Journal of Theoretical and Applied Information Technology
31st March 2015. Vol.73 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DETECTING SUSPICIOUS PROFILES USING TEXT ANALYSIS WITHIN SOCIAL MEDIA

[1]SALIM ALAMI, [2]OMAR EL BEQQALI

LIIAN Laboratory, Sidi Mohammed Ben Abdellah University, Fez, Morocco

E-mail: [1]Salim.alami@usmba.ac.ma, [2]Omar.elbeqqali@usmba.ac.ma

ABSTRACT

The exponential advancement in information and communication technology has fostered the emergence of new channels for online discussion and has also reduced distances between people. Unfortunately, malicious people take advantage of this technological achievement in the sense that they use it for illegal purposes. In social media, the users produce several and various formats of suspicious posts (text, image, video…) and exchange them online with other people. The data in most social media sites are stored in text format, so in this work we will focus only on text posts. Text mining is an effective way to add semantics aspect to this communication's form presenting a significant research challenge. Similarity approach is used in text analysis to detect suspicious posts in social media. The evaluation of our proposed approach is done within real posts.

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

1. INTRODUCTION

Web 2.0 has emerged a huge mass of important data on the web, 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 websites play a key role in current web applications of which three websites in the top 10 sites, as shown in table 1.

Table 1: Internet Traffic Report By ALEXA[1] Website On November 2014

Rank	Website	Rank	Website
1	Google	6	Amazon
2	**Facebook**	7	Wikipedia
3	**Youtube**	8	Taobao
4	Yahoo	**9**	**Twitter**
5	Baidu	10	Qq

We note that Facebook, Youtube and Twitter (the ones with underline bold font in table 1) are the most requested social media websites. We may categorize these social media websites into three main categories: social networking for Facebook, media sharing for Youtube, and micro blogging for Twitter.

Unfortunately, these kinds of social media are also called new communication channels are increasingly used by malicious users who take advantages of this technological feat to publish illegal and suspicious contents (images, videos, texts ...) in order to exchange data online and share ideas that could affect the security of countries or institutions. As an example, the media sharing websites for Youtube allow to publish videos in relation to "how to create a bomb". The social network Facebook and the micro blog Twitter also help criminals to coordinate and manage online suspect actions.

Social media has been used as a communication channel in several circumstances, namely the Egyptian revolution, BOSTON'S attack, etc. Social media offers several and different kind of services allowing users to exchange different formats of data, including image, video, text etc. The text content plays a role of paramount importance. Microblogging services allow users to share news, information and to participate in events through text messages still called "TWEETS".

[1] www.alexa.com

Journal of Theoretical and Applied Information Technology
31st March 2015. Vol.73 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Social media contains valuable evidence [1], it becomes a true means to remove the reality if only there is a good analysis of data flowing across these media. It could be rich sources of digital evidence reflecting the thinking and activities of the real world. In this important flow of data, allowing to detect suspicious profiles around the social media, several problems arise:

-How to discover the suspicious published contents and publications posted by users?

-How to analyze the users' behavior in the social media?

-How to detect malicious people and suspicious posts?

Using text analytics to detect suspicious user in social media presents an important way to resolve these problems.

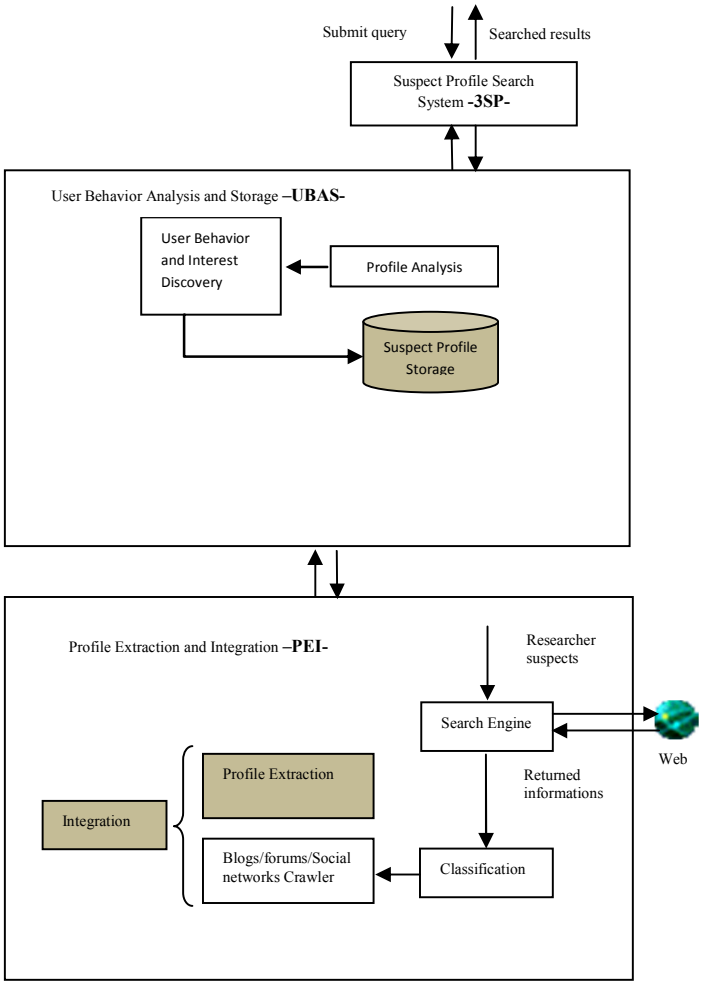

Figure 1: Overview Of Our Research Project

The main objective behind our research project is to develop an automatic system (**S**uspect **P**rofile **S**earch **S**ystem **-3SP-**), shown in figure 1, allowing an efficient investigation of suspicious profiles in the internet, especially in the social media, exploiting all the hints (comments, posts, visited web sites...) through which we may uncover suspicious behavior and interests of users as well.

This research work is mainly focused on two components: the first one is "Profile Extraction Characteristics and Integration" and the second is "User Behavior Analysis and Storage Management".

The first target the extraction and integration of information related to a given profile from the web, and the second target the analysis of user's behavior.

In this work, we focus on publications and text analysis posted generally on social media by users in order to discover the suspicious published contents with the deduction of the suspicious behavior users on the web. This 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 analyze the data published on the internet; the semantic dimension is also involved.

The paper is organized as following: in section 2, we present related works for the text analytics in Social Media. In section 3, we present our system including approach to analyze posts. In section 4, the evaluation results are given; finally we provide a conclusion and perspectives in section 5.

2. 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.

Adamic et al. [2] present a work based on the evolution of responses to well-defined issues. They analyze in a first time the activity of knowledge sharing in yahoo responses. And then divide the discussions according to their characteristics. Some users are interested to sharing expertise forums, others focus on specific topics. In second time they use these functions to trace a map linking the categories and characteristics of the entropy of user interests.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Moreover, Lin et al. [3] was interested in determining the events that are of interests to social networks' users based on their texts data. In this study they collected information from the internet, online communities, and social networks.

Sakaki et al. [4] analyze the real-time interaction of micro blogging events especially on Twitter. In their opinion the user may be considered as a sensor to monitor tweets posted recently and to detect different events.

With the rapid evolution of content and communication styles in social media, text is changing too. Different from traditional textual data, the text in social media is not independent and identically distributed (i.i.d.) data anymore. A comment or post may reflect the user's interest, and a user is connected and influenced by his friends.

Based on internet users's feelings that were measured from their texts, Connor et al. [5] was able to investigate their political opinion as well as their confidence.

On the other hand Lerman et hogg [6] were able to predict news popularity using a model that measures the interaction between users in social networks based on several criteria such as the design of websites and previous votes.

Lu et al. [7] have improved operation of review's quality predictions. Previous works considered the review as a standalone document and defined characteristics from this one. Against, Lu et al. use contextual data on the social network of the author to predefine the quality of the examination.

Gesche et al. [8] present their experiences with large scale semantic databases. They use these to semantically enrich small French texts for which statistical methods show a poor performance. Besides, the use of multilingual resources allows us to circumvent the fact that most resources are in English language.

All the presented techniques propose different and several methods for text's analysis, the table bellow (refer to table 2) shows the comparison within some relevant criteria:

Table 2: All Methods Comparison Of Text's Analysis

	C1	C2	C3	C4	C5
Our Approach	*X*	*X*		*X*	
Knowledge Sharing and Yahoo Answers.			*X*		
A Statistical Model for Popular Events Tracking in Social Communities.	*X*		*X*		
Earthquake shakes twitter users: real-time event detection by social sensors.			*X*		
From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.			*X*		
Using a Model of Social Dynamics to Predict Popularity of News.				*X*	
Exploiting Social Context for Review Quality Prediction.				*X*	*X*
Supervised semantic classification of French newspapers.					*X*

Table legend:
C1: Analyzing text by using text representation.
C2: Using similarity approach to compare texts.
C3: Predictions based on simply extrapolating from users' interactions.
C4: Using learning approach to detect nature of text.
C5: Using data mining approach to analyze texts.

The analysis of this table shows that presented techniques propose typical characterization methods for text's analysis, all these presented approaches take into account just one or two criteria.

Our proposed approach is based on using similarity approach to compare texts (C2), in addition to the use of the analyzing a text by using text representation (C1), and the use of learning approach to detect nature of text (C4).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

All these criteria C1, C2 and C4 will be used in our approach to perform text mining and to detect suspicious posts in social media. We mention that similarity approach to compare texts (C1) is mainly based on similarity calculation distance to distinguish suspicious posts.

3. PROPOSED APPROACH

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

-Text corpus.
-Corpus processing.
-Classification process using similarity approach.

We note that the similarity between words is calculating with a predefined suspicious words database.

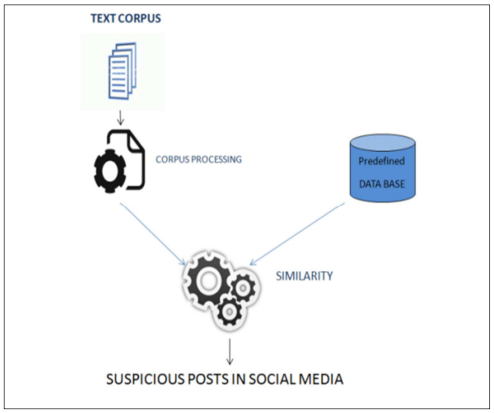

Figure 2: Our Proposed Approach

3.1 Processes

3.1.1 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 [9] 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.

3.1.2 Corpus processing

This stage consists to remove stop words and stemming.

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 [10] can simplify the processing and reduce errors.

3.1.3 Classification process using similarity

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

-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.

To measure the similarity or dissimilarity between objects, the notion of distance is mandatory and depending on the nature of the data [11].

To measure the similarity or dissimilarity between objects, a notion of distance, depending on the nature of the data is requisite. If these objects are expressed by numeric variables such as age, number of subjects, etc, distances such as Euclidean distance, Manhattan distance (city block), Chebyshev distance, Minkowski distance, etc, are preferably used [12].

However, to represent simple distances between variables, same category as colors, animal families, etc, it will turn to the Jaccard distance or Hamming [13].

3.2 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) [14].

The obtained distance of similarity is expressed by:

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

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

Journal of Theoretical and Applied Information Technology
31st March 2015. Vol.73 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

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

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.

We decompose each post in terms and compare them automatically to suspicious terms.

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 figure 3) shows an example of detecting of suspicious post using similarity processing.

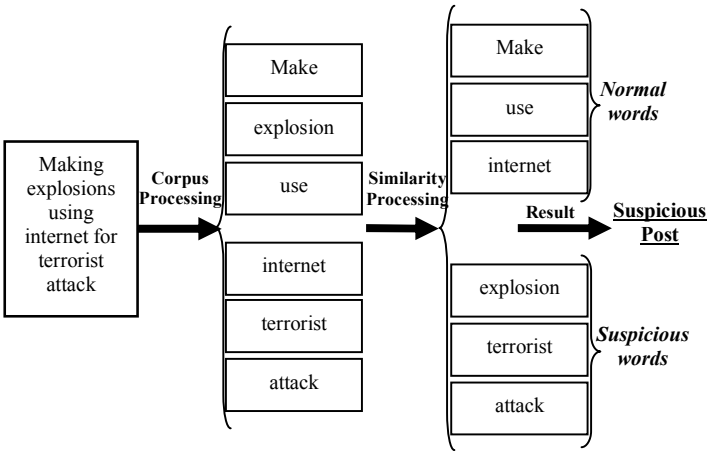

Figure 3: Example Of Detecting Of Suspicious Post Using Similarity Processing

4. EVALUATION OF RESULTS

We consider this example "Making explosions using internet for terrorist attack". After Corpus Processing step, we tested our system using this sentence and we detected three suspicious words which are: explosion, terrorist and attack.

This table (refer to table 3) shows the obtained result of NCD calculations, between words posted and predefined words in our database.

Table 3: Results Of NCD Calculating Between Similar Words

Term 1	Term 2 (Database)	NCD
explosion	explosion	0
terrorist	terrorist	0
attack	attack	0

If we compare these three words with other words that aren't similar we find that NCD tends to 1, like shown in table 4.

Table 4: Results Of NCD Calculating Between Different Words

Term 1	Term 2 (Database)	NCD
make	explosion	0,3
internet	terrorist	0,34
use	attack	0,8

In evaluation of results, presented in table 3 and 4, 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.

5. CONCLUSION AND PERSPECTIVES

The advances in digital and multimedia technology are significantly impacting human behaviors and social interactions.

The main idea of our global research project is to develop an automatic system for detecting suspicious profiles in the social media, through which we can uncover suspicious behavior and interests of users as well.

With the lack of information retrieval analysis from social media, the real performances of social media retrieval data analysis remain debatable. Our 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 our approach is to decompose each post in terms and compare them

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

automatically to predefined suspicious terms database by using similarity distance calculation.

In this paper, we have focused to present a system for detecting suspicious posts in social media using similarity approach in text analysis. Our approach is based on similarity with comparing social media seized posts with a suspicious predefined database.

For future work, we plan to improve the system in term of execution time, developing automated classification and using other knowledge resources in order to improve the precision rates, the semantic of exchanged information will be used to identify more significant suspicious profiles.

REFERENCES:

[1] Mark Pollitt, PhD Associate Professor, "The Narratives of Digital Evidence". AAFS 66th Annual Scientific Meeting, Seattle, WA February 17, 2014.

[2] L. Adamic, J. Zhang, E. Bakshy, and M. Ackerman. "Knowledge sharing and yahoo answers: everyone knows something". In Proceeding of the 17th international conference on World Wide Web, ACM, 2008.

[3] C. Lin, B. Zhao, Q. Mei, and J. Han. Pet: "a statistical model for popular events tracking in social communities". In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.

[4] T. Sakaki, M. Okazaki, and Y. Matsuo."Earthquake shakes twitter users: real-time event detection by social sensors". In Proceedings of the 19th international conference on World wide web. ACM, 2010.

[5] B. Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith."From tweets to polls: Linking text sentiment to public opinion time series". In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media 2010.

[6] K.Lerman and T. Hogg." Using a model of social dynamics to predict popularity of news". In proceedings of the 19th international conference on World wide web WWW'10, pages 621-630, New York, NY, USA, 2010. ACM.

[7] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. "Exploiting social context for review quality prediction". In Proceedings of the 19th international conference on World wide web,WWW'10 New York, NY, USA, 2010. ACM.

[8] Samuel Gesche, Elöd Egyed-Zsigmond, Sylvie Calabretto, Guy Caplat, Jean Beney " Classification supervisée sémantique d'articles de presse en français". Seconde édition de l'atelier Recherche d'Information Sémantique 2010.

[9] 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.

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

[11] Samuel Marquiz "classificateur de Kolmogorov sur le web " 7 Juin 2004

[12] Vincent Levorato ,Thanh Van Le , Michel Lamure, and Marc Bui " Distance de compression et classification prétopologique ".

[13] KAUFMAN L., ROUSSEEUW P. J., "Finding groups in data: An introduction to cluster analysis", WILEYInterscience, 1990.

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