Cybercrime Profiling: Text mining techniques to detect and predict criminal activities in microblog posts

Salim ALAMI
LIIAN laboratory
Sidi Mohammed Ben Abdellah University
Fez, Morocco
salim.alami@usmba.ac.ma

Omar ELBEQQALI
LIIAN laboratory
Sidi Mohammed Ben Abdellah University
Fez, Morocco
omar.elbeqqali@usmba.ac.ma

***Abstract*— The exponential development in online social media allows users around the globe the possibility to share and communicate information and ideas freely in different formats of data via internet. This emerging media has become a dominant communication tool and it has been used as a communication channel in several events, especially "The Arab Spring" and BOSTON'S attack etc. In order to develop useful profiles of different cybercriminals, text mining techniques is an effective way to detect and predict criminal activities in microblog posts taking account the problems of data sparseness and semantic gap. The hashtags used on Twitter (e.g., #arabspring, #BostonAttack) contains outstanding indicators to detect events and trending topics especially to target and detect suspicious topics and eventual illegal events. Similarity approach is used in text analysis to detect suspicious posts in microblog publications. The evaluation of our proposed approach is done within real posts.**

***Keywords*—Cybercrime, Semantic Web, Social Media, Text Analysis, Text Mining, Similarity, NCD Normalized Compression Distance, Profiling, Suspicious Profile.**

I. INTRODUCTION

A large number of malicious people moved to cyberspace, establishing thousands of websites that promoted their criminal activities. Many suspicious web sites were targeted by law enforcement agencies, attacked some of them, and forced their operators to seek new online alternatives.

Malicious people turn from traditional media Web 1.0 to Web 2.0 (shown in Fig. 1), people being previously only readers are becoming contributors to contents, namely the user is no longer a simple consumer of information, but he is also involved in its production using mainly social media within different ways (Wiki, Blog, Micro Blog, Social Network ...).

Social media enables anyone to publish or access information; it offers highly interactive platforms through which communities create, co-create, modify, discuss and share content. Virtual communities using social media are increasingly popular all over the world, it is the primary way in which suspicious people take advantage to communicate and coordinate in the sense that they use this technical achievement for illegal purposes. In this vein, establishing the appropriate solution that will be the foundation stone to detect and predict
criminal activities in social media, has become a necessity and a priority, with exploiting all reactions from each replies, retweets, comments, likes, posts…etc., through which we may uncover suspicious behavior and interests of users as well.

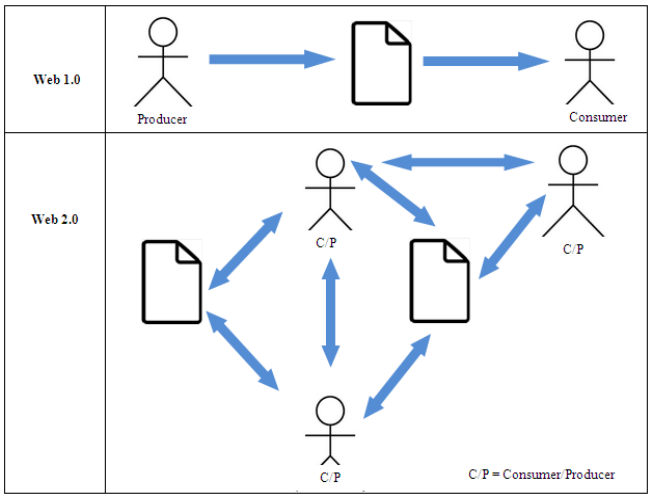

Fig. 1 From Web 1.0 to Web 2.0

This paper presents an extension and improvement of the already proposed solution by S. ALAMI and al. [1]. This already proposed solution presents an idea of our global research project including an automatic system for detecting suspicious profiles in the social media, through which we can uncover suspicious behavior and interests of users as well. The proposed approach is based on the calculation of a similarity distance to distinguish suspicious posts for detecting suspicious profiles within social media. The purpose of the approach is to decompose each post in terms and compare them automatically to predefined suspicious terms database by using similarity distance calculation.

In this work we are going to include a disambiguation step, because many resources can be matched to the same entity that lead to synonymy and polysemy problems in order to add semantics of exchanged information to identify more

978-1-4799-7560-0/15/$31 ©2015 IEEE

significant suspicious profiles and also to improve the system in term of execution time.

To overcome the problem of data sparseness and the semantic gap in shorts text, various approaches have been proposed for adding semantics to text contained in tweets. In our work the hashtags used on twitter contain outstanding indicators to detect events and trending topics especially to target and detect suspicious topics and eventual illegal events.

The paper is organized as following: in section II, we present an over view about cybercrime profiling. In section III, we present related works for methods to detect the meaning of expressions; we also present our proposed approach to analyse posts in section IV. In section V, the evaluation results are given; finally we provide a conclusion and perspectives in section VI.

II. CYBERCRIME PROFILING

In order to develop useful profiles of different cybercriminals, social media contains a large amount of data that must been used to improve the reporting of cybercrime. To understand the new developments of the cybercrime and also establishing the appropriate solution that will be the foundation stone to investigate and prosecute criminal activities, there is an urgent need for cooperation and harmonization of public (e.g. law enforcement) and private sectors (e.g. Facebook, Twitter, YouTube...) to encourage cybercrime reporting. Understanding the steps in the process of committing crime, and understanding the conditions that facilitate its commission, helps us to see how we can intervene to frustrate crime" [2].

Criminal profiling is the process of Investigating and examining criminal behaviour in order to help identify the type of person responsible for wrongdoing [3].

(Johnson, 2005), defined profiling as - An educated attempt to provide…specific information as to the type of individual who committed a certain crime…. A profile based on characteristics patterns or factors of uniqueness that distinguishes certain individuals from the general population [4].

To date, all the national security organizations depend on data and text mining techniques to detect and predict criminal activities, while data mining refers to the exploration and analysis of large quantities of data to discover meaningful patterns and rules. Text mining, sometimes refers to as text data mining, is the process of analysing naturally occurring text for the purposes of extracting and non-trivial patterns or knowledge from unstructured text [5].

S. ALAMI and al. [1] proposed solution in relation to use data and text mining techniques to detect and predict the suspicious published contents with the deduction of the suspicious behavior users on the web. The proposed solution is mainly obtained by representing a major challenge using techniques of text mining, based on the calculation of a similarity distance to detect suspicious posts, which is an effective way to analyse the data published on the internet. The objective of this intelligence data analysis projects is to use data mining to find association and discover relationships among suspect entities based in historical data, in order to predict criminal activities. Data mining is a powerful tool that enables criminal investigator who may lack extensive training as data analysts to explore large database quickly and efficiently. In this publication the authors haven't considered disambiguation step in their proposed framework. In fact, providing an effective way to add semantics to this form of communication requires a disambiguation step, because many resources can be matched to the same entity that lead to synonymy and polysemy problems.

III. RELATED WORKS

Using text analytics to detect suspicious user in social media presents an important challenge. There are various methods to detect the meaning of expressions; many works have been done in this context showing several techniques for text analytics. The lexical matching suffers from many drawbacks; including ambiguity (polysemy and synonymy) and possible lack of specificity(less "meaningful" concepts are identified). Short texts have the characteristics of sparsity, and noisy due to their limited length. When using "bag of words" model to represent short text, contextual information is neglected and hence often leads to synonymy and polysemy problems [6].

To overcome the problem of data sparseness and the semantic gap in shorts text, various approaches have been proposed for adding semantics to text contained in tweets.

Meij et al. [7] proposed an approach to link n-grams to Wikipedia concepts based on various features. Their approach is divided in two steps; in the former they generate a ranked list of candidate concepts for each n-gram in a tweet by applying a various kind of features. In the later, they aim to improve precision by applying supervised machine learning.

Tang and al. [8] have presented a framework for enriching short text for clustering purpose in which they performs multi-language knowledge integration and feature reduction simultaneously through matrix factorization techniques.

Mendes and al. [9] proposed Linked Open Social Signals, a framework that includes annotating tweets with information from Linked Data. Their approach is rather straightforward and involves either looking up hashtag definitions or lexically matching strings to recognize (DBpedia[1]) entities in tweets.

Banerjee and al. [10] proposed a method to enrich short texts representation with additional features from Wikipedia. This method only used the titles of Wikipedia articles as additional external features; it showed improvement in the accuracy of short texts clustering.

Liu and al. [11] focus on Named Entity Recognition (NER) on tweets and use a semi-supervised learning framework to identify four types of entities.

[1] Is a project aiming to extract structured content from the information created as part of the Wikipedia project.

978-1-4799-7560-0/15/$31 ©2015 IEEE

Stephen Guo and al. [12] proposed a structural SVM method to address the problem of end-to-end entity linking on Twitter. By considering mention detection and entity disambiguation together.

Hachey and al. [13] found that it is useful to divide the entity linking task into two phases: search and disambiguation. During the former the system proposes a set of candidates for a named entity mention to be linked to, which are then ranked by the disambiguation. They have also found that much of the variation between NEL systems explained by the performance of their searchers, and the literature on named entity linking has focused almost exclusively on disambiguation.

Laniado and al. [14] have explored the use of hashtags in Twitter and the relation to (Freebase) concepts. Really that hashtags are good indicators to detect events and trending topics. Using manual annotations, they find that about half of the hashtags can be mapped to freebase concepts, most of them being named entities. In a few cases, more general hashtags are mapped to concepts. Assessors showed high agreement on the task of mapping hashtags to concepts. The authors make the assumption that hashtags are mainly used to "ground" tweets, an assumption that we adopt in our work, enabling us to add semantics to tweets with the use of hashtags definition.

IV. OUR PROPOSED APPROACH

Textual data in social media has the problems of data sparseness and semantic gap. One effective way to solve these problems is to integrate semantic knowledge, which has been found to be useful in dealing with the semantic gap [15]. This work takes better into account indicators to detect events and trending topics to detect and predict criminal activities.

Our proposed approach is mainly based on the calculation of a similarity distance to distinguish suspicious posts. The figure below (Fig. 2) shows three stages of our proposition:

- **Text corpus;**
- **Corpus processing** taking into account indicators to detect events and trending topics to detect and predict criminal activities;
- **Classification process** using similarity approach.

We note that the similarity between words is calculating with a predefined suspicious words database. This solution taking into account a hashtag in order to specify the context and organize and search tweets by subject.

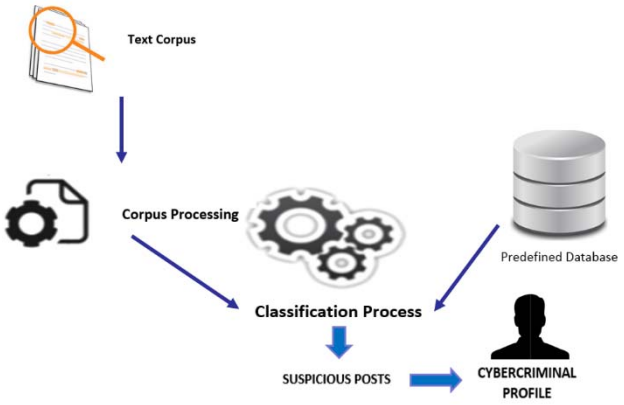

Fig. 2 Our proposed approach

A. Processes

1) Text corpus

Text corpus is a huge and structured set of texts posted in the social media, and different techniques can be employed in this step. In this stage we use dataset [16] of Twitter[2]. It contains 284 million following relationships, 3 million user profiles and 50 million tweets. This dataset was collected at May 2011 and it's very rich of data users' posts.

2) Corpus processing

This stage consists to remove stop words and stemming, in this step we are going to put a mechanism to analyze the "Hashtags".

In computing, stop words are words which are filtered out prior to, or after, processing of natural language data (text). To simplify the study we have to eliminate stop words[3] that contains no useful information, as stop word remove stemming [17] can simplify the processing and reduce errors.

- **Hashtags**

A hashtag is a string of characters preceded by the hash (#) character and it is used to build communities around particular topics. To outside observers, the meaning of hashtags is usually difficult to analyze, as they consist of short, often abbreviated or concatenated concepts (e.g., #Arabspring).

In our work the hashtags used on twitter contain outstanding indicators to detect events and trending topics especially to target and detect suspicious topics and eventual illegal events (e.g., #Arabspring, #BostonAttack).
A hashtag specifies the context and organize and search tweets by subject.

3) Classification process using similarity

The classification stage aims to well organize a set of texts in two classes:

2https://wiki.cites.illinois.edu/wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012

3http://www.lextek.com/manuals/onix/

978-1-4799-7560-0/15/$31 ©2015 IEEE

-Automatic classification method is based generally on the following idea of similarity;
-Two close elements are in the same class and two distant elements are into different classes.

The evaluation of similarities between textual entities (documents, sentences, words...) is one of the central issues for the implementation of efficient methods for tasks such as description and exploration of textual data, information retrieval or knowledge extraction.

B. Mathematical Formulation

In this paper we use the Normalized Compression Distance to detect the similarity between terms that a post contains and suspicious terms collected in a data base.

C (xy) will have the same number of bytes as C(x) when x = y. The more y looks like x the more redundancy will be met by the compressor, resulting in C (xy) bytes coming closer to the number of bytes of C(x) [18].

The obtained distance of similarity is expressed by:

$$NCD(x,y) = \frac{C(xy) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}}$$

Where **0 <= NCD(x, y) <= 1.**

If **NCD(x,y) = 0**, then x and y are similar, if **NCD(x,y) = 1**, they are dissimilar. The distance is used to cluster objects.

The idea of our approach is to analyze sentences posted by users in social media.

First and foremost we extract the characters preceded by the hash (#) character (usually used to build communities around particular topics). We analyze this hashtags for detecting events and trending topics especially to target and detect suspicious topics and eventual illegal events. Then, we decompose each post in terms and compare them automatically to suspicious terms.

If the hashtag contains suspicious information (Suspicious topics), we defined a threshold that we call "a" determining the maximum values of the distance comparison allowing us to conclude that the two terms are similar.

If a sentence contains two terms (suspicious words) which presents similarity with the terms of our database we classify as suspicious post.

The figure below (refer to Fig.3) shows an example of detecting of suspicious post using similarity processing.

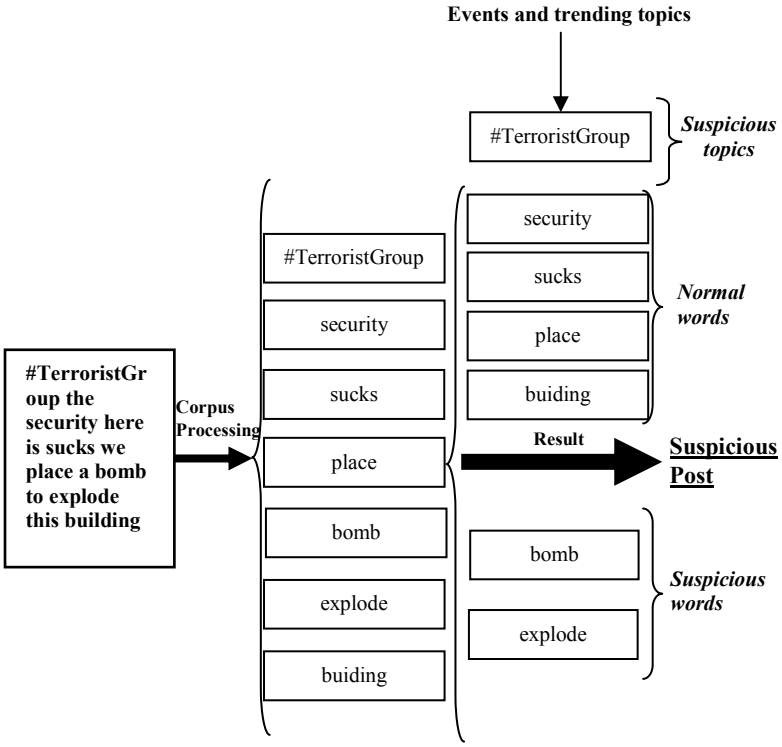

Fig. 3 Example of detecting of suspicious post using similarity processing including hashtags analyses.

V. EVALUATION OF RESULTS

We consider this example "**#TerroristGroup the security here is sucks we place a bomb to explode this building**". After Corpus Processing step, we tested our system using this sentence and we detected that the hashtag contains illegal information "**#TerroristGroup",** its mean that this hashtag talks about suspicious topics, then two suspicious words are detected which are: bomb and explode.

This table (refer to table 1) shows the obtained result of NCD calculations, between words posted and predefined words in our database.

TABLE 1: RESULTS OF NCD CALCULATING BETWEEN SIMILAR WORDS

Term 1	Term 2 (Database)	NCD
#TerroristGroup	**TerroristGroup**	0
bomb	Bomb	0
explode	Explode	0

In evaluation of results, presented in table 1, are conducted based on textual description of each terms, we note that the similarity distance is important when the two terms are not similar and tends to 0 if the two terms are equals. The purpose of our approach is to decompose each post in terms and compare them automatically to predefined suspicious terms database by using similarity distance calculation.

978-1-4799-7560-0/15/$31 ©2015 IEEE

VI. CONCLUSION AND PERSPECTIVES

Our proposed approach is based on the calculation of a similarity distance to detect and predict criminal activities in microblog posts. The purpose of our approach is to decompose each post in terms and compare them automatically to predefined suspicious terms database by using similarity distance calculation.

This paper presents an extension and improvement of the already proposed solution. The already proposed solution presents an idea of our global research project including an automatic system for detecting suspicious profiles in the social media, through which we can uncover suspicious behavior and interests of users as well.

In this paper we are focused on including a disambiguation step, because many resources can be matched to the same entity that lead to synonymy and polysemy problems in order to add semantics of exchanged information to identify more significant suspicious profiles and also to improve the system in term of execution time.

For future work, we plan to improve the system in term of execution time, developing new scoring methods for disambiguation and using other knowledge resources in order to improve the precision rates.

REFERENCES

[1] S. ALAMI, O. ELBEQQALI «DETECTING SUSPICIOUS PROFILES USING TEXT ANALYSIS WITHIN SOCIAL MEDIA" Journal of Theoretical and Applied Information Technology, Vol.73 No.3, p. 405–410, 31stMarch 2015.

[2] Wilson, R. 2006, Understanding the Perpetration of Employee Computer Crime in the Organizational Context. Working paper no.04-2006.

[3] Turvey, B., (2002). Criminal Profiling, An Introduction To Behavioural Evidence. UK, Elsevier.

[4] Johnson, T. A., (2005). Forensic Crime Investigation. USA, CRC Press.

[5] Kanellis P., Kiountouzis E., Kolokotronis N., and Martakos D., (2006). Digital Crime and Forensic Science in Cyberspace, Idea Group Inc. (IGI), USA.

[6] J. Tang, X. Wang, H. Gao, X. Hu and H. Liu, "Enriching short text representation in microblog for clustering," Journal of Frontiers of Computer Science in China, pp. 88-101, 2012.

[7] E. Meij, W. Weerkamp and M. d. Rijke, "Adding Semantics to Microblog Posts," in WSDM '12 Proceedings of the fifth ACM international conference on Web search and data mining, 2012.

[8] X. W. H. G. X. H. H. L. J. TANG, "Enriching short text representation in microblog for clustering," springer, 2012.

[9] P. N. Mendes, A. Passant, P. Kapanipathi and A. P. Sheth, "Linked open social signals," WI-IAT '10 Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 01, pp. 224-231, 2010.

[10] S. Banerjee, K. Ramanathan and A. Gupta, "Clustering short texts using Wikipedia," SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, p. 787–788, 2007.

[11] X. Liu, S. Zhang, F. Wei and M. Zhou, "Recognizing named entities in tweets," HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 359-367, 2011.

[12] G. Stephen, C. Ming-Wei and K. Emre, "To link or not to link? a study on end-to-end tweet entity linking," In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,Atlanta, Georgia,June. Association for Computational Linguistics, p. 1020–1030, 2013.

[13] H. Ben, R. Will, N. Joel, H. Matthew and R. C. James, "Evaluating Entity Linking with Wikipedia," journal of Artificial Intelligence, vol. 194, p. 130–150, 2013.

[14] D. Laniado and P. Mika, "Making sense of twitter," ISWC'10 Proceedings of the 9th international semantic web conference on The semantic web, vol. Volume Part I, pp. 470-485, 2010.

[15] X. Hu, N. Sun, C. Zhang and T.-S. Chua, "Exploiting Internal and External Semantics for the Clustering of Short Texts Using World Knowledge," CIKM '09 Proceedings of the 18th ACM conference on Information and knowledge management, pp. 919-928, 2009.

[16] R. Li, S. Wang, H. Deng, R. Wang and K. C.-C. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home locations," in KDD '12: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, USA, 2012.

[17] M. F. Porter. An algorithm for suffix stripping.Program, 14(3):130–137, 1980.

[18] Marc Dommers, "Calculating the normalized compression distance between two strings". January 20, 2009.

978-1-4799-7560-0/15/$31 ©2015 IEEE