Big data and textual analysis: a corpus selection from Twitter. Rome between the fear of terrorism and the Jubilee

Francesca della Ratta, Maria Elena Pontecorvo,
Carlo Vaccari, Antonino Virgillito[1]

Istat – Istituto Nazionale di Statistica, Rome – Italy

Abstract

The exponential growth of web technologies makes Big Data a field of great interest for textual analysis. Twitter, among the social media, best suits the analysis of ideas and contents for its "openness" and "horizontality". However, extracting a textual corpus from Twitter is not an immediate task. The Big Data Sandbox project, promoted as part of the High Level Group at UNECE, aims to check the possibility of using Big Data in the official statistics. The project, started in 2014 and attended by about twenty national and international statistical organizations, focused in 2015 on the analysis of four different sources of Big Data. In particular one group focused on the collection of geo-located tweets. The public interface provided by Twitter is used to extract tweet generated within defined geographic coordinates. Within this project, all tweets generated in the territory of Rome starting from November 2015 are stored, to monitoring activities related to the Jubilee. The dramatic events of November 13 in Paris, quickly attracted the attention of users in Rome: in the context of the global threat of terrorist, the attack on a European city has deeply affected the imagination of Twitter users, also in view of the forthcoming Jubilee, which increased worldwide media exposure of the city. This suggested the opportunity to investigate the connections between Jubilee and terrorism to understand whether among Twitter's users the global threat of terrorism could affect the way of telling the Jubilee.

The aim of this work is to apply some techniques of textual analysis on a corpus extracted from Twitter, to describe its contents and to investigate possible ties between technologies for Big Data analysis and Text Mining. Despite the selected corpus shows a poor connection between the two phenomena in the period of analysis, the analysis supplied interesting possibilities.

Key words: big data; text mining; Paris attacks; Jubilee; sentiment analysis, social media.

1. Introduction

The term "Big Data" is used when a dataset is so large that cannot be processed within reasonable time using conventional tools. A possible definition often used is the following: "Big Data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information." (Techamerica; 2012).

The collection of data coming from Social Networks has focused the interest of researchers working on Big Data - the diffusion of such websites, where users generate a considerable amount of information not otherwise available, makes them one of the most important potential sources for data, including textual data.

Specifically, Twitter is an online social network that enables users to send and read short 140-character messages called "tweets". Tweets are publicly visible by default, but users may

[1] This work comes from a common effort; paragraphs 1 and 7 are written by all authors; par. 2 by Carlo Vaccari; par. 3 by Antonino Virgillito, par 4.1 and 5 by Maria Elena Pontecorvo and 4.2 and 6 by Francesca della Ratta.

subscribe to other users' tweets (i.e. they become "followers"), forward individual tweets to their own followers ("retweet"), or they can "like" (formerly "favorite") them. Created in 2006, Twitter soon became one of the most popular social network, surpassing in 2015 500 million users. According to Alexa (2016), currently Twitter is the ninth most visited site in the world.

Aim of this work is to apply textual analysis to a corpus extracted from Twitter, in order to describe its contents and join together the both worlds of big data and textual analysis. In the context of the Big Data Sandbox project, promoted by the High Level Group on the Modernisation of Official Statistics, we extract a sample of tweets generated in the city of Rome, in order to monitoring the opinions of Twitter users about the Jubilee, started in Rome in December 2015. The dramatic events of 13th November in Paris suggested the idea of studying how these were seen from users in Rome and also finding a possible connections between the two events, in order to understand if the global threat of terrorism could influence the way the Jubilee was perceived and described. Despite the fact that the analysis revealed a loose connection between the two phenomena, it also provided interesting outcomes.

2. Big Data

2.1 Introduction to Big Data

Many authors adhere to the “Four Vs definition” that points to the four characteristics of Big Data, namely Volume, Variety, Velocity and Veracity (Vaccari, 2014). For what concerns the Volume, the quantity of data available continues to increase at an unprecedented rate: in the world more than 2.5 exabytes [10^{18}] bytes of data are generated every day and 90% of the data in the world has been created in the last two years alone. Considered that all mail delivered by the US postal service amounts to 5 petabytes, the same amount of data is handled by Google in just one hour. Still, everyone can agree that whatever is considered “high volume” today, will be even higher tomorrow.

Variety indicates the many different forms that data coming from sensors or social networks or smart devices, can take today: web data, tweets, audio, video, click streams, log files and more. The Velocity at which data are generated today makes impossible the use of traditional systems to capture, store and analyze them. And the quest for high data quality requires to give attention to the level of reliability (Veracity) associated with these new types of data.

The UNECE task team on Big Data in 2012 proposed the following taxonomy to classify Big Data (UNECE, 2012):

- **Social Networks** (human-sourced information): this information is the record of human experiences, previously recorded in books and works of art, and later in photographs, audio and video.
- **Traditional Business systems** (process-mediated data): these process record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc.
- **Internet of Things** (machine-generated data): derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world.

In this work we are interested in data coming from Social Networks, and, among them, Twitter, which, for its "openness" and "horizontality", best suits the analysis of ideas and texts.

2.2 The Big Data Sandbox project

High-Level Group on Modernisation of Official Statistics (HLG-MOS) was set up by the Bureau of the Conference of European Statisticians in 2010 to oversee and coordinate international work relating to statistical modernisation. HLG-MOS started in 2014 a project called Big Data in Official Statistics (UNECE 2014). One of the most important results of the project was the implementation of the “Sandbox”, a web-based environment established at Irish Centre of High-End Computing (ICHEC) to better understand how to use the power of "Big Data" to support the production of official statistics.

In last two years, more than 40 experts from national and international statistical organizations worked together to identify and tackle the main challenges of using Big Data sources for official statistics. The Sandbox gave participating statistical entities the opportunity to test how existing statistical standards / models / methods can be applied to Big Data, determining which Big Data software tools are most useful for statistical organizations and learning more about the potential uses, advantages and disadvantages of Big Data sets.

In 2015 the Sandbox operated working in four main task teams. Every team, composed by multinational members, worked to understand which statistical results could be obtained from a specific Big Data source.

The four selected sources were:

- Wikipedia page views on specific encyclopedia entries;
- Trade data loaded from United Nations global ComTrade Database;
- Enterprise Websites, to verify the potential collection of data about job vacancies;
- Social data from Twitter, loading geolocated data to analyze the mobility of people.

The Sandbox Research focused on the latter dataset, collected first in Mexico and then in UK, to study people mobility inside the country and “sentiment” of people. In this work we reused their data collection method and applied it to data generated in the area of Rome, as described in the following.

3. Collecting Data from Twitter

In November 2015, a data collection process was activated on the Sandbox environment (currently already active) acquiring tweets generated in the city of Rome, with the aim to analyze the movements of tourists during the 2015-2016 Catholic Jubilee.

As mentioned above, two different solutions can be used for collecting data from Twitter: accessing subsets of the published data using the developer tools made available directly by Twitter or either purchasing entire sets of tweets through a service provider. The obvious difference is that the free solution presents limitations with respect to the commercial one in terms of the amount of data that can be accessed. Despite this, the public tools are widely used by researchers, being the commercial service mainly targeted to enterprises that decide to exploit social data to back their business. We will discuss in the following the public APIs highlighting their limitations and those aspects in the terms and conditions that are relevant with respect to data collection for research.

Public access to Twitter data is achieved through two distinct categories of APIs:

- Streaming APIs: allow to monitor tweets in real-time by keeping an connection on which tweets are constantly streamed as long as they are published. It is possible to define filters to restrict the tweets that are returned in the stream.
- REST APIs: allow to perform historical search queries on recently posted tweets and to retrieve lists of users, friends, followers etc.

Access to the API is associated to the personal Twitter account of the developer. Users have to request a personal key for accessing the API and this key will be required at any access, and also used for monitoring the behavior of the user.

The REST API poses limits in the number of requests that can be issued from a same user (currently the limit is maximum 15 requests in a 15 minutes time window) and in the number of the tweets that are returned. Twitter generically states that "Not all Tweets will be indexed or made available via the search interface", although not specifying exactly the size of the sample, however also adding that only recent (last 6-9 days) tweets are considered. In the Streaming API each user account can open only one connection at time to the streaming endpoints. Users are also warned against making excessive connection attempts, a behavior that if detected may cause the user to be banned. Thus the data returned by the Streaming APIs is a non-well defined "sample" of all the public tweets with limitations also present on the possible filters that can be used, with a limit of 400 keywords and 5,000 users that can be tracked in a single connection.

3.1 Structure of a tweet

The data returned by both Twitter APIs is a collection of tweets in JSON format, where each single record is a composite object describing all the aspects of a tweet. All the properties that are "visible" on the Twitter timeline can be accessed, such as user, time and date of creation and obviously the full text. Besides this, other interesting attributes complement the data structure. In particular, each record contains figures related to the user activity are included, such as the number of users that respectively follow and are followed by the author of the tweet. These give an indication of the popularity and relevance of the user and can be exploited to weight the actual impact of a tweet. Other attributes related to geolocation, i.e. the place from where the tweet has been posted, are set according to the settings of the client used and the kind of device. For example, if a tweet is posted from a smartphone and geolocation is enabled in the Twitter client, the latitude and longitude where the tweet was generated are included in the data. If this level of precision is either not possible or not allowed by the user, a generic place (for example the name of the city) is specified.

Geolocated tweets are very interesting for research purposes, and could also potentially allow to track the position of users, although this practice is discouraged by the policies of use of the developer tools. Geolocated tweets can be filtered, limiting the places they are originated from, either by specifying the name of the place (e.g. "Rome") or, on a more refined level, by defining a bounding rectangle on a map whose extremes are latitude-longitude pairs.

3.2 Architecture of the Data Collection

The data collection has been implemented through a set of programs written in R that directly access the Twitter Streaming API. Filters have been set on both the language (Italian) and the location of the tweets, by specifying a bounding rectangle that wraps the area of the city of Rome and some of its surroundings, including the airport of Fiumicino. The downloaded data has been stored in a storage and indexing engine named Elasticsearch. This is an open-source tools that allows to store generic documents in JSON format and automatically builds an index that makes it possible to query the document database in an extremely efficient way. Data was partly processed in R before being stored in Elasticsearch, to make limited processing such as adjustment of the format of time and geolocation attributes. Some auxiliary variables were also computed, such as quantiles for followers, friends and statuses[2].

An open-source visualization tool named Kibana represents a visual front-end to Elasticsearch, allowing to discover and visualize data in a very quick way. Kibana was used to perform a preliminary exploratory analysis and then data was extracted from Elasticsearch using R and downloaded in CSV files that were fed to Taltac2, the tool that has been used for text analysis (Bolasco, 2005).

4. A first application on a corpus extracted from Twitter

4.1 Corpus extraction

Tweets of interest were all those generated in the city of Rome in the seven weeks between November 1st and December 17th. These constitute a big corpus, composed by about 240,000 elements and 3.1 million token. In order to reduce redundancy of the corpus (Ratinaud, 2014), only the original messages (i.e. non-retweeted[3]) were considered. As mentioned above, each tweet included the date, the week of publication and the name of the author. Information about the number of followers and friends were coded according to the quartiles they belong to.

Ranke	Types	Token
1	#XF9	2053
2	#Roma	1790
3	#Rome	1703
4	#GF14	1200
5	#roma	1166
6	#1DIT	1077
7	**#Parigi**	**1039**
8	#trndnl	948
9	#farmaciauno	843
10	#ASRoma	640
11	#rome	615
13	#MTVStars	410
14	#Amici15	407
15	#buongiorno	384
16	#Renzi	383
17	#chilhavisto	380
18	**#ParisAttacks**	**350**
19	**#PrayForParis**	**349**
20	#italy	340
21	**#Giubileo**	**336**
22	#Natale	334
23	#Stigmabase	310
24	#Italia	304
25	#uominiedonne	299

Table 1 - List of first 25th hashtags

The set of selected tweets belongs to a total of 15,682 distinct users, with a narrow distribution: half of the users posted at most two tweets, 40% of users posted a number of tweet between 3 and 24 tweets. Only one tenth of users posted about 77% of tweets, thus confirming the well-known power-law behavior typical of web-related phenomena (Anderson, 2006). A first attempt of selecting the tweets of interest was conducted considering the presence of keywords identified (a priori). Subsequently we decided to exploit the semantic information associated to hashtags[4].

All the hashtags were extracted from the whole set of tweets, obtaining an interesting list that, just by itself, offers an indication of the main themes of discussion

[2] Particularly, for the number of followers, the first quartile is up to 249, the second is in the range 250-826, the third in the range 827-2,252 and the fourth from 2,253 to the maximum value (5,763,988).

[3] A peculiar form of repetition emerged during the analysis, in which a user posted several times the same message, sent to different users. Even if these tweets were not eliminated by the corpus, the terms recognized as belonging to duplicated tweets were not considered.

[4] The hashtag is a labelling system used by some social networks as a theme aggregators, that allows user to find easily messages on a specific topic. Hashtags are denoted by the symbol #.

in the analyzed corpus (***Table 1***). The top of the list, besides *Roma* or *Rome*, is occupied by references to TV shows or bands. The first reference to *Parigi* is in 7th position, while references to *ParisAttacks* or *PrayForParis* appear below in the list (1,738 token in overall). The *Giubileo* (that started in Rome on December 8th) is in 21st position with 336 token. From this list, 138 hashtags were selected, that were connected to either the events in Paris or the Jubilee (111 and 27 respectively). Tweets in which one of such hashtags appeared were selected with the "Entity Search" function of Taltac2: the selected corpus (referred in the following as PARILEO) is composed by about 5,000 tweets, that is the 2% of the total, raising to 7.5% in the week from 13th to 19th November. The corpus includes 76,000 tokens, a vocabulary of 14,982 types and a type/token ratio of 19.6%.

Tweets in the corpus were published by 1,462 users, the 9.3% of the total. Also in this case the distribution is narrow, with an even longer tail with respect to the overall dataset: half of the users posted only one tweet, while the more active 10-tile of the distribution posted 53% of the tweets. This is in line with the reticular dynamics of mass communication, that occur especially in the case of exceptional events or emergencies, where a restricted number of key users emerges rapidly as a driver of the communication exchange, while a long tail of users "will participate only in retwitting or commenting in crisis tweets from time to time, almost randomly" (Bruns and Burgess, 2014). Among the more active users, communication media such as radios or TV channels can be found, as well as experts, opinion leaders or non-profit foundations. With respect to the total distribution of tweets, those belonging to the PARILEO corpus appear to be associated to more "followed" users, as the percentage of the messages posted by users in the fourth quartile (more than 2,253 followers) rises to 32%.

4.2. The PARILEO corpus description

A semantic label has been subsequently associated to each fragment of the corpus, starting from the two lists of hashtags (Paris or Jubilee). Tweets referring to Paris events are the majority (84%) and are concentrated specifically in the week from 13th to 19th November (***Figure 1***), while those dedicated only to the Jubilee are about the 15% of the total. Only a small amount of tweets are dedicated to both themes (1,1%).

The trend of the number of tweets per date highlights the strong emotional impact of the occurrences of 13th November: almost 40% of the total of the analyzed tweets has been published between 13th and 14th November. Conversely, those dedicated to the Jubilee are by far less numerous, and are concentrated around the initial date of the Jubilee (8th December).

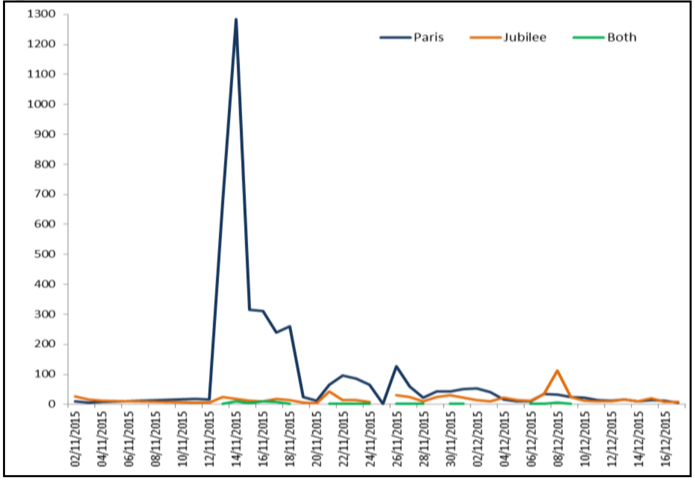

Figure 1 – Temporal trend of the tweets in the PARILEO corpus, divided by theme

These first results provide an answer to one of the questions behind our work, showing a weak connection between the two themes and a modest impact on Twitter of the Jubilee theme. Nevertheless, the selected corpus can inspire interesting reflections about the content on the tweets or, in general, for textual analysis of this kind.

Specifically, we applied traditional

tools of textual analysis to the corpus. One aspect of particular interest was the analysis of repeated segments, that generally offer a first description of the contents of a text. More relevant segments referred to the only events in Paris were classified in 5 categories: news and information (*blitz a Saint Denis, 127 morti, invita a non lasciare pacchi)*; opinions and questions (*chiudere le frontiere, basta buonismo)*; media emotion (*non ho parole, sto piangendo)*; places (*al Bataclan, Piazza Farnese)*; subjects (*foreign fighters, vittime, Valeria Solesin)*. Different and lower informative content, also because of the smaller size of the text, derived from the repeated segments on Jubilee, mainly referred to the person of the Pope or his speeches, or either to other Vatican-related topics (*apertura della porta Santa*, *Basilica di San Pietro*, *pochi pellegrini).*

Informazioni/News		Commozione mediatica/ Media Emotion		Temi di discussione/ Discussion themes		Luoghi/Places	
Segment	Tokens	Segment	Tokens	Segment	Tokens	Segment	Tokens
iran usa francia si uniscono contro Isis	25	sotto attacco	20	finendo il neocolonialismo in mediorente	18	al #Bataclan / al bataclan	43
attacchi terroristici / attacco terroristico	16	persone innocenti	13	chiudere le frontiere / chiusura delle	15	in #Siria	41
strage di #Parigi	13	siamo tutti francesi/ siamo tutti parigini	9	siamo in_guerra	14	piazza Farnese / Palazzo Farnese	17
degli attacchi di #Parigi / attacchi a parigi	13	fa paura	8	armiamoci e partite	13	a bruxelles / grand place	13
un_minuto di silenzio	13	esseri umani	7	in_nome_di Dio / nome di Dio	12	stade de france	12
#Madonna a #Parigi rende omaggio	12	non ho parole	7	maggiore #sicurezza	10	piazza del popolo	11
attentati di parigi / Attentati parigi /	12	nostri fratelli	6	contro il #terrorismo / lotta al #terrorismo	10	interno_del teatro / sala concerti	10
#Parigi sotto attacco	9	parole per descrivere	6	guinzaglio per i cani #infedeli	8	a #Molenbeek / a molenbeek	8
blitz a #SaintDenis / blitz a Saint_Denis	9	sto piangendo	6	terza guerra	6	a Saint_Denis	6
bombardamenti in #Siria	5	vivere nel terrore	6	chi finanzia	5	arabia saudita	5
#PorteOuverte #Paris	4	andare avanti	5	futuro dell' #Africa	5	medio oriente	4
#ValeriaSolesin morta	4	fratelli francesi	5	loro hanno le armi	5	spazio aereo	4
60 ostaggi	4	mie preghiere	5	per una maggiore #sicurezza	5	**Soggetti/Subjects**	
diffonde foto	4	parigi ci appartiene	5	terza guerra mondiale	5	foreign fighters	25
in segno di lutto	4	questa barbarie	5	vorrei un monumento italiano illuminato	5	vittime degli attentati / vittime del	20
invita a non lasciare #pacchi	4	vicino_ai #cittadini	5	vende armi	4	Valeria #Solesin / #ValeriaSolesin /	14
Isis minaccia	4	vite spezzate	5	basta buonismo	4	forze_dell' ordine	12
morta nella discoteca	4	commozione e solidarietà	4	per la #sicurezza	4	vittime di #Parigi / vittime di parigi	16
stato di emergenza	4	coraggio di essere umani	4	prevalere la morte sulla vita	4	Charlie hebdo	8
testimone dell' attentato di parigi	4	fa male /mi viene da piangere	4	vive la france	4	Oriana fallaci	8
		non si può morire	4			polizia belga / polizia francese	8
		grande commozione	4			popolo francese	6
		nostro dolore	4			attentatori di #Parigi	5
		siamo con te	4			famiglie delle vittime	5
		siamo impotenti	4			servizi segreti francesi	4
		sotto assedio	4			abdeslam salah	4
		tante persone innocenti	4			fondamentalisti islamici/ Stato islamico	4

Table 2 – Relevant repeated segments about terrorism

The analysis of segments also allowed to explore some peculiar structural features of the texts extracted from Twitter, such as the use of the # symbol denoting hashtags: while for some users the hashtag is detached from the content and is used only for categorizing the message, in some tweets it becomes an integral and meaningful part of the text (e.g. "*Madonna a #Parigi rende omaggio alle vittime degli attentati*"). In order to reduce the redundancy of the text in the analysis presented in the following, the symbol # in the parsing has been considered as a separator, making the number of types drop down to 14,162.

5. The specific contents of the corpus

A more detailed description of the text can be achieved through the extraction of the keywords, that allow to describe the most relevant themes discussed in the tweets, through the computation of the standardized ratio, a measure of the under- or over-representation of a word (in terms of relative frequency), with respect to a dictionary of frequency chosen as a model (Bolasco, 2013).

Table 3 shows significant keywords, obtained by comparing the text with the lexicon of standard Italian available in Taltac2, mapped to theme categories: report, players (divided into the makers of the terrorism acts, the victims, the society, policemen, countries); emotions (empathy, anger); discussions; places and Jubilee. Moreover, the table includes the main verbal forms, over-represented and in infinite tense.

EVENTI/FACTS			ATTORI/PLAYERS			EMOZIONI/EMOTIONS			DISCUSSIONI/DISCUSSIONS		
Types	Token	Stand. ratio	Types	Tok	Stand. ratio	Types	Toke	Stand. ratio	Types	Toke	Stand. ratio
Le parole della cronaca/ Report			*Gli artefici/ Makers*			*Paura e compassione/ Empathy*					
terrorismo	304	210.6	terroristi/a	148	154.0	carneficina	11	65.0	frontiere	37	32.0
attentati/o	168	170.4	kamikaze	18	86.9	strage/i	63	52.9	psicosi	8	23.3
sparatorie/a	30	100.6	attentatori	16	77.2	terrore	49	43.1	accoglienza	18	21.2
blitz	43	95.8	infedeli	13	48.4	paura	132	40.2	vicinanza	13	21.0
raid	22	74.9	arrestati	11	25.0	guerra	166	37.5	contro	138	18.4
attacchi/o	118	60.1	cellula	6	16.3	brividi	15	35.8	immigrazione	14	15.7
sicurezza	160	49.0	assassini	11	15.7	orrore	26	25.4	petrolio	15	15.6
video	57	43.3	fondamentalisti	6	15.4	sgomento	13	23.1	pace	70	14.3
passaporti	15	37.4	fuggitivo/i	5	14.3	commozione	13	19.8	religione	25	12.7
bombardamenti/o	25	33.9	killer	6	6.6	dolore	50	19.6	asilo	7	12.1
bombe/a	36	24.9	nemici/o	17	5.0	barbarie	11	19.4	guerre	16	12.1
armi	49	22.1	*Le vittime/ Victimes*			spezzate	5	18.4	religioni	9	10.5
tributo	8	19.7	ostaggi	38	84.6	silenzio	47	18.2	colori	17	8.5
esplosioni/e	13	19.7	vittime/a	121	82.8	cordoglio	7	18.1	fanatismo	5	8.4
arresti/o	25	19.1	morti	112	59.6	stragi	17	16.7	fermezza	8	7.5
granate	6	18.6	feriti	22	28.6	lutto	12	15.4	monumento	6	5.9
allarme	22	16.7	ragazzi	88	25.5	scempio	5	14.4	laico	5	5.5
assalto	11	15.4	innocenti	38	25.1	tristezza	13	13.8	tolleranza	5	5.4
arrestato	12	13.8	civili	23	12.3	shock	5	13.5	educazione	12	5.3
controlli	21	13.6	*Forze dell'ordine / Policemen*			uccisi	15	13.3	confine	7	5.3
pacchi	5	12.1	polizia	54	25.6	abbraccio	13	12.4	democrazie	5	5.1
arma	13	11.4	militari	22	14.3	panico	7	11.8	riflessioni	7	4.7
redazione	11	11.3	guardie	6	10.3	vite	12	11.8	sanzioni	6	4.4
informativa	6	10.9	volontari	6	9.3	massacro	7	10.7	**GIUBILEO/JUBILEE**		
ucciso/a	20	10.6	militare	18	6.9	umani	25	10.2	Forma	occ.	Scarto stand.
gatti	10	10.0	truppe	5	5.5	incubo	8	10.1	basilica	39	188.6
violenza	32	10.0	soldati	6	4.2	coraggio	24	10.0	Immacolata	29	98.9
immagini	22	10.0	*Società civile/ Society*			parole	75	9.8	misericordia	62	91.0
esteri	15	9.8	tifosi	7	10.7	candele	5	9.5	pellegrini	22	55.1
chiusura	15	9.5	passanti	5	10.6	tragedia/e	16	8.9	presepe	16	44.3
caccia	13	9.3	popolo	33	9.6	lacrime	12	8.9	vaticani	7	33.6
porte	15	8.7	spettatori	6	6.9	ferita	7	8.6	guide	13	23.1
bandiera	10	8.6	testimone/i	15	6.5	sconvolto	6	8.2	preghiera	22	20.4
armato/i	11	8.4	persone	81	6.4	morta	51	8.0	tour	9	17.7
sangue	19	8.1	*Popoli e paesi/ Countries*			sorriso	11	7.4	Natale	13	7.5
emergenza/e	21	7.9	amici	24	5.4	preghiere	6	6.8	chiesa	25	6.2
corteo	6	7.2	cittadini	27	4.5	ansia	10	6.6	**VERBI/VERBS**		
news	6	7.1	commissario	5	4.1	amarezza	5	5.3	Forma	occ.	Scarto stand.
fiori	10	5.3	rifugiati	39	86.8	morte	35	5.2	bombardare	30	102.3
strategia	13	4.6	parigini	22	82.2	dramma	7	5.2	armare	18	56.6
Luoghi/ Places			francesi/e	145	53.1	assurdo	7	5.1	combattere	52	24.5
stadio/i	33	32.3	islamici/o/che	38	27.4	poveri	13	5.1	colpire	43	17.4
ambasciata	16	31.0	occidente	21	25.8	speranza	17	4.7	annientare	8	16.1
trattoria/ristorante	20	29.2	belga	16	23.6	cuore	31	4.3	reagire	22	15.1
piazza/e/strade	88	28.5	curdi	9	23.4	*L'odio e la rabbia/ anger*			uccidere	55	14.3
palco	19	28.2	musulmani	21	22.6	odio	45	31.6	commuovere	6	12.9
metro	24	23.9	saudita	6	18.6	vigliacchi	5	29.4	sconfiggere	12	12.0
moschee/a	12	17.6	profughi	11	16.3	bestie	13	17.8	piangere	30	11.3
concerti/o	27	15.1	turchi	7	14.6	sveglia	9	14.0	pregare	24	10.0
muri	14	13.9	immigrati	16	14.0	pazzi	7	13.6	bloccare	15	9.0
luoghi	25	11.2	romani	10	11.3	follia	10	11.8	morire	32	8.7
hotel	9	11.1	arabo	10	10.1	rabbia	14	11.2	sparare	9	8.0
mondo/pianeta	121	10.9	occidentali	11	7.8	ignoranza/ignoranti	16	10.7	mobilitare	7	7.9
rifugio	6	8.5	arabi	7	7.6	vendetta	9	7.6	fermare	23	7.8
teatro	21	6.6	clandestini	5	7.4	basta	39	7.5	finanziare	16	7.8
musei	8	5.8	russi/o	13	5.1	vergogna	9	7.3	distruggere	17	7.7
stazioni	5	4.8	oriente	5	4.1	condanna	8	4.2	emozionare	2	7.2

Table 3 – Categorised keyword (ordered by theme and standardized ratio)

The words that we classified as emotions were “dense”, and return in a very vivid way the feelings of those days (*brividi*, *commozione*, *sgomento, terrore*) alternated to strong feelings of rage and blame (*vigliacchi, vendetta, follia*). Finally there are some of those days’ most discussed themes, and particularly the matters related to immigration policies in Europe (*accoglienza, immigrazione, tolleranza, pace, fermezza*) or to the general destiny of democracy.

The availability of variables associated to tweets allows to extract the characteristic words for each of them (Tuzzi, 2003). Focusing the analysis exclusively on the tweets dedicated to Paris events (PARIS[5] corpus) it is interesting to observe the differences by number of followers

[5] The corpus is constituted by 65,000 tokens and 12,180 types. # symbols were considered as separators.

and by evolution in time. The analysis of characteristic words by number of followers highlights a polarization in the behavior of users with a low and high number of followers, respectively. In the former there is a higher occurrence of types that are more related to emotions (*io_non_ho_paura*, *pray*, *voi*/*io*), while the latter - that can be typically identified as professional profiles (newspaper, politicians, bloggers, etc.) – present a more frequent use of links and neutral words that recalls facts, places and subjects of politics (*ultimora*, *https*, *Renzi*, *Belgio* ecc.). The analysis of characteristic words per week well represent the chronological evolution of events. Obviously, the references to the attacks are mainly concentrated in the week from 13th to 19th November (week 2), as well as the most popular hashtags (*parisattacks, prayforparis*, etc.) or the references to news (*Bataclan, Saint Denis, morti, attacco* etc.).

Types	Token tot. (*10.000)	Token Week2 (*10.000)	p.value (*100.000)
Parigi/paris	256.0	309.9	0.000000
prayforparis	78.7	107.6	0.000000
parisattacks	57.2	76.8	0.000000
prayers4paris	19.2	26.7	0.000000
parigisottoattacco	14.3	20.0	0.000000
raqqa	11.4	15.7	0.000120
Dio	8.5	11.7	0.039595
francesi	12.6	16.6	0.078649
orianafallaci	8.1	11.0	0.098123
fare	17.9	22.7	0.114330
morti	17.0	21.7	0.120443
parigiinfiamme	7.6	10.4	0.242438
bataclan	18.8	23.6	0.359111
porteouverte	5.8	8.1	0.525933
ostaggi	5.8	8.1	0.525933
saintdenis	5.8	8.1	0.525933
parisshooting	5.8	8.1	0.525933
sarebbe	7.0	9.6	0.805559
siamo	20.1	24.6	1.846171
ho	14.1	17.8	2.447038
spiego	5.0	7.0	2.604672
prayforpeace	5.0	7.0	2.604672
attacco	8.5	11.3	2.951280
parisattack	4.9	6.8	3.586823
dolore	7.3	9.8	4.027251

Table 4 - Characteristic word by week from November 13th to 19th (ordered by increasing p-value)

Table 4 presents the characteristic words ordered by increasing p-value, with the indication of the normalized occurrences in the global text and in the specific part being analyzed. Characteristic words per day from week 2 also allow to describe the evolution of the topics of discussion: the words that describe what happened in Paris on the night of 13th leave space to participation and emotion, to the attention to *Valeria Solesin* – the only Italian victim – to the beginning of the French air strikes in *Raqqa*, to the raid in the *Molenbeek* district in Brussels. In the last two days of the week, characteristic words explain the blitz in Saint Denis. Week from 20th to 25th November (Week 3) is focused again on the lockdown in Brussels (*Bruxelles*, *foreign fighters*, *brusselslockdown*, *terrorismo*, *allerta*, *sicurezza*, *gatti*), while in the following weeks tweets in the corpus start to deviate from the terrorism theme.

6. The positive-negative dictionary to detect the mood of the text

With the explosive growth of social media on the web, individuals and organizations are increasingly using the content in these media to analyze behavior and opinions. Opinion mining or sentiment analysis analyzes people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics, and their attributes (Liu 2006). Starting from 2001, awareness of opportunities that sentiment analysis and opinion mining raise and subsequently there have been literally hundreds of papers published on the subject (Pang and Lee 2007). Twitter has already made a fertile ground for the analysis of moods and emotional attitudes of users. It is a "digital socioscope" (Mejova, Weber, and Macy, 2015) to study aspects of the society's functioning that are hard to capture in other ways. Different techniques and software are developed to face the difficult question of measuring the sentiment. The main difficulties regard the ambiguities of language and the intrinsic cultural factors in the text, reason why appropriate training activities are required (Bolasco, 2013). Aware of these limitations and ambiguities, in this section the tone of our corpus is evaluated using the positive/negative dictionary available in Taltac2: our task is not to classify the texts but to explore in broad terms the change of user's mood over the weeks and following the events. The use of this method has proved useful for this purpose.

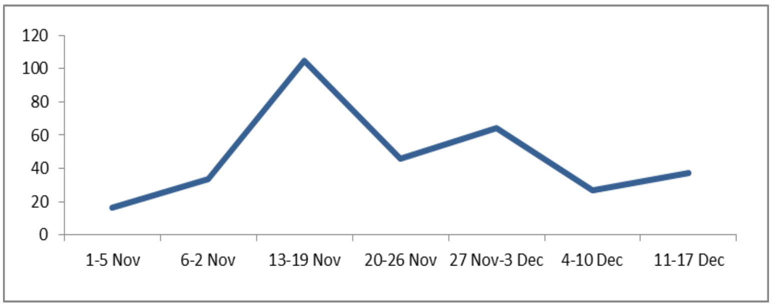

Figure 2 – Index of negativity of the PARIS corpus per week

The dictionary contains positive and negative adjectives present in General Inquirer (an instrument for content analysis developed in 1966 by Stone and Kelly) translated into Italian and integrated with lemmas from other sources (overall about 6,000 different forms). Applying the dictionary to a specific corpus it is possible counting the number of positive and negative forms. The ratio of negative to positive (tokens Neg/Pos*100) is useful to measure the level of negativity in a text. Applying the dictionary to an Italian frequency dictionary (Bolasco, della Ratta, 2004) a threshold point was defined: texts with negative index higher than 40% can be considered with a negative connotation. In the global set of the 240,000 tweets the general level of the index is 38.5%, a level substantially aligned with what detected using the dictionary of frequency. This values raises to 78.9% in the PARILEO corpus, with a peak of 92.1% in the tweets related to terrorism and a value of 25.5% in the Jubilee ones. Along the seven analyzed weeks, a peak of 105% is reached during week 2 (Figure 2).

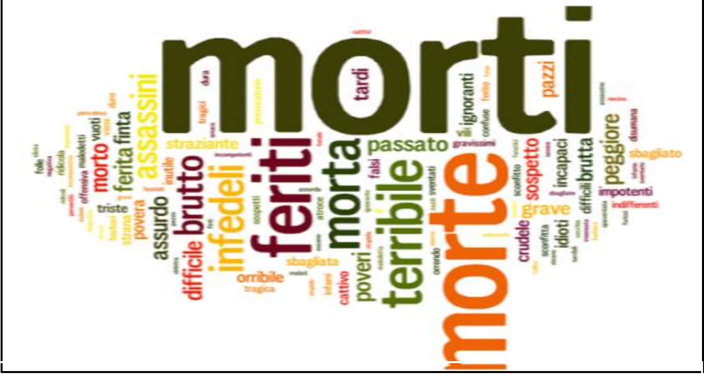

Figure 3 – Tagcloud of negative adjectives in PARIS

The level of negativity in the Paris-only tweets is 95% in those authors that are placed in the first quartile in terms of followers, the one corresponding to non professional authors use Twitter to manifest their feelings about the events, while reaching only 78% among the most followed users, that probably use a more neutral language, as explained above in the analysis of characteristic words. Besides computing the index, it is also possible to identify the most frequent negative adjectives. In the Paris-only corpus the most frequent form is *morti* (dead – 122 occurrences) besides *feriti, terribile, infedeli, brutto, assassino, assurdo, pazzi e crudeli*. Figure 3 depicts the tag cloud of the negative adjectives that well represents the shock of the Twitter community.

7. Conclusions

This paper describes an example of integration between big data and text analysis techniques that can provide inspiration for future research. The possibility of quickly extracting a (well-defined) selection of tweets and computing variables associated to them is a promising starting point for research.

Analysis techniques that do not require language resources, such as repeated segments detection and extraction of characteristic words, can be applied more straightforwardly as they do not pose limitations on the language to select. Among the resources that are more specific to text analysis the use of the positive-negative dictionary included in Taltac2 was particularly useful for classifying the mood of the text, especially in the presence of events of high media coverage and emotional impact.

The recognition of hashtags can be considered an added value also in the analysis of content and user profiles, for example analyzing the number of # in a text or the way they are used (attached to words in the text or as separated tags). The same can be said about the use of other special characters, such as @ for targeting other users or links to images and web pages.

The results of the analysis are unavoidably influenced by the use of a geographical filter on the source. However, the concentration of major media and political influencers makes the selected sample highly representative for what concerns the different types of users in an explorative analysis. Nevertheless, future work could explore the effect of different or wider geographical selections of publication place.

References

Alexa (2016). Twitter site overview, available at http://www.alexa.com/siteinfo/twitter.com

Anderson C. (2006). *The long tail: Why the future of business is selling less of more*. NY: Hyperion.

Bolasco S. (2005). Statistica testuale e Text Mining: alcuni paradigmi applicativi". In *Quaderni di Statistic*a. Liguori, 7.

Bolasco S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Roma, Carocci.

Bolasco S. and della Ratta-Rinaldi F. (2004). Experiments on semantic categorisation of texts: analysis of positive and negative dimension, in Purnelle G., Fairon C. e Dister A. (eds). *Le poids des mots, Actes des 7es Journées internationales d'Analyse Statistique des Données Textuelles*. UCL, Presses Universitaires de Louvain.

Bruns A. and Burgess J. (2014). Crisis Communication in Natural Disasters: The Queensland Floods and Christchurch Earthquake. In Weller K., Bruns A., Burgess J., Mahrt M., and C. Puschmann C. (Eds.) *Twitter and Society*. New York. NY: Peter Lang

Liu B. (2006). *Web data mining. Exploring hyperlinks, contents, and usage data*. Springer.

Mejova Y., Weber I. and Macy M. (2007). *Twitter: A Digital Socioscope*. Cambridge University Press.

Pang B. and Lee. L. (2007). Opinion Mining and Sentiment Analysis. Foundations and Trends. In Information Retrieval 2(1-2):1-135.

Ratinaud P. (2014). Visualisation chronologique des analyses ALCESTE: application à Twitter avec l'exemple du hashtag #mariagepourtous, in E. Née, J.M. Daube, M. Valette, S. Fleury, (eds). *Actes des 12es Journées internationales d'Analyse statistique des Données Textuelles*. Paris Sorbonne Nouvelle – Inalco.

Techamerica (2012). Demystifying Big Data: A Practical Guide to Transforming the Business of Government. TechAmerica Foundation's Federal Big Data Commission, available at http://mim.umd.edu/2012/11/demystifying-big-data-a-practical-guide-to-transforming-the-business-of-government/

UNECE (2012). Classification of Types of Big Data, available at http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data

UNECE (2014). Big Data in Official Statistics, available at http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics

Tuzzi A. (2003). *L'analisi del contenuto. Introduzione ai metodi e alle tecniche di ricerca*. Roma. Carocci.

Vaccari C. (2014). Big Data and Official Statistics. PhD Thesis, School of Science and Technologies - University of Camerino, available at https://www.academia.edu/7571682/PhD_Thesis_on_Big_Data_in_Official_Statistics_