JADT' 18

PROCEEDINGS OF THE
14TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

JADT' 18

PROCEEDINGS OF THE
14TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

(Rome, 12-15 June 2018)

Vol. I

UniversItalia
2018

ISBN 978-88-3293-137-2

Program Committee

Ramón Álvarez Esteban: Univ. of León, E
Valérie Beaudouin: Telecom ParisTech, F
Mónica Bécue: Poly. Univ. of Catalunya, E
Sergio Bolasco: Sapienza Univ. of Rome, I
Isabella Chiari: Sapienza Univ. of Rome, I
François Daoust, UQÀM, Montreal, CDN
Anne Dister, FUSL, Bruxelles / UCL, Louvain, B
Jules Duchastel: UQÀM, Montreal, CDN
Serge Fleury: Univ. Paris 3, F
Cédrick Fairon: UCL, Louvain, B
Luca Giuliano: Sapienza Univ. of Rome, I
Serge Heiden, ENS, Lyon, F
Domenica Fioredistella Iezzi, Univ. of Tor Vergata, I
Margareta Kastberg, Univ. of Franche Comté, F
Ludovic Lebart: CNRS / ENST, Paris, F
Jean-Marc Leblanc: Univ. of Créteil, F
Alain Lelu: Univ. of Franche Comté, F
Dominique Longrée, Univ. of Liège, B
Véronique Magri: Univ. of Nice Sophia-Antipolis, F
Pascal Marchand: Univ. of Toulouse, F
William Martinez: Univ. of Lisboa, P
Damon Mayaffre: CNRS, Nice, F
Sylvie Mellet: CNRS, Nice, F
Michelangelo Misuraca: Univ. of Calabria, I
Denis Monière: Univ. of Montréal, CDN
Bénédicte Pincemin: CNRS, Lyon, F
Céline Poudat: Univ. of Nice Sophia-Antipolis, F
Pierre Retinaud: Univ. of Tolouse, F
André Salem: Univ. Paris 3, F
Monique Slodzian: Inalco, F
Arjuna Tuzzi: Univ. of Padua, I
Mathieu Valette: Inalco, F

Organising Committee

Domenica Fioredistella Iezzi: Univ. of Tor Vergata, I
Sergio Bolasco: Sapienza Univ. of Rome, I
Livia Celardo: Sapienza Univ. of Rome, I
Isabella Chiari: Sapienza Univ. of Rome, I
Francesca della Ratta: ISTAT, I
Fiorenza Deriu: Sapienza Univ. of Rome, I
Francesca Dolcetti: Sapienza Univ. of Rome, I
Andrea Fronzetti Colladon: Univ. of Tor Vergata, I
Francesca Greco: Sapienza Univ. of Rome, I
Isabella Mingo: Sapienza Univ. of Rome, I
Michelangelo Misuraca: Univ. of Calabria, I
Arjuna Tuzzi: Univ. of Padua, I
Maurizio Vichi: Sapienza Univ. of Rome, I
Francesco Zarelli: ISTAT, I

Local Organisation

Francesco Alò, Giulia Giacco,
Paolo Meoli, Vittorio Palermo, Viola Talucci

Table of contents

Abstracts

Improving Collection Process for Social Media Intelligence: A Case Study

Luisa Franchina[1], Francesca Greco[2], Andrea Lucariello[3], Angelo Socal[4], Laura Teodonno[5]
[1]AIIC (Associazione Italiana esperti in Infrastrutture Critiche) President – blustarcacina@gmail.com
[2]Sapienza University of Rome – francesca.greco@uniroma1.it
[3]Hermes Bay Srl – a.lucariello@hermesbay.com
[4]Hermes Bay Srl – a.socal@hermesbay.com
[5]Hermes Bay Srl – l.teodonno@hermesbay.com

Abstract
Social Media Intelligence (SOCMINT) is a specific section of Open Source Intelligence. Open Source Intelligence (OSINT) consists in the collection and analysis of information that is gathered from public, or open sources. Social Media Intelligence allows to collect data gathering from Social Media web sites (such as Facebook, Twitter, YouTube etc…). Both OSINT and SOCMINT are based on the Intelligence Cycle. This Paper aims to illustrate advantages gained by applying text mining to collection phase of the intelligence cycle, in order to perform threat analysis. The first step for detecting information related to a specific target is to define a consistent set of keywords. Web sources are various and characterized by different writing styles. Repeating this process manually for each source could be very inefficient and time consuming. Text mining specific software have been used in order to automatize the process and to reach more reliable results. A partially automatized procedure has been developed in order to gather information on specific topic using the Social Media Twitter. The procedure consists in searching manually a set of few keywords to be used for a specific threat analysis. Then TwitteR of R Statistics was used to gather tweets that were collected in a corpus and processed with T-Lab software in order to identify a new list of keywords according to their occurrence and association. Finally, an analysis of advantages and drawbacks of the developed method.

Abstract
La Social Media Intelligence (SOCMINT) è una sezione specifica di Open Source Intelligence. L'Open Source Intelligence (OSINT) consiste nella raccolta e analisi di informazioni da fonti pubbliche o aperte. La Social Media Intelligence consente di raccogliere dati da siti Web di social media (come Facebook, Twitter, YouTube ecc.). Sia l'OSINT che la SOCMINT sono basate

sul ciclo di Intelligence. Il presente documento intende illustrare i vantaggi ottenuti applicando tecniche di text mining alla fase di raccolta del ciclo di intelligence, al fine di eseguire analisi delle minacce. Il primo passo per individuare le informazioni relative ad un obiettivo specifico è definire un insieme coerente di parole chiave. Le fonti Web sono varie e caratterizzate da diversi stili di scrittura. La ripetizione manuale di questo processo per ciascuna fonte potrebbe essere molto inefficiente e dispendiosa in termini di tempo. Sono stati utilizzati software specifici di text mining per automatizzare il processo e ottenere risultati più affidabili. È stata sviluppata una procedura parzialmente automatizzata al fine di raccogliere informazioni su argomenti specifici utilizzando il Social Media Twitter. La procedura consiste nella ricerca manuale di un gruppo di poche parole chiave da utilizzare per un'analisi specifica delle minacce. Quindi il pacchetto TwitteR di R Statistics è stato utilizzato per raccogliere i tweet che sono stati raccolti in un corpus ed elaborati con il software T-Lab al fine di identificare un nuovo elenco di parole chiave in base al loro verificarsi e associazione. Infine viene fornita un'analisi dei vantaggi e degli svantaggi della procedura sviluppata.

Keywords: Social Media Intelligence, Twitter, text mining, data collection

1. Introduction

"Open Source Intelligence [OSINT] is the discipline that pertains to intelligence produced from publicly available information that is collected, exploited, and disseminated in a timely manner to an appropriate audience for the purpose of addressing a specific intelligence requirement" (Headquarters Department of the Army, 2010, p. 11-1). OSINT is mainly used in the framework of national security, by law enforcement to conduct investigations, and in business field to gather important information. Social Media Intelligence (SOCMINT) is a specific section of OSINT which focuses on Social Media.

In recent years, with the spread of Internet, and the high amount of readily accessible data, which give a picture of the actual state of things, the importance of OSINT and SOCMINT has grown, becoming a key enabler of decision and policy making. To bring the best out of such flow of data, the intelligence process must take place as a systematic approach structured around clear steps: planning and direction; collection; processing; analysis and production; dissemination. These stages, each of which is vital, create the Intelligence Cycle (CIA - Central Intelligence Agency, 2013). In order to automatically collect data from both the web and the Social Media, OSINT dashboards are being developed (Brignoli et Franchina, 2017).

This paper describes the contribution provided by automated support tools in the collection phase of the Intelligence Cycle from a Social Media (Twitter) on the phenomenon of interest. To capture the real essence of text available and turn data publicly collected into valuable and reliable knowledge, text mining techniques were implemented. To this aim, text mining plays a relevant role as it enables the detection of meaningful patterns to explore knowledge from textual data. As stated by Feldman and Sanger: "Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. In a manner analogous to data mining, text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns" (Feldman et Sanger, 2007, p. 1).

2. The use of Twitter

Twitter is a common Social Media, a microblog mainly for real time information and communication. With Social Media becoming the main tool for informational exchange, in October 2017, Twitter reached about 330 million users (Statista, 2018*).*

Twitter's specific characteristics makes such a social particularly suitable for SOCMINT purposes. Contents can be accessed by anyone, with no need to create an account. Its users interact with short messages called "tweet", whose length is limited to 280 characters and can be embedded, replied to, liked and unliked. Tweet quick nature, which can then be easily compared to SMS (Short Messaging Service) messaging, fosters the use of acronyms and slang, providing a real-time feel as they bring the first reaction to an event. Phrasing can be simple in structure or imply a large amount of hapax.

With Twitter becoming one of the most important web application, it provides a big amount of data and therefore it constitutes a vital source for Social Media Intelligence. Thanks to its characteristics (potential reach, one-on-one conversation, promotional impact), Tweeter gained importance over years in different social fields, from policy, to media communication and terrorism. As a result, it is commonly considered a valuable source to monitor social phenomena and their changing pattern.

3. Case Study

This paragraph illustrates how text mining tools can be integrated into the SOCMINT data collection phase. The aim of the procedure is to select a suitable and limited list of keywords allowing for an effective and efficient information retrieval in order to support the analyst work.

In this case study the analyst was interested in collecting tweets on the criminal and antagonist threat macro thematic that is related to many specific topics as, for example, critical infrastructures or telecommunications. The collection process has to identify a list of keyword able to collect the messages concerning, for example, "the criminal and antagonist threat in relation to critical infrastructures". The process can be illustrated by a cycle of four different steps: selection of keywords related with the specific tropic performed by the analyst; tweets collection; text mining; and verification and list of keywords definition (figure 1).

Figure 1: illustration of automatic process for Twitter's data collection four steps cycle

3.2. Keywords selection

The first step is performed by the analyst and consists in defining a suitable list of words which could be used in order to collect tweets related to a specific thematic, which in our example could be *Critical Infrastructures*. To each X topic there is a set of keywords defining it (X_1, X_2, … X_n), e.g., *railway*, *station*, *airport*. The same topic is made by all possible sets, given by the formula:

$$\forall X \exists \{X_1, X_2 \dots X_n\}: X = \bigcup X_i$$

3.1. Tweets collection

Once the keywords are selected, the second step consists collect data from Twitter repository, e.g. using the twitteR package of R statistics (Gentry, 2016), in order to identify the keywords allowing for the collection of a certain amount of tweets, that in our example was more than one hundred in a day. That is, a word could perfectly represent the topic but could be rarely used in the messages, resulting in a collection of a small sample of tweets. The aim of this step is to find these words that allows for an effective data collection ($n \geq 100$), eliminating those words that are rarely used in the messages (n < 100). That makes information retrieval more effective as the number of keywords that can be used is limited.

3.3. Text Mining

After the keywords' data collection efficacy was checked, a ten day messages collection was performed including the retweets (49,3%), which is the data retrieval maximum limit of the Twitter repository. The large size corpus (token = 284.253) of 19.491 tweets was cleaned and pre-processed by the software T-Lab (Lancia, 2017) in order to build a vocabulary (type = 19.765; hapax = 8.947) and a list of content words (nouns, verbs, adverbs, adjectives) (table 1). Then the list of content words was checked in order to identify the new keywords and to implement the list.

Table 1: List of the first 20 lemmas of the list

Word	n	Word	n	Word	n	Word	n	Word	n
stazione	6066	elettrico	2226	treno	1198	via	825	ferrovia	659
aeroporto	4734	nuovo	1581	regione	1025	Milano	731	repubblica	632
impianti	3605	rifiuti	1536	Zingaretti	1022	autorizzare	720	giorni	627
Roma	3337	comune	1317	aiutare	896	Italia	679	centrale	605

In order to perform a content analysis, keywords were selected. In particular, we used lemmas as keywords filtering out the lemmas below ten occurrences. Then, on the tweets per keywords matrix, we performed a cluster analysis with a bisecting k-means algorithm (Savaresi et Boley, 2004) limited to twenty partitions, excluding all the tweets that did not have at least two keywords co-occurrence. The eta squared value was used to evaluate and choose the optimal solution.
The results of the cluster analysis show that the keywords selection criteria allow the classification of 98.53% of the tweets. The eta squared value was calculated on partitions from 3 to 19, and it shows that the optimal solution is 13 clusters ($\eta 2$ = 0,19) (figure 2). Then, the analyst controlled for the lexical profile of each cluster in order to detect the words useful to focus data collection by means of the Boolean operators.
This procedure allows for the identification of a short list of most used words (about 20) with regard to both the macro thematic and the related topic. The list of keyword was then further reduced and it was reached a set off five meaningful words for each intersection of the macro thematic with a specific topic. Such a reduction stems from the fact that the use of a bigger amount of words led to an exponential increase of false - positive production rate.

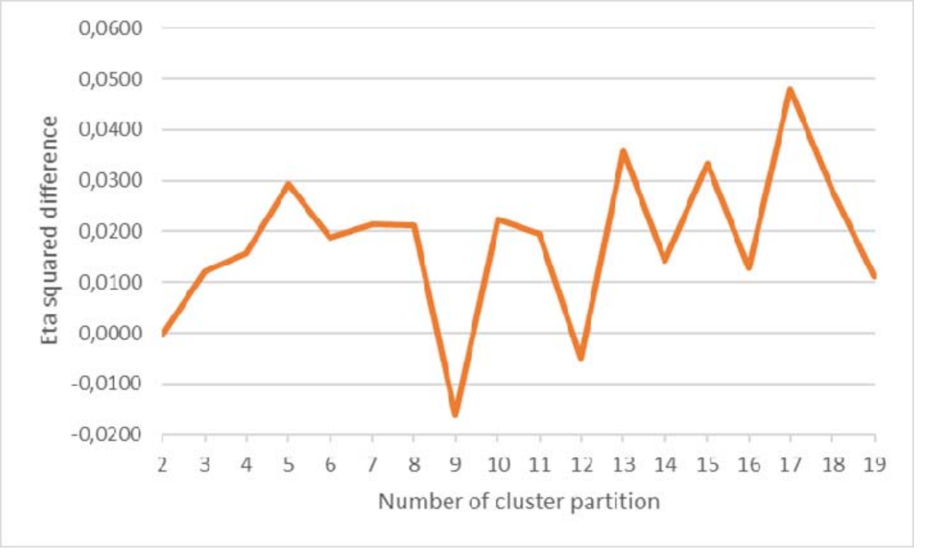

Figure 2: Eta squared difference per partition

As abovementioned, though such a work methodology effectively enables to extract more often used words, with regard to Twitter it is still necessary to test keywords to delete "noise" they produce, which however will not be eliminated entirely. In other words, this methodology affects keywords' amount on the basis of redundancies used by users. However, keywords' quality should be tested in Twitter search engine in order to reach a level of acceptance which includes both false and positive negative. Such words made up the vocabulary to be used to identify intersection between the macro thematic and a specific topic, i.e in the first case "criminal and antagonist's threat with regard to critical infrastructure", in the second case "criminal and antagonist's threat with regard to telecommunication" etc. Between words identified there is an OR relationship. Example: terrorism OR attack OR attack at station OR airport OR railway. Intersection between cluster "criminal and antagonist's threat" and "critical infrastructure is synthetized by the following formula:

$$C = A \cap B = | \quad |(A_i \cap B_i) \neq 0$$

Where A is the cluster "criminal and antagonist's threat", B is "critical infrastructure" and C is the intersection, which is "criminal and antagonist's threat with regard to "critical infrastructures". The following image shows an example.

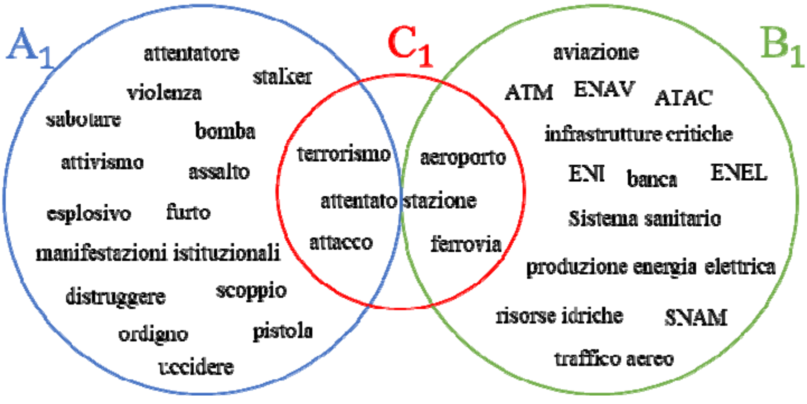

Figure 3: an example of a possible set of words defining the intersection of the cluster "criminal and antagonist's threat", with the topic "critical infrastructure"

3.4. Verification test

Finally, the list of keywords was tested on the Open Source Intelligence dashboard. Collected Tweets were analyzed in order to identify the level of its reliability to monitor the desired phenomena.

4. Conclusion

The developed process reflects the reliability of text mining software in supporting information gathering process for Social Media Intelligence purposes. The vocabulary identified for four different clusters, each of one covering a specific topic, is being tested at this very moment on an advanced dashboard in order to evaluate reliability. However, the role of the analyst is still fundamental. The relationship between OSINT dashboard and analysts must be complementary: dashboard plays a key role in gathering a big amount of tweet, but it is still necessary the analyst support in choosing the suitable keywords to be upload in the database, in order to render information collection more effective. Indeed, OSINT dashboard can't understand Twitter users' use of metaphors and similarities: keywords choice must be made in accordance with monitoring targets. It should be recalled that Italian language is really complex and it might occur that users' language don't refer to chosen target. Let's see a practical example: some keywords which usually refer to criminal threats (bomba - bomb or furto - theft) can be used in Italian language also to refer to synthetic concepts with regard to football or business offers ("bomba" might be used to mean a goal scored through a powerful strike; "furto" might be used to mean that a particular business offer is uneconomical). Another very important issue, which can't be solved without analysts, regard ironic tweets: dashboard collects all information uploaded into database but it can't subdivide tweets into ironic and non-ironic by means of interpretation. To conclude, as dashboards don't understand textual meaning of words, analysts are required to support dashboards' capabilities, being the only ones to interpret the specific meaning of words.

References

Brignoli M. A., and Franchina L. (2017). Progetto di Piattaforma di Intelligence con strumenti OSINT e tecnologie Open Source. Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, pp. 232-241.

CIA, Central Intelligence Agency (2013). Kids' Zone. CIA, https://www.cia.gov/kids-page/6-12th-grade/who-we-are-what-we-do/the-intelligence-cycle.html

Feldman R. and Sanger J. (2006), *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press.

Gentry J. (2016). *R Based Twitter Client*. R package version 1.1.9.

Headquarters Department of the Army (2010). FM 2-0 Intelligence: Field Manual. USA Army, https://fas.org/irp/doddir/army/atp2-22-9.pdf

Lancia F. (2017). *User's Manual: Tools for text analysis.* T-Lab version Plus 2017.

Savaresi S.M. and Boley D.L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis, 8(4):* 345-362.

Statista (2018). Twitter: number of monthly active users 2010-2017. Statista, https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/