The Significance of Online Monitoring Activities for the Social Media Intelligence (SOCMINT)

Elena Șușnea, Adrian Iftene

Abstract

In the last years, social networks have increased in active members who upload and share postings about daily activities (pictures, comments, news, likes etc.), but also use these networks to get informed. During a crisis situation, the occurrence of an unexpected event can generate immediate reactions from active members, which further result in a huge amount of text and other digital resources related to that event. This event can be the trigger factor for crisis escalation with unwanted consequences.

Keywords: SOCMINT, OSINT, Web mining, early warning, crisis, Twitter.

1 Introduction

A great advantage of social networks consists in the timely dissemination of information about the new event, as opposed to the classical sources of news, such as television, radio, and newspapers. From this perspective, the monitoring of social networks is a very important activity for both crisis managers and the intelligence community, because getting timely information about an event may reduce the undesirable effects, through specific actions that lead to crisis stabilization and even to its regression. For example, one unexpected event may occur during night time. In such situations, most open sources information is not updated, as in the case of TV channels which usually replay different shows until morning when

new information arrives. Instead, social networks begin to "buzz" by providing information about what happened. Despite the fact that this information is not very accurate, it can be an early warning for both crisis managers and the intelligence community as it offers a snapshot of the event.

2 The Challenges of Social Media for Intelligence

Social media intelligence (SOCMINT) is "the latest member of the intelligence family" [1], joining HUMINT, SIGINT, IMINT, MASINT and OSINT. SOCMINT is recently coined term for confluence of ideas from open source intelligence (OSINT) and Web mining technique (machine learning and database methods) applied to social media data in order to identify and understand those situations from social media environment characterized by behavior of individuals that would affect national security, and accordingly try to make rational decisions to bring the situation to the desired state.

A combination of factors like increasing scale of threats of violence such as terrorism, and the economic and political instability in the Middle East and North Africa have contributed to the increasing flow of migrants and consequently, they will communicate more and more with each other using social media. Also, in the thinking of many European citizens, terrorist threats and the crisis of refugees and migrants are largely tied to each other. A series of terrorist attacks - most claimed by the Israeli Islamic and Muslim Islamic Group (ISIS), a terrorist group in which most refugees and migrants have left or will leave when in their countries of origin - have eroded European public support for refugees and migrants. The results of the survey conducted by the Pew Research Center in July 2016 [2] highlighted the decline in support from European nations of refugee and migrant flows. In eight of the ten European nations participating in the survey, more than 50% of respondents consider that the influx of refugees and migrants will increase the prospects of terrorism. The same study points out that the threat of terrorism and the crisis of refugees and migrants is largely linked to one another in the thinking of many European citizens.

The series of terrorist attacks that have taken place in Europe over the past three years and the UK's decision to leave the European Union (EU)

have diminished the public's confidence in European security. This anxiety manifests itself not only in real life but also in the virtual environment. Social media sites - blogs, micro-blogs, and social networking sites, among others - are very useful both for migrants and residents due to the facilities available for social interactions and content generating. Applications like Twitter, Facebook, YouTube, etc. abound in distributed materials on this theme. Furthermore, social media sites allow active users to immediately react to unexpected events, like the veritable human sensors. The occurrence of such events (protests and terrorist attacks) can be identified much faster by analyzing social networks rather than through classical news sources.

Therefore, the rapid development of information and communications technology and easy Internet access have not only yielded indisputable benefits but also it brings some vulnerabilities to security environment caused by lack of borders, dynamism and anonymity. Because active members are growing in numbers and they upload and share a variety of electronic resources, it is very difficult for intelligence analysts to identify and monitor all those individuals who would affect national security. It's called big data and is characterized by "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" [3]. From this perspective, applying data mining techniques on social media will conduct to extract timely and relevant information needed for intelligence process.

3 Social Media Mining: Content, Usage, Structure

In a crisis situation, social media complexity becomes the limiting factor when intelligence analysts want to extract finished intelligence product. Supplying accurate, timely and usable information to those who make national security decisions is essential.

Although there are currently many benefits to using these data mining techniques in the field of information, especially for social media analysis, there are some limitations on usability. One limitation is that, although data mining techniques can help in discovering patterns, they do not tell the user the value or significance of these models. These types of determinations must be made by the information analyst. For example, in

order to assess the validity of a data mining project designed to identify a person suspected of terrorist activities, the user can test a pattern generated on data that includes information about known terrorists. But even if the pattern confirms the profile of this terrorist, it does not mean that the person has to be labeled as terrorist.

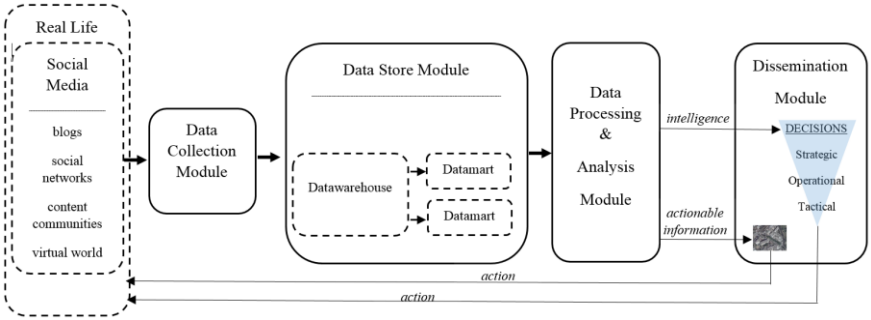

Figure 1. Information system architecture for social media monitoring

The model of the information system for monitoring the crisis of refugees and migrants based on the exploitation of open source resources comprising four individual modules (Figure 1):

(1) The data collection module that allows automatic or semi-automatic collection of social media data (social networks, blogs, microblogs, wiki) for the purpose of digital data collections. From a large data perspective, data collection is difficult to achieve. Not just the huge volume of data is a problem, but their format also generates difficulties in the collection process. Finally, digital collections must contain metadata, text, image, audio, and video files.

(2) The data storage module has the role of cleaning the data collected from multiple open sources and ensuring their efficient access by storing in a data repository. Before the data reaches the data warehouse, it goes through a specific process called extract-transform-load (ETL). Even if there is a centralized warehouse, it does not mean that the data cannot fit in and stored in specific warehouses for each open source category.

(3) The data analysis and intelligence product module convert data into intelligence information. There are three types of tasks for social

media mining. Firstly, Web content mining supports intelligence process to extract actionable information from Web page content. In contrast to one decade ago, the information in the digital universe today is predominantly unstructured (text, images, voice, video, etc.) and increasing the content of the web page was structured (data was organized in tables or relational databases) trace. The techniques used for mining Web content include classification, clustering, language processing, and decision tree. Secondly, Web usage mining focuses on the discovery of patterns from Web usage logs, which stores every click made by a user, such as IP addresses, page references, and date and time of access. In this case, there are some techniques applicable such as association rules, sequence analysis, and information extraction. Lastly, Web structure mining consists of extracting insights from the Web using social network analysis and PageRank. Of the three types of social media mining, we aim to highlight how to use language processing on web content available on the Twitter platform.

(4) The intelligence reporting and dissemination module supports the interaction between policy makers and military decision-makers and intelligence analysts, providing a collaborative virtual space. Final users will access the computer system via a Web interface. The interface will allow access to the computer system for any authorized user by using a web browser. Collaborative virtual space is very important for the creation of the finished product by information analysts.

4 Real Time Mining Twitter Content for SOCMINT

Using Twitter for social media monitoring offers some benefits, such as: The Twitter platform is used by different people to express their opinion on the various topics or events that take place, so it is a valuable source of opinion for the intelligence community and implicit for policy-makers; the Twitter app contains an enormous number of tweets and expands every day so that the volume of text you collect can be extremely high in this way the accuracy of intelligence analysis increases; the Twitter audience differs from other social network users because it is much more varied, from simple people to celebrities, representatives of military or non-

military organizations, politicians, heads of state; the Twitter audience is made up of users from multiple countries. As the Twitter platform publisher grows every day and the services offered expand, the data from this open source can be used in the sentiment analysis.

Although the Twitter app is somewhat new, the scientific research on its use in crisis situations has expanded rapidly. For example, the Twitter platform has been successfully used by the general public in mass protests that took place in Iran, Tunisia and Egypt, especially for the exchange of information. Thus, on June 12, 2009 in Iran, or held presidential elections whose results dissatisfied with Iranian citizens. So the next day, an estimated crowd of several million people gathered in the streets around the Azadi Square in Tehran to protest. During the weeks following the disputed elections, the government has tightened control over data networks and Internet gateways, so traffic has been greatly reduced. In this context, the use of Twitter has been particularly important. First, governments blocked access to the Twitter site, so it was accessible only through a proxy server or a text message on a mobile phone. Secondly, street protesters often needed updated real-time information to avoid confrontations with law enforcement. Thus, they used their mobile phones not only to take pictures or videos, but also to disseminate information about what was happening at that time and the specific location, such as the streets that should be avoided due to police presence. This was also the case for Tunisia and Egypt.

We will use Twitter to analyze two very important indicators: geolocation and sentiment analysis. Tweets feeds are important for analysis because data generation can have a very high frequency and algorithms that process them need to do so under very strict storage and time conditions. Twitter users generate about half a billion tweets a day. Some of these tweets are available to researchers and developers via the Twitter API [4]. To analyze these indicators, we will go through the following steps: first we will collect real-time tweets on the basis of which we will analyze geolocation and feelings analysis, then process data in real time with data mining techniques and save them in a database, and at the end we will use the information for the viewing component.

5 Collection data from Twitter

For collecting tweets, we used a web crawler that connects to the Twitter stream and receives, filters and sends tweets continuously for analysis. The visual representation of the geographical locations from which messages were distributed will be done by filtering the received tweets so that we only process those that have the geographic coordinates attached. In addition, we will perform a language filtering because the annotator used in this case study can only recognize emotions in tweets written in English, French, German, and Arabic. The key words used were: NATO, European Union, refugee, civilian, struggle, and peace. We will also take into account the lexical field of each keyword. Data collection took place between 18 March 2017 and 04 April 2017 and resulted in a total of 66,975 tweets. It is impossible for these data to be analyzed individually, but we will propose two examples to be discussed in the following section.

6 Real-time processing and viewing based on keyword search

At this stage, tweets are available as text data and each row contains a tweet. We will analyze some existing tweets in the monitoring system database to exemplify the multitude and variety of information that can be extracted. The first example, a tweet received in March 2017, which contains the geographical coordinates of a reception center for asylum seekers in eastern Slovakia (Figure 2). The message was identified based on the keyword refugee. The huge amount of data we have is compelling us to a more complex approach, namely a big data analysis. In this regard, we used a Naive Bayes classifier [5] that framed this tweet, in terms of polarity, into the neutral class. Messages labeled neutral do not indicate the occurrence of events leading to escalation of the crisis.

1942 { "_id" : { "$oid" : "58ddf591a66d812220b884a" }, "author" : { "$numberLong" : "15852281" }, "id" : { "$numberLong" : "847639541635860529" }, "createdAt" : { "$date" : "2017-03-31T06:22:02.000 +0000" }, "text" : "PUMPED for day 2 in the refugee camps! This dude hannigan_r will get the chance to share his… https://t.co/3S[illegible]16v", "language" : "english", "polarity" : "neutral", "location" : { "longitude" : 21.912222, "latitude" : 48.930556 }, "keywordsFound" : [{ "original" : "refugee", "matches" : ["refugee"] }] },

Figure 2. Example of tweet in the iSIAD database

Another example of the tweets we propose for analysis is presented in the picture below. It contains the words terror, attack, killed and has been automatically identified as having negative polarity (Figure 3).

514 { "_id" : { "$oid" : "58d2fda3a6d81220b33a3" }, "author" : { "$numberLong" : "8041215667836 9" }, "id" : { "$numberLong" : "844680480600" }, "createdAt" : { "$date" : "2017-03-22T22:41:34.000+0000" }, "text" : "Four Killed, 20 Injured In UK "Terror" Attack", "language" : "english", "polarity" : "negative", "location" : { "longitude" : 7.39379883, "latitude" : 9.10616471 }, "keywordsFound" : [{ "original" : "attack", "matches" : ["attack"] }] },

Figure 3. An example of a tweet that refers to a terrorist attack

Indeed, the message is classified correctly from the point of view of polarity because it refers to the terrorist attack that took place on March 22, 2017 in the British Parliament, located in the vicinity of Westminster Palace in London, (Figure 4).

Figure 4. Visualization of results using Google Maps

7 Visualization and decision maker interaction

For visual representation of data, we first clean or remove duplicate tweets because they will induce biases in the classification process. Then we need to remove the punctuation marks and other non-useful symbols (e.g. emoticons, links), as they may reduce the efficiency and affect the accuracy of the overall process. MapReduce [6] is a new parallel programming model, so the classic Naive Bayes algorithm based on sentiment analysis is adjusted to suit the MapReduce model.

In order to classify the collected tweets by opinions we used Sentiment140 API [7], which internally uses a (semi-supervised) trained model. Unfortunately Sentiment140 API only works with Spanish and English languages. In this situation we would have been forced to discard a big part of our collected data. To overcome this situation we decided to make use of a translation API (Yandex) [8] which would translate the tweet's content from its native language into English. Once translated in English the tweet is sent to Sentiment140 API [7].

Also, in order to improve the quality of results, we chose to use the Naive Bayes classifier along with an English lexical SentiWordNet [9] to improve the accuracy of the tweet classification. The classes used to analyze feelings are the following three: positive, negative, neutral.

The map above shows custom bookmarks using hexagons (Figure 4) or circles (Figure 5). The size of the hexagons increases in proportion to the number of published tweets that contain the specified keywords in that particular location. The visualization of hot spots facilitates the understanding by political and military decision-makers of the distribution of the population interested in a particular subject/event. Instead of placing a single marker for each hot spot, we used variable-level markers to represent the data distribution.

When a cluster grows, it affects other clusters on the map, which will decrease proportionally, so the largest clusters being the most visible. To speed the display the clusters who associated a circle with a radius less than a value set by the user are displayed as dots. If the user will use the option of zooming in a given region (Figure 5 left), where he spotted a cluster higher, it will be divided into smaller clusters, until we see only clusters composed of a single tweet (Figure 5 right).

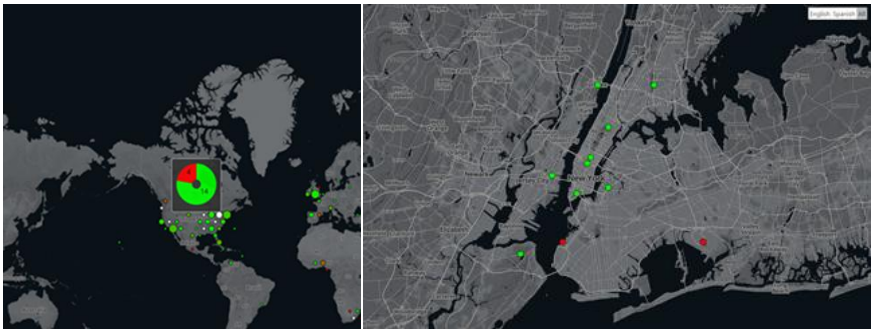

Figure 5. (a) View details of a larger cluster, (b) zoom in for a region from the map

8 Conclusion

In conclusion, real-time analysis of data available from open sources on the Web contributes significantly to the crisis management process, since both the intelligence community and policy and military decision-makers can quickly understand the context of changes that could lead to degeneration of the situation. Tweets analysis can prevent decision-makers in charge of crisis management from occurring unexpected events and develop models that can be used to make proactive decisions to mitigate unwanted consequences.

At the strategic level, SOCMINT can provide indications and warnings about both hostile intentions of crisis-engaging entities and opportunities that should be exploited by decision-makers. The analysis of various sources of information such as regional newspapers in the Middle East, comments posted on various social networks by people coming to Europe or those residing in the countries of destination of refugees and migrants, extraneous audio and video materials distributed in social networks often represent a more robust basis for estimating stability or instability than reports from clandestine sources with a limited accessibility level and a personal perspective that influences the objectively assumed character of the report.

Acknowledgments. POC-A1-A1.2.3-G-2015 program, as part of the PrivateSky project (P_40_371/13/01.09.2016) has supported part of the research for this paper.

References

[1] S. D. Omand. *Introducing Social Media Intelligence (SOCMINT)*. Intelligence and National Security, vol. 27, no. 6, (2012), pp. 801-823.

[2] Pew Research Center, Europeans Fear Wave of Refugees Will Mean More Terrorism, Fewer Jobs, July, 2016, accessed 26 April 2018 at www.pewresearch.org

[3] Gartner IT Glossary, accessed 26 April 2018 at http://www.gartner.com/it-glossary/big-data/.

[4] Twitter API, accessed 26 April 2018 at https://dev.twitter.com/rest/public

[5] S. Russell, P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, (1995)

[6] R. Lämmel. *Google's MapReduce programming model*. In Science of Computer Programming, vol. 70, issue 1, (2008), pp. 1-30, ISSN 0167-6423, accessed 26 April 2018 at https://doi.org/10.1016/j.scico.2007.07.001

[7] Sentiment140 API, accessed 26 April 2018 at http://help.sentiment140.com/api

[8] Yandex, accessed 26 April 2018 at https://www.yandex.com/

[9] SentiWordNet, accessed 26 April 2018 at http://sentiwordnet.isti.cnr.it/

Elena Șușnea[1], Adrian Iftene[2]

[1]"Carol I" National Defense University, Bucharest, Romania
E-mail: esusnea@yahoo.com

[2]"Alexandru Ioan Cuza" University, General Berthelot, No. 16, Iasi, Romania
E-mail: adiftene@info.uaic.ro