# An extraction and unification methodology for social networks data: an application to public security

**Conference Paper** · December 2017

DOI: 10.1145/3151759.3151836

Amal Abid
Faculté des Sciences Économiques et de Gestion de Sfax
**5** PUBLICATIONS **2** CITATIONS
SEE PROFILE

Hanen Ameur
Institut Supérieur d'Informatique et de Multimédia de Sfax
**10** PUBLICATIONS **41** CITATIONS
SEE PROFILE

Atika Mbarek
Faculté des Sciences Économiques et de Gestion de Sfax
**4** PUBLICATIONS **1** CITATION
SEE PROFILE

Salma Jamoussi
University of Sfax
**94** PUBLICATIONS **268** CITATIONS
SEE PROFILE

# An extraction and unification methodology for social networks data: an application to public security

Amal Abid
abidamal90@gmail.com

Hanen Ameur
ameurhanen@gmail.com

Atika Mbarek
mbarek.atika91@gmail.com

Amal Rekik
rekik.amal91@gmail.com

Salma Jamoussi
salma.jammoussi@isims.usf.tn

Abdelmajid Ben Hamadou
abdelmajid.benhamadou@isimsf.rnu.tn

*Multimedia InfoRmation systems and Advanced Computing Laboratory, MIRACL*

*Digital Research Center of Sfax, CRNS*

*Technopark of Sfax, Tunisia*

## ABSTRACT

Social media are invaded our daily life where millions of users are subscribed. Those online sites provide great tool for users to communicate with others from all over the world, share information, and express their opinion. Unfortunately, this ease of access and availability of information is exploited by malicious users to spread their radical ideas and prepare for terrorist attacks. In this paper, we have proposed a new methodology for data extraction, annotation, and unification in order to identify suspicious content from online social media threatening public security. To the best of our knowledge, there is no specific data collected method from social media devoted to suspicious user profile extraction. Also, implying an expert for data annotation for a security purposes remain a new research task that necessitate expertise and knowledge. In this paper, we tackle this research area by collecting suspicious content from different social media.

## CCS CONCEPTS

• **Web mining** → **Content Analysis**; *Social media data extraction*

## KEYWORDS

Data collection, XML structure, Social data extraction, Structure unification, Social media, Facebook, Twitter, YouTube.

## 1 INTRODUCTION

In the last few years, we are witnessing to a rapidly evolving world in terms of technology and innovative applications. In particular, real time services such as social media sites have widespread in all over the world producing a large amount of highly diverse information. The number of subscribed users is incrementally increasing day by day. The ease of use and the facility of communication make those sources valuable for people to produce, spread, and exchange information in a simple way. Users' profile contains many obvious information such as demographic features (gender, age, native language and location), published content (tweet, comment, or status), etc. Accordingly, user generated content is rich of information and hidden knowledge such as user's opinion and behavior.

Due to this explosive growth of social services and their big information relevance, fetching social media data constitute an emerging research direction that has gathering much more attention in the last few decades. Many Web data extraction techniques have been proposed to extract data from the vast amount of social media [1] to detect and discover new knowledge and interesting patterns for different domain applications [2, 3, 4]. While almost data extraction techniques from social media are targeted for consumer interest analysis and business purposes, an important research area for suspicious content detection and identification from social media is still in its infancy. Specially, with this novel technology that allow extremist persons to use it for malicious purposes. Extremists are increasingly integrated in social media sites that allow them to easily spread their radical ideas and attract more users to their networks. Cyber-security becomes a sensitive domain where it is important to detect and predict abnormal content from social media to protect public security [7]. Social media sites become commonly used by terrorist individuals and organizations that tend to spread their propaganda and recruit more persons to their network. Terrorist organizations, like ISIS (Islamic State of Iraq and Syria) and Al-Qaeda, have proved their online presence on social media, e.g., YouTube, Twitter and Facebook. Particularly, Twitter has recently become among preferred sites for terrorist organizations to disseminate their propaganda. The flexibility of Twitter offers a good environment for terrorists to disseminate their messages and

*iiWAS '17*, December 4–6, 2017, Salzburg, Austria

ACM ISBN 978-1-4503-5299-4/17/12…$15.00
https://doi.org/10.1145/3151759.3151836

spread their ideas through hashtags. Similarly to Twitter, Facebook is a commonly used in many countries. Facebook postings as well as YouTube videos are used to teach the use of explosives [11], the creation of bombs and many other suspicious activities. The YouTube social media is widely used to share video content. The ability to exchange comments about videos and to communicate with other users through sending private messages help users to easily disseminate their ideas and improve the number of their subscribers.

The aim of this paper is to propose a data collection methodology for suspicious content from social media sources. In this work we have focused on the security domain given that there is limited data on this research topic. Our proposed data collection methodology is novel in terms that it collects data from different social media based on a novel keyword search strategy. Also the strength of our method resides on the proposition of a unification model that combines different structure of social media that have different characteristics.

To the best of our knowledge there didn't exist any social data sets for terrorist user profiles. The novelty of our collection method reposes on the following points:

- The collected data is from multiple social media sources that are commonly used by terrorist relying on an incremental list of suspicious keywords.
- Data abstraction and the unification process is novel in terms that it associates different social media sites that have heterogeneous characteristics.
- Terrorist user profile collection is a new research area that allows data scientists (data mining experts) to analyze and participate in fighting against terrorism [5].
- The strength of the annotation step that is the fruit of the all collected data for researchers to understand social media use in critical situations, extract useful information, and predict for suspicious events that can threat public security [6].

This paper is organized as follows: Section 2 describes our data collection methodology. Section 3 is devoted for the data output structure and the proposed unification method. Finally, in the last section (Section 4), we conclude this paper.

## 2 DATA COLLECTION METHODOLOGY

We mainly adopt the Arabic language for our data collection process. Meanwhile, we can consider the Latin words used to express the Arabic language. We are mainly interested by the Islamic terrorism adopted by ISIS also known as Daesh.

We present here the proposed strategy of data collection which contains three major steps: (1) suspicious data extraction from three social media sites (Twitter, Facebook and YouTube), (2) filtering the collected data by keeping relevant information about user profile, and (3) data annotation step realized by an expert "sociologist".

### 2.1 Data Extraction

The data extraction step consists of collecting suspicious content generated by malicious users in social media. This can be published as a tweet, comment, video or Facebook post. Collecting data from social sites is a very challenging task since it is hard to find suspicious users or users supporting ISIS.

In our methodology, we adopt a keyword-based method to extract suspicious data. This method focuses on searching data related to the predefined keywords. To do so, we initially prepare a set of keywords which their use is judged dangerous according to the expert (Islamic state, Sharia, ISIS, etc.). This set of keywords will be automatically incremented and enriched by other keywords most frequently occurring with the basic keywords initially selected. In fact, we keep a list of keywords that are above a given threshold and we re-select content talking about these new keywords.

#### *2.1.1 Twitter data collection.*

For accessing Twitter services, we have used the open source Twitter4j[1] java library. In addition to the keywords based method described above to collect suspicious posts, we have proposed another interesting method designed to collect data in critical situations. Collecting useful information during crises and critical situation is important to understand the behavior and the reactions of users towards a given attack. However, some users are interested just in spreading information and among them we found generally media which is interested to inform the public. From the other side, some users are involved to express their opinion, influence other, or to gloat about the attack victims. This last type of users is very interesting to consider because generally those persons are radical and are sometimes involved in such attack. Those malicious users are targeted and their profile is considered as a dangerous profile. Keywords are mainly extracted from the free encyclopedia Wikipedia [2] that offers different pertinent information from multiple sources. In the last few years, many terrorist attacks are occurred in different places in the world. The first step was to choose the corresponding attack event. After that a simple research on Wikipedia provides relevant keywords related to this attack (place, date, names of attackers, etc.).

In order to target suspicious events from Twitter, we adopt other strategy for Tweet collection based on occurred terrorist events. We have selected attacks happened in: Bardo, Istanbul, Nice, Sousse. For each event, we select tweets seven days before and one month after. The Twitter API limit is it doesn't allow extracting historic data. Hence, we perform the collection of historical data manually through Twitter Advanced Search[3].

This strategy helps in determining implicated users in the event and user that may prepare for that attack (days before).

#### *2.1.2 Facebook data collection.*

An easy way for collecting data from Facebook is provided by Facebook Graph API Explorer[4]. In order to access Facebook API and collect social data, we use Rfacebook-package [10]. This

[1] http://twitter4j.org/en/
[2] http://www.wikipedia.org
[3] https://twitter.com/search-advanced
[4] https://developers.facebook.com/docs/

package provides series of functions that allow R developers to get available information about public pages and groups.

However, in Facebook data collection methodology, our aim in targeting pages or groups is to collect user profiles presenting malicious activities. To do so, we adopted the keyword-based method already described above. As we collected tweets containing these keywords, we collect Facebook data from name of pages and groups or their description that may contain such keyword. In the meantime, we construct an incremental list of keywords that always appear with the initial list. As pages or groups may contain much diversified content, we aim to extract only suspicious posts that contain those keywords. Moreover, some pages or groups are inactively posting content and there is limited number of posts. For this reason, we also adopted event-based method presented in the Twitter data collection section.

After that, we extract only the most active user profiles according to their interactions (shares, comments, and likes). This filtering process is described in the following (Section 3.2). These profiles constitute our corpus where each user is described by his available profile information.

### *2.1.3 YouTube data collection.*

In order to analyze users' channels on YouTube, we have used the Tuber Package on RStudio framework. In fact, the Tuber package [8] provides access to the YouTube API V3 via R. It permits researchers to benefit from several features such as searching for videos with particular content, getting their statistics (number of likes, number of dislikes…) and consulting their comments. Our choice of this package is mainly due to not only the efficiency of his results but also the simplicity of his manipulation.

Using the Tuber package, we adapted a strategy which consists on several steps to analyze potential malicious users' channels on Youtube. In order to target suspicious videos, we searched for the videos that their titles, descriptions or tags contain one of the predefined keywords related to terrorism. Then, we performed a processing step to delete duplicated videos (videos that contain more than one keyword). For each obtained video, we collected the list of videos' comments.

In the following step, through the filtering process, only users that have a clear attitude to the terrorism (either with or against) are kept. In fact, since the user YouTube channel do not contains any personal information about the user and that information may be available on his Google Plus account, we performed a matching step between the user YouTube channel and Google plus account. To do that, we explored the plusser package [9] that provides an API interface to Google Plus. We extracted users' personal information from Google Plus account (sex, language, current location …) as well as his shared posts.

## 2.2 Filtering Data

As our ultimate objective is to collect the available user profile information, we extract users that publish and feedback on suspicious content. However, one of the specificities of the obtained data is the presence of very heterogeneous users with a large amount with millions of extracting profiles. It often contains user in the form of media account which share news about suspicious event and terrorism attacks. Hence, it is very important to accentuate on users who express their personal opinion and their points of view towards suspicious content, as well as, illegitimate users who aim to spread their radical ideas, and plan criminal acts and terrorist attacks.

In order to keep these kinds of users, we perform a filtering step of relevant and active users. The filtering step is carried out in two stages; the first one aims to eliminate media accounts. We prepare an automatic program which uses a predefined list of media names. Furthermore, an expert is intervening to eliminate all irrelevant content. The expert must analyze either the name of the profile and its content.

In the second stage, we specify and extract only active users who are more reacting to the suspicious content and are not media accounts. In our case, we determine active users using a summation of total user activity like share, like, comment or tweets (see equation 1) and we choose users above a defined threshold.

An active user can be defined as follows (Formula 1):

$$UserActivness = \sum \mathrm{User}_{\mathrm{Activity}} \quad (1)$$

Where $\mathrm{User}_{\mathrm{Activity}}$ can be share, like, comment or tweet.

## 2.3 Data Annotation

After filtering the collected data and extracting the available profile information of the chosen active users, we proceed to an annotation step. In this step, we appeal to a sociologist as an expert who plays a key role in defining an annotation strategy based on his observations. He analyzes the information of each user profile and deeply understands his beliefs and behavior in order to determine the degree of his violence and terrorism, his psychological state and his intentions. In some cases, the expert completes missed information not available from the API such as age, gender, and language.

In Table 1, we summarize and describe all annotation labels that are decided after discussion with sociologist. We distinguish four kinds of labels: linguistic feature, psychological feature, user interest and expert report.

# 3 DATA ABSTRACTION AND UNIFICATION

In this paper, we have adopted different data sources for data collection. This heterogeneous content improves the corpus on one side but on the other side it is difficult to homogenize it.

The goal is to transform the extracted data from different social media into structured data understandable by the machine. Even each social media has its structure but the consolidation of different data structure is possible. The first step is to create a comprehensible structure for the machine easy to parse and generate. For this purpose, we have targeted the XML schema to produce a semi structured structure for the collected data. The power of XML resides on its structure that has no predefined tags.

So in order to organize our data, we have to define our specific semantic tags and the document structure.

Table 1: **Data annotation features**

| Features | Annotation labels | Values |
|---|---|---|
| **Linguistic feature** | Native land | The native land is the place where user is from like Tunisia, Egypt, etc. |
| | User Native Arabic | It has a binary value: 1 if the native language of user is Arabic and 0 if not |
| | User Dialect language | The dialect language is user spoken language such as Tunisian Dialect, Egyptian dialect, etc. |
| **Psychological feature** | Psychologist state | Five degrees of user's psychological states: psychiatric disorder, acute disorder, balanced, mental disorder and schizophrenia |
| | Terrorist | Seven terrorism degrees: sympathetic, collaborator, viable to terrorism, against, terrorist, dangerous terrorist and dissident. |
| | Violence | It has a binary value: 1 if user is violence and 0 if not. |
| | Leader | It has a binary value: 1 if user is a leader and 0 if not. |
| **User interest** | User Interests | The preferences and interests of user like religion, politics, sport, etc. |
| **Expert report** | Terrorist Keywords | A list of keywords that design the terrorist aspect of user. |
| | Violence Keywords | A list of keywords that design the violence aspect of user. |
| | Dangerous Content | A list of dangerous tweets with which sociologist can detect the terrorism degree. |

After obtaining different XML data structure from different social media, the goal of the second step is to unify them into common structure. In the following subsections, we present the XML for each social network and the corresponding outlet of the unified structure. Data extracted through these three social media sites (Twitter, Facebook and YouTube) have archived in different XML structures where each one has its own specificities. In order to unify the treatments using these XML files, it seems very important to generate a generic structure for the whole data collected. To do so, we model unified XML-tags by concatenating all tags presented in the three generated XML files of each social media at the head level or body level.

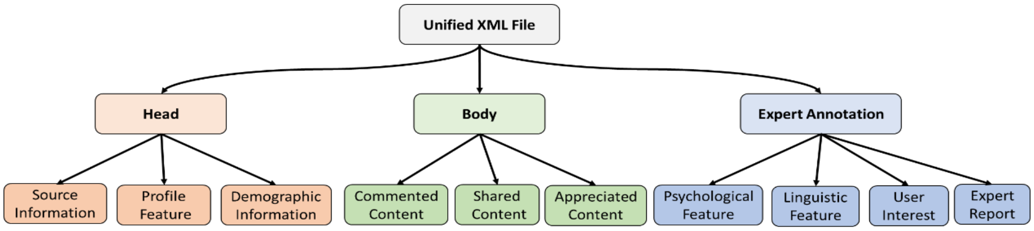

Figure 1: **Unified data structure from Twitter, Facebook and YouTube**

Fig. 1 shows the generic tree of the unified structure. This structure contains three major parts: The first one is the personal information about a user, called "Head". The head part presents source information, demographic information, and profile features. The second one is the body of XML file which contains commented content, shared content, and appreciated content. The third part is dedicated for the expert annotations to define and present psychological feature, linguistic feature user interest, and some expert report or remarks (see Table 1).

Table 2 presents the unified Head part between the three different XML structures of Twitter, Facebook, and YouTube. The goal of this representation is to get homogenous structure for data analysis tasks. The absence of a tag in a social media and its presence in another social media implies its presence in the unified structure. For example, the tag <User_Screen_Name> in Table 2 is present in Twitter and Facebook but not available in YouTube. The same unification process is applied to the Body part related to Commented, Shared and Appreciated content tags.

## 4 CONCLUSIONS

Social media are now involved in our daily life, in which we express our opinion and sentiments. The sparsity of online data makes it easier for malicious users to exploit those social media sites to spread their beliefs and influence others. Extracting data from social media become an interesting task since it is full of knowledge.

In this paper, a data collection methodology is presented targeting suspicious content circulating on the media. The goal of this methodology is to extract malicious content that can threat public security. This paper proposed a novel method to perform this task through the proposition of an incremental list of keywords. Adopting the search keyword method is efficient to find content that is relevant to a particular topic. This strategy helps in improving the extracted content and thus in finding a diversity of user profiles that can be cooperating with terrorist. In addition, we have adopted event-based method for suspicious content extraction.

**Table 2: The unified tags of "Head" part in XML file**

| Unified | Unified Tag | Twitter Tag | Facebook Tag | YouTube Tag |
|---|---|---|---|---|
| **Source Information** | User_ID | User_Id | User_Id | Channel_Id |
| | User_Screen_Name | User_Screen_Name | User_Screen_Name | -- |
| | User_Name | User_Name | User_Name | Channel_Name |
| | Created_Profile | Created_Profile | -- | Date_inscription |
| | Picture_Profile | Profile_Img | Picture | Picture _Channel |
| | Url_Profile | Url_User | -- | Channel_Url |
| | Facebook_Account | -- | -- | Facebook_Account |
| | Twitter_Account | -- | -- | Twitter_Account |
| | Instagram_Account | -- | -- | Instagram_Account |
| | GooglePlus_Account | -- | -- | GooglePlus_Account |
| **Demographic Information** | Range_Age | User_Age | Birthday | Birthday |
| | User_Gender | User_Gender | User_Gender | Sex |
| | Location | Location | -- | Current_Location |
| | Biography | -- | -- | Biography |
| **Profile Feature** | Activity_Count | Activity_Count | Number_Of_Activity | Activity_Count |
| | Keywords | Keywords | Keywords | Keywords |
| | Description_Profile | Description_Profile | -- | Description_Channel |
| | User _Lang | User_Prefered_Lang | -- | Lang |
| | TagLine | -- | -- | Tagline |
| | Skills | -- | -- | Skills |
| | Followers | Followers | -- | Subscribers |
| | Friends | Friends | -- | -- |
| | Likes | Likes | -- | -- |
| | Viewers_Count | -- | -- | View_Count_Channel |
| | Post_Count | Tweet_Count | Shared_Post_Count | Video_Count_Channel |
| | Comment_Count | Retweet_Count | Comment_Count | Comment_Count_Channel |

## ACKNOWLEDGMENT

This publication was made possible by NPRP grant #9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

## REFERENCES

[1] E.Ferrara, P.De Meo, G.Fiumara, and R.Baumgartner. 2014. Web data extraction, applications and techniques: A survey. Knowledge-Based Systems, 70 (2014), 301-323.

[2] M. J.Paul, A. Sarker, J. S. Brownstein, A. Nikfarjam, M. Scotch, K. L. Smith, and G. Gonzalez. 2016. Social media mining for public health monitoring and surveillance. In Pacific Symposium on Biocomputing 2016 (PSB'2016). World Scientific Publishing Co. Pte Ltd, Singapore, 468-479.

[3] H. Wu, Z. Shenghua and L. Ling. 2013. Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33 (2013), 464-472.

[4] F. Grajales, S. Sheps, H. Novak-Lauscher and G. Eysenbach. 2014. Social Media: A Review and Tutorial of Applications in Medicine and Health Care. Journal of Medical Internet Research, 16 (2014), 11-16.

[5] S.Gupta and A. Tiwari. 2016. Terrorism in the Cyber Space: The New Battlefield. International Journal of Advanced Research in Computer and Communication Engineering, 5 (2016), 218-222.

[6] S. Alami and O. El Beqqali. 2015 .Cybercrime profiling: Text mining techniques to detect and predict criminal activities in microblog posts. InProceeding of the 10th International Conference on Intelligent Systems: Theories and Applications (SITA'2015), IEEE, Rabat Morocco.

[7] E. Ferrara, W-Q. Wang, O. Varol, A. Flammini and A. Galstyan. 2016. Predicting online extremism, content adopters, and interaction reciprocity. International Conference on Social Informatics, 9 (2016), 22-39.

[8] G. Sood. 2017. Package 'tuber'. http://cran.r-project.org/web/packages/tuber/index.html. Last checked on 28 Mai 2017.

[9] C. Waldhauser. 2014. Package 'plusser'. http://cran.r-project.org/web/packages/plusser/index.html. Last checked on 27 Avril 2014.

[10] P. Barbera 2017. Package 'Rfacebook'. https://cran.r-project.org/web/packages/Rfacebook/Rfacebook.pdf. Last checked on 24 Mai 2017.

[11] W. Gabriel. 2010. Terror on facebook, twitter, and youtube. *The Brown Journal of World Affairs*, *16*(2), 45-54.