied the default configurations under WEKA for each of the algorithms.

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.

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 theorised 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 in the situation, we first analysed 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 (64,835% women and 35,165% men, between 14 and 18 years old) who were followers and/or followings of Gossip Girl, and 43 tweets from the trolling profile, Gossip Girl.

Table 2 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 subject from the whole collection of messages (43 tweets) published by Gossip Girl. It can be appreciated that three subjects appear among the top 4 in the fourth classifiers

SMO PolyKernel PolyKernel		J48		SMO NormalizePolykernel NormalizePolykernel		Random Forest	
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%**

Table 2. Results of the authorship identification for the 92 users followers and/or followings of Gossip Girl's trolling profile. 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 subjects are presented in this table, as most of them have absolutely no indication of being behind the trolling profile.

(highlighted cells). These results made us realise 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.

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. With fear in the body, the perpetrators revealed their identity, which matched with the three names we identified.

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 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. Firstly, 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 bast 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. Boyd, D., Ellison, N.: Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication **13**(1) (2007) 210–230
2. Sanz, B., Laorden, C., Alvarez, G., Bringas, P.G.: A threat model approach to attacks and countermeasures in on-line social networks. In: In Proceedings of the 11th Reunion Española de Criptografía y Seguridad de la Información (RECSI), 7-10th September, Tarragona (Spain). (2010) 343–348
3. Laorden, C., Sanz, B., Alvarez, G., Bringas, P.G.: A threat model approach to threats and vulnerabilities in on-line social networks. In: Computational Intelligence in Security for Information Systems 2010. Volume 85 of Advances in Intelligent and Soft Computing. (2010) 135–142

4. Smith, P.K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., Tippett, N.: Cyberbullying: Its nature and impact in secondary school pupils. Journal of Child Psychology and Psychiatry **49**(4) (2008) 376–385
5. Bishop, J.: Scope and limitations in the government of wales act 2006 for tackling internet abuses in the form of 'flame trolling'. Statute Law Review **33**(2) (2012) 207–216
6. Palfrey, J., Sacco, D., Boyd, D., DeBonis, L., Tatlock, J.: Enhancing child safety & online technologies. Accessed Online: cyber. law. harvard. edu/sites/cyber. law. harvard. edu/files/ISTTF_Final_Report. pdf (2008)
7. Vandebosch, H., Van Cleemput, K.: Defining cyberbullying: A qualitative research into the perceptions of youngsters. CyberPsychology & Behavior **11**(4) (2008) 499–503
8. Laorden, C., Galán-García, P., Santos, I., Sanz, B., Hidalgo, J.M.G., Bringas, P.G.: 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, Springer (2013) 261–270
9. Kontostathis, A.: 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. (2009)
10. Smets, K., Goethals, B., Verdonk, B.: Automatic vandalism detection in wikipedia: Towards a machine learning approach. In: AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy. (2008) 43–48
11. Tan, P.N., Chen, F., Jain, A.: Information assurance: Detection of web spam attacks in social media. In: Proceedings of Army Science Conference, Orland, Florida. (2010)
12. Simanjuntak, D.A., Ipung, H.P., Lim, C., Nugroho, A.S.: Text classification techniques used to faciliate cyber terrorism investigation. In: Advances in Computing, Control and Telecommunication Technologies (ACT), 2010 Second International Conference on, IEEE (2010) 198–200
13. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. In: International Conference on Weblog and Social Media-Social Mobile Web Workshop. (2011)
14. De Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining e-mail content for author identification forensics. ACM Sigmod Record **30**(4) (2001) 55–64
15. Holmes, G., Donkin, A., Witten, I.H.: Weka: A machine learning workbench. In: Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on, IEEE (1994) 357–361
16. Garner, S.R., et al.: Weka: The waikato environment for knowledge analysis. In: Proceedings of the New Zealand computer science research students conference, Citeseer (1995) 57–64
17. Wilbur, W.J., Sirotkin, K.: The automatic identification of stop words. Journal of information science **18**(1) (1992) 45–55
18. Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint Conference on artificial intelligence. Volume 14., Lawrence Erlbaum Associates Ltd (1995) 1137–1145
19. Salton, G., McGill, M.: Introduction to modern information retrieval. McGraw-Hill New York (1983)
20. Baeza-Yates, R.A., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1999)