UNIVERSITÀ DI PISA

Scuola di Dottorato in Ingegneria “Leonardo da Vinci”

Corso di Dottorato di Ricerca in

INGEGNERIA DELL’ INFORMAZIONE

Tesi di Dottorato di Ricerca

Social Media Monitoring and Analysis: Multi-domain Perspectives

Davide Gazzè

Anno 2015

UNIVERSITÀ DI PISA

Scuola di Dottorato in Ingegneria “Leonardo da Vinci”

Corso di Dottorato di Ricerca in
INGEGNERIA DELL’INFORMAZIONE

Tesi di Dottorato di Ricerca

Social Media Monitoring and Analysis: Multi-domain Perspective

Autore:

Davide Gazzè ______________________

Relatore:

Ing. Alessio Bechini ______________________

Dott. Maurizio Tesconi ______________________

Dott. Andrea Marchetti ______________________

Anno 2015
SSD ING-INF/05

Sommario

I Social Media sono potenti strumenti di comunicazione entrati nella vita di tutti i giorni. Essi permettono di ridurre le barriere geografiche e temporali tra le persone, di migliorare la propria immagine e di condividere facilmente informazioni. Per questo motivo, sempre più professionisti e aziende hanno un account su almeno un Social Media. Data la diffusione di queste piattaforme, ogni giorno su di esse viene generata un'enorme quantità di dati. Le relative informazioni possono giocare un ruolo fondamentale in svariati processi di decision-making e per questo motivo le tematiche di monitoraggio e analisi di Social Media stanno assumendo un'importanza sempre crescente. Partendo dalle metodologie di raccolta dati da Social Media, si tratteremo le problematiche connesse alla memorizzazione di dati semi-strutturati e ai possibili gap d'informazione legate alle politiche di privacy. Tali argomenti saranno analizzati sia per dominio di studio (Online Reputation, Social Media Intelligence e Opinion Mining) sia per Social Media (Facebook, Twitter, etc.). Successivamente sarà presentata un'architettura teorica per la raccolta dati da Social Media, progettata sulla base delle problematiche e dei requisiti per il monitoraggio. Saranno inoltre presentate tre versioni semplificate della suddetta architettura su tre domini di studio: Online Reputation, Social Media Intelligence e Opinion Mining per dominio turistico. Infine, la tesi tratterà delle tematiche di analisi nei suddetti domini. Inizialmente verrà presentato *SocialTrends*, una web application che permette di monitorare personaggi su Facebook, Twitter e YouTube. Per il secondo ambito di studio, verrà presentata una metodologia di analisi delle interazioni tra utenti eseguite negli spazi pubblici (public-by-design) di Facebook. Infine, verrà presentata *Tour-pedia*, una web application che mostra su una mappa l'opinione degli utenti su luoghi appartenenti a varie categorie (accommodation, restaurant, point of interest e attraction) di varie città del mondo.

Abstract

Social Media Platforms, such as Facebook or Twitter, are part of everyday life as powerful communication tools. They let users communicate anywhere-anytime, improve their own public image and readily share information. For this reason, a growing number of individuals such as professionals as well as companies have opened an account in one or more Social Media platforms. Due to the widespread use and growing numbers of users, a huge amount of data is generated every day. This information may play a crucial role in various decision-making processes. In this setting, research topics connected to monitoring and analysis of Social Media data are becoming increasingly important. The present work stems from data collection methodologies from different Social Media sources. It introduces the problems involved in storing semi-structured data, and in possible information gaps due to privacy policies. These facets are described according to the application domain as well as the Social Media platform. Subsequently, a theoretical generic architecture for handling data from Social Media sources is presented. We present three simplified versions of this architecture in three different domains: Online Reputation, Social Media Intelligence, and Opinion Mining in tourism. In the last part of the work, we introduce Social Media Analysis in these three domains. For the first, we present the project SocialTrends, a web application able to monitor “public” people on Facebook, Twitter, and YouTube. In the second, we introduce an innovative approach for measuring the interactions between users in public spaces such as Facebook (public-by-design). Finally, we present Tour-pedia, a web application that displays a sentiment map of tourist locations in several cities according to different categories (accommodation, restaurants, points of interest and attractions).

This thesis is dedicated to my past, present and future family

“The best thing for being sad," replied Merlin, beginning to puff and blow, "is to learn something. That’s the only thing that never fails. You may grow old and trembling in your anatomies, you may lie awake at night listening to the disorder of your veins, you may miss your only love, you may see the world about you devastated by evil lunatics, or know your honour trampled in the sewers of baser minds. There is only one thing for it then — to learn. Learn why the world wags and what wags it. That is the only thing which the mind can never exhaust, never alienate, never be tortured by, never fear or distrust, and never dream of regretting. Learning is the only thing for you. Look what a lot of things there are to learn.”

T.H. White, The Once and Future King

Contents

List of Figures

List of Tables

1

Introduction

1.1 Rationale

According to the definition of Kaplan and Haenlein [64], Social Media platforms are “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of “user-generated content”.

Nowadays, Social Media platforms are very important instruments for enabling communication between users, and for sharing opinions, photos, videos, and documents. After observing this, many companies have developed a strong interest in Social Media data, especially personal data, for the related marketing implications. Also in the scientific community, a growing number of researchers are currently using these data for investigations in various areas, from sociology to ICT.

In recent years, the rapid growth of Social Media platforms has attracted interests across different domains. Typically, the focus is on how to exploit social media data for specific purposes: just to mention a few, Online Reputation, prevention of organized crime, political analysis, better tackling of social crises, support of e-Health services, effective marketing analysis, provision of early warnings/alerts, and so on.

Investigation of Social Media data is aimed at obtaining knowledge, from User Generated Content (UGC), on the “wisdom of crowds” (term coined in a 2012 report [79]). This knowledge is valuable for various decision making processes, and it is the main rationale for monitoring and analysing Social Media data. In general, Social Media monitoring and analysis can be carried out by following some basic steps: data capture, data analysis, and results reporting.

Any rigorous investigation in this field requires both thorough knowledge of the application domain and a proper technical support. From a technical point of view, the main issues involve the selected methodology for data capturing, the constraints imposed by specific policies adopted in the Social Media platform, and the solutions for efficient data storage and access. The success of studies of Social Media data usually relies on the knowledge of users' behaviour. Moreover, a preliminary analysis is fundamental in order to properly choose the reference Social Media. For example, eBay is not a good data source for studying the reputation of political candidates, but it may be perfect for marketing goals.

In recent scientific literature, typically papers do not present any general approach to Social Media data capture, focusing instead on data analysis problems. Furthermore, the data capture phase is often carried out in a semi-automatic way, with manual intervention, or by means of very specific applications.

According to the observations reported so far, it is possible to define the following key points for a successful analysis using Social Media data:

- Knowledge of the application domain;
- Choice of one/multiple adequate Social Media platform(s);
- Use of a proper system (developed according to a well-defined architecture) to collect, manage, and analyse social data.

1.2 Contribution

The successful exploitation of Social Media data is possible only after overcoming the technical issues involved in data capture, storing, management, and analysis. To the best of our knowledge, in the scientific literature so far investigations have been mostly biased towards specific application domains, thus lacking a uniform approach to shaping a general system architecture for capturing Social Media data. Although in principle the main steps of a study are always the same, in practice each application domain underpins specific requirements and goals. For example, for Brand Online Reputation, it is crucial to monitor how the number of fans changes over time. Instead, for the identification of criminals on Social Media, information about people, relationships, and text content plays a central role.

The main goal of this study is to offer general contributions to the field of Social Media monitoring and analysis by answering the following basic questions:

i) What specific capabilities on the Social Media platform are required to enable data capture by a structured tool?
ii) Given such enabling means, how can the data capture tool deal with the related limitations?

iii) What is the impact of flexibility and robustness requirements in data capture on the tool architecture? What are the required components?
iv) Is it possible to derive a simplified, lean version of the proposed general architecture to implement efficient tools to address specific domains?
v) What kind of analyses can be carried out on the captured data?

To answer such questions, a bottom-up approach is pursued, and problems related to massive data capture are tackled starting from the study of Social Media data in different domains. Techniques for capturing data from Social Media are first described, subsequently discussing the advantages and disadvantages of each methodology. These considerations represent the answers to questions i) and ii).

Question iii) concerns shaping a generic theoretical architecture for data capture from a Social Media platform. Since building up a full-featured implementation of the generic architecture might be very demanding, we studied different approaches to create simplified versions from it. We use the simplified architectures in the following domains: Online Reputation, Social Media Intelligence, and Opinion Mining in tourism. Furthermore, for each domain, we detail the advantages and disadvantages of the simplified implementation. This part of the work provides answers to question iv).

Regarding the last question, methods fo analysis for the aforementioned domains are presented. In the specific field of Online Reputation, we present *Social-Trends*[1], a web application that collects, elaborates, and visualizes social media data from Facebook, Twitter, and YouTube. In the SOCMINT field, we describe the European Project CAPER, showing analysis solutions for investigating Facebook interactions between people in terms of strength, frequency, and duration. For the last domain, we present *Tour-pedia*, a web application developed in the context of the European Project OPENER. Tour-pedia exploits the OPENER's linguistic pipeline to extract the sentiment in reviews about accommodations, attractions, points of interest, and restaurants.

Summing up, the main contributions of this work can be enumerated as follows:

- Discussion of problems related to capture, storage, management, and analysis of Social Media data;
- Design of a general architecture for capturing and managing Social Media data;
- Implementation of simplified versions of the general architecture in different domains (Online Reputation, Social Media Intelligence, and Opinion Mining in tourism);

[1] http://www.social-trends.it

- Presentation of different types of analysis for each of the above domain.

1.3 Thesis overview

Chapter 2 first describes the state of the art of opportunities offered by Social Media platforms. Next, we introduce methodologies for capturing data from Social Media, namely crawling, scraping, and Web API. For each technique, we underline both the advantages and disadvantages. Regarding the last (the most frequently used) technique, two kinds of Web API are addressed: REST and Streaming. Moreover, we present a complete overview of the methodologies for capturing data from Facebook and Twitter. In the last part of this chapter, a literature review of Social Media Analytics introduces this strategic research field.

Chapter 3 is devoted to the design of a generic modular architecture for data capture from different Social Media platforms. Step by step, we present a simplified implementation of such an architecture, able to collect data in different application areas. In particular, we present three implementations for the domains of Online Reputation, Social Media Intelligence, and Opinion Mining in tourism. These applications are able to collect data from different Social Media platforms (Facebook, Twitter, YouTube, etc.), thus assessing the flexibility of the proposed architecture.

Chapter 4 introduces the problem of how to analyse data, after their collection. In particular, for Online Reputation, we present *Social Trends*, a web application able to compare brands and public people over three different platforms (Facebook, Twitter, and YouTube). For Social Media intelligence, we present an innovative approach for the exploitation of the public-by-design space of Facebook, in order to help Law Enforcement Agencies (LEAs) spot criminals. As a final case study, we present *Tour-pedia*, which features an interactive map able to display the sentiment of reviews on touristic places (categorized as accommodations, attractions, points of interest, and restaurants). In this case, reviews are captured from four different Social Media (Facebook, Foursquare, Google Places, and Booking.com).

In Chapter 5, we discuss the main findings and draw proper conclusions.

2

State of the Art

2.1 Data Revolution

In today’s society, *Data* are “at the centre of the future knowledge economy and society” [1]. In accordance with the ISO/IEC 2382-1, data are “a reinterpretable representation of information in a formalized manner, suitable for communication, interpretation or processing”. People or machines can produce different types of data (geospatial information, weather data, reviews, etc.).

There are three big classes of data: Structured, Unstructured, and Semi-structured. For the first class, we refer to all kinds of data that follow a rigid schema. This kind of data is self-describing; some examples of Structured Data are a SQL Database or an XML file (with an XML-schema).

With the term *Unstructured data*, we refer to typical texts where there is no fixed schema, and the data can contain substructures like dates or currencies. For the vast variety of this kind of data, Unstructured data are the most difficult to analyse.

The term Semi-structured data refers to a kind of data with a model where same parts are not present in all instances. This typology has recently emerged [24] as an important topic for different motivations:

- Support for hierarchical or nested data;
- Representation of relationships of two or more instance.

Today, the term Big Data is very popular. In [1], the authors define it as a “huge quantity of different types of data produced with high velocity from a high number of various types of sources. Handling today’s highly variable and real-time datasets requires new tools and methods, such as powerful processors, software, and algorithms‘’.

In [61] and [9], the focus point is the inability "to extract insight from an immense volume, variety, and velocity of data, in context, beyond what was previously possible" because of the inefficiency of traditional tools [61].

Three characteristics (known also as V3) define Big Data: volume, variety, and velocity (as shown in Fig. 2.1). The volume of data is growing. In the year 2000, according to a book [61], all the data in the world ware estimated to be 800,000 petabytes (PB). In 2020, the prediction is 35 zettabytes (ZB). As support for this hypothesis, there are Social Network platforms. For example, Facebook generate over 500 Terabytes of data every day[1].

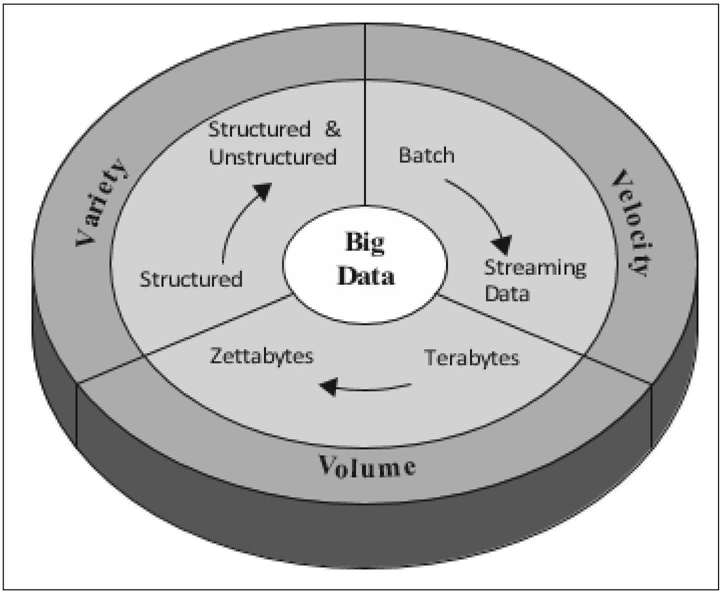


Figure 2.1: Properties of Big Data (source: IBM [61])

Another challenge is the data's variety. In fact, different sources, such as companies, sensors, smart devices, and Social Media platforms produce data with different formats. Moreover, most of these data are in an unstructured or semi-structured [61] form.

The last characteristic is the velocity of the creation of data. In the paper [61], authors suggested changing the idea of velocity from growing rates associated

[1] https://gigaom.com/2012/08/22/facebook-is-collecting-your-data-500-terabytes-a-day/

with data repositories to the “speed at which the data are flowing”. The importance of this research field is the predictive and descriptive power of data.

The term *Big Digital Divide* refers to the gap between whom generates content and who collects and stores it [9]. Usually the first actors are the end users, while the second actors are the companies. Big Data are important, in public affairs as well, since “our democracy relies on the quality of data in the public domain, and the public’s trust in it”.[2]

Closely connected to the Big Data revolution, there is the issue of the management of personal and sensitive information that companies tend to store. The term *personal data revolution* refers to the management of personal information, especially if sensitive. In an interview with “The Guardian”, Tim Berners-Lee says[3], “My computer has a great understanding of my state of fitness, of the things I’m eating, of the places I’m at. My phone understands from being in my pocket how much exercise I’ve been getting and how many stairs I’ve been walking up and so on.”

2.2 Social Media Platforms

Social Media platforms increasingly pervade everyday life: they represent virtual places where people share contents (text, video, and photos) and sensitive information. In [64], Kaplan and Haenlein (2010) define a Social Media platform as a web or mobile application that allows user to create, access, and share user-generated content (UGC).

Examples of Social Media are services like Social Networks (e.g. Facebook, Twitter, and YouTube), Really Simple Syndication (RSS) Feeds, blogs, and wikis [17].

The authors classify the Social Media platforms into six groups.

Blogs: the earliest form of Social Media, personal web-sites;

Collaborative projects: platforms that allow many users to create and modify the same content (e.g. Wikipedia);

Content communities: platforms that allow sharing of media content (e.g., Book-Crossing, YouTube, Slideshare, Flickr);

Social network: applications that allow users to connect with others (e.g. Facebook);

[2] http://blogs.lse.ac.uk/impactofsocialsciences/2015/01/22/data-manifesto-democratic-debate-royal-statistical-society/

[3] http://www.theguardian.com/technology/2012/apr/18/tim-berners-lee-google-facebook

Virtual worlds: 3-dimensional environments where users can interact and play (e.g. World of Warcraft);

Virtual social worlds: applications where users have virtual life similar to their real life (e.g. Second Life).

Figure 2.2[4] shows the vast landscape of most popular Social Media.

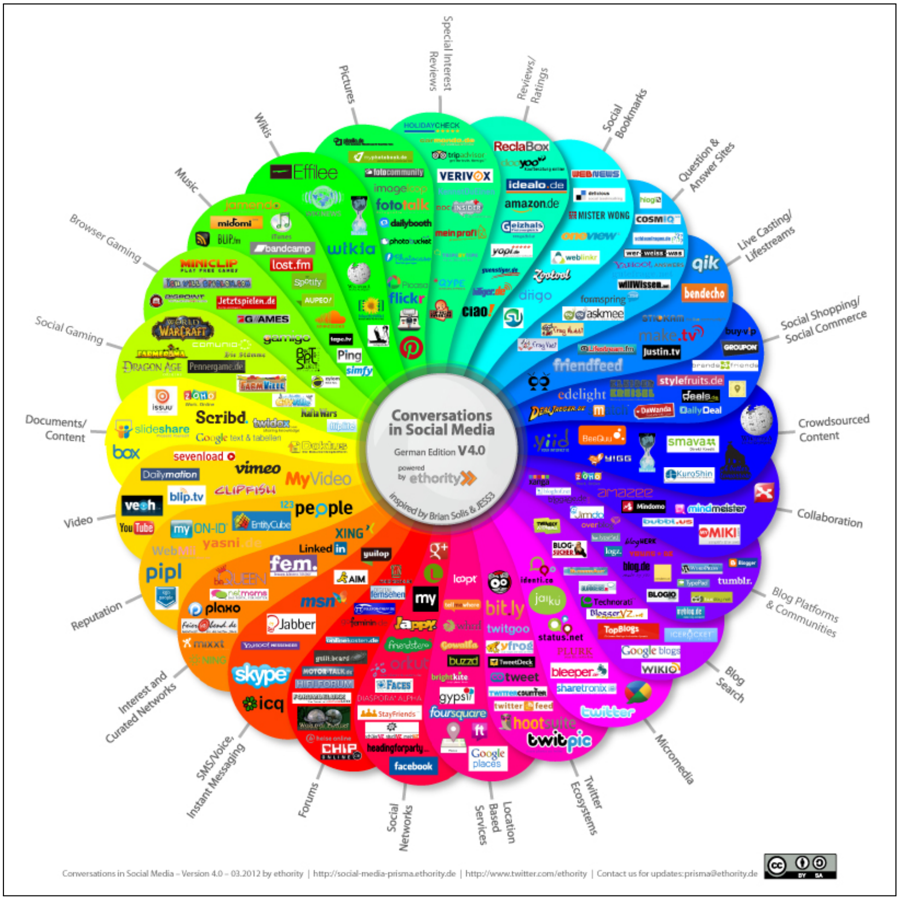

Conversations in Social Media
German Edition V4.0
powered by ethority
inspired by Brian Solis & JESS3
Special Interest Reviews
Reviews/ Ratings
Social Bookmarks
Question & Answer Sites
Live Casting/ Lifestreams
Social Shopping/ Social Commerce
Crowdsourced Content
Collaboration
Blog Platforms & Communities
Blog Search
Micromedia
Twitter Ecosystems
Location Based Services
Social Networks
Forums
SMS/Voice, Instant Messaging
Interest and Curated Networks
Reputation
Video
Documents/ Content
Social Gaming
Browser Gaming
Music
Wikis
Pictures
Conversations in Social Media – Version 4.0 – 03.2012 by ethority | http://social-media-prisma.ethority.de | http://www.twitter.com/ethority | Contact us for updates: prisma@ethority.de

Figure 2.2: Social Media Landscape

[4] http://usabilitynotes.typepad.com/usabilitynotes/2012/06/many-social-media-services-ethnoritys-lovely-taxonomy.html

Table 2.1, taken from [64], shows the richness of Social Media platforms in accordance with the personal details that users tend to insert.

		Media Richness		
		Low	**Medium**	**High**
Self Presentation	**High**	Blogs	Social Network	Virtual Social World
	Low	Collaborative project	Content communities	Virtual game worlds

Table 2.1: Classification of Social Media by media richness and self-presentation

Over the last few years, Social Media Platforms have been increasing continuously. Figure 2.3[5] shows the Social Media Timeline from 1994 to 2011.

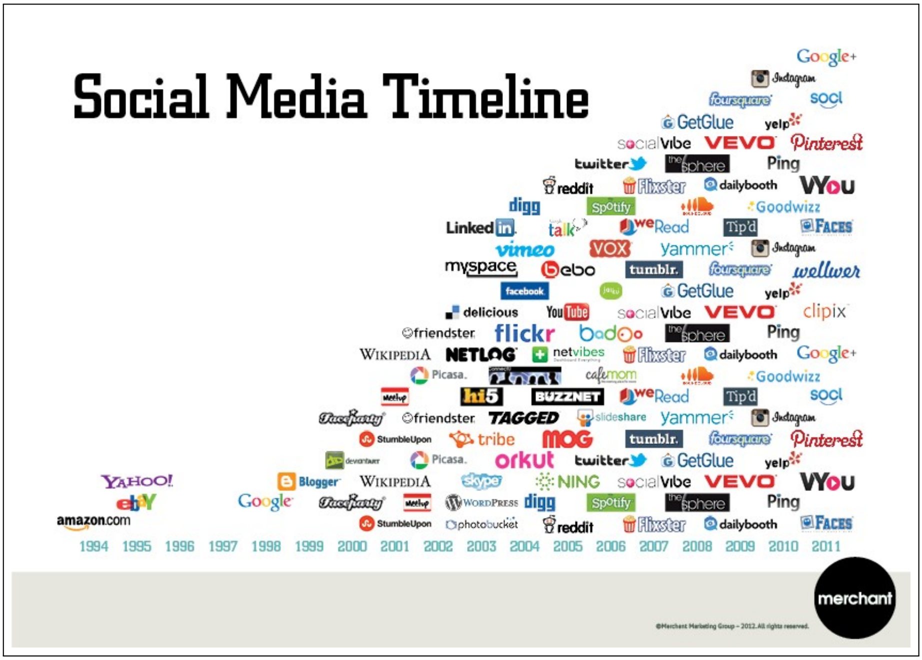


Figure 2.3: Social Media Timeline

[5] http://ltpublicrelations.com/a-decade-of-social-media/

A short specification about the role of Social Networks is necessary. In accordance with [23], a social network site is a web-based service that allows people to:

- Construct a public or semi-public profile;
- Create a list of users with whom they share a connection;
- View their list of connections.

An example of popular Social Media is Facebook. This Social Network had approximately 865 millions daily active users in September 2014 and 1.35 billion/month[6].

For all reasons, the Social Media platforms are a rich source of content for different investigations [90].

To understand the typology of data in Social Media, some considerations about the user's behaviours are necessary. A user tends to create connections with others. These connections have different names, some popular terms are *Friend*, *Follower*, *Fan*, *Subscriber*, and others.

As described in [23] and [93], a relationship is bi-directional or symmetric, like the *friendship* on Facebook or the *colleague* on LinkedIn when the relation from user A to user B, A->B, implicates that B->A. Other times, the relationship is unidirectional or asymmetric. In this case, the presence of A->B does not implicate B->A (but it can exist). A relationship is multimode if the connection is between actors of different types (Corporations employ People, People are fans of a Band, etc). More details of typologies of relationships are available at [93] and [92].

There are various research challenges connected to Social Media. The survey in [17] defines the following:

Data Capture: the step of capturing data from a platform;

Data Cleansing: the raw data are not ready for other computations. For this, the raw data must be cleaned and/or transformed into a normalised format;

Holistic Data Sources: the task of bring together different data sources;

Data Protection: this research area regards the protection of a Big Data resource. The access to data needs to be granted by IP, and with different levels of access;

Data Analytics: all the kinds of analysis such as Opinion Mining, Social Network Analysis, Group Detection, Group Evolution, etc;

Analytics Dashboards: many Social Media platforms are accessible via APIs. Non-programming interfaces can provide deep access to raw data for non-developers;

[6] http://newsroom.fb.com/company-info/

Data Visualization: representation of data with the goal of communicating information clearly through graphical means.

In this thesis, we will approach the challenges of *Data Capture* and *Data Analytics*. Moreover, we will give some details on the challenges of *Data Cleansing* and *Holistic Data Sources*.

2.3 Methodology of capture of Social Media Data

As described in [9], there are two main categories of analysers: those who have the data and those who not. For the analyser in the second category, a capture phase is necessary to gather these data. Historically, the action of acquiring data from Web is not novel, because many datasets are collections of objects connected by links. The original field of study is Link Mining. This field includes object ranking, group detection, link prediction and subgraph discovery [45].

By the term *capture*, we mean all the techniques for retrieving data from Social Media Platforms. The choice of the correct techniques or tools for capturing data is critical. The National Archives and Records Administration (NARA)[7] claims that the use of Social Media Platforms in federal agencies is increasing. For this reason, the authors present some best practice and tools for successfully capturing data. Unfortunately, each tool is able to capture only one type of data (e.g. a video from YouTube).

The authors of [17] propose a classification based on the availability of information:

1. Freely available databases: repositories that can be downloaded free (e.g. Dump of Wikipedia);
2. Data access via tools: software that provides controlled access to Social Media data. These tools can be free or commercial (GNIP or DataSift);
3. Data access via Web APIs: Social Media Platforms provide programmable REST to provide access to Social Media data (e.g., Twitter, Facebook, and YouTube).

The above classification is only theoretical because each Social Media provides a platform with proper rules. Obviously, the first class is the most comfortable for end users, since all the necessary data are available in a simple format. Unfortunately, this case is rare and this way is impracticable if there are personal data. There are different formats of Social Media data [17]:

[7] http://www.archives.gov/records-mgmt/resources/socialmediacapture.pdf

SQL Dump: The SQL export of a Relational Table or Database;
HTML: HyperText Markup Language (HTML), the markup language for web pages;
XML: Extensible Markup Language (XML), the markup language for structuring textual data;
JSON: JavaScript Object Notation (JSON) is a text-based standard. It is designed for human-readable data interchange;
CSV: Comma-separated values (CSV) contains values as series of lines organized such that each row starts a new line and each column value is separated from the next by a comma or semi-comma.

Nowadays, the JSON format is the most frequently used because it is lightweight and fully supported by every modern programming language. A little specification is necessary. Until now, we have referred to data as all kinds of information from Social Media. We can divide Social Media data into two categories:

Data: Designs all the information that users insert in Social Media;
Metadata: Data about data, for example the time when the user posts a content, the ID of a user, etc.

There are different kinds of applications for capturing data. Historically, we mentioned the Crawler and the Scraper. A crawler is an application that systematically browses the Web with the goal of Web Indexing. These programs use HTML's hyperlinks to visit the complete graph of the Web. In addition, a scraper uses several techniques, such as DOM parsing, Regular Expression or xPath query, for extracting information from a website. However, some platforms disallow the use of these applications due to the presence of sensitive information.

Due to the complexity of modern web technologies (e.g. Javascript and Ajax), it is impossible to capture data using a crawler or a scraper. Moreover, the Social Media business model needs to exploit the user's data. Regarding the second point, most Social Media platforms provide a set of free routines for capturing data.

There routines are the Web API (Application programming interface). In contrast to applications like crawlers and scrapers, which can go across public pages, the Web APIs also provide access to private information (using legacy authentication mechanisms).

Web APIs are important in research for both quantitative and qualitative reasons. There are quantitative implications because Social Media data are Big Data [21]. The qualitative reasons are the opportunities for studying the communication patterns from content [21].

Unfortunately, sometimes Web APIs are commercial tools. In fact, because of the importance of data for marketing implications, most Social Media platforms sell their data through companies (called resellers). For example, GNIP, now Twitter's partner, is a Social Media API aggregation and it provides commercial access to different Social Media platforms.

2.3.1 Issues related to management of Social Media data

In the previous paragraphs, we explained the basic techniques for capturing data from Social Media. However, the creation of a module for capturing data is not a trivial task because the software must overcome various obstacles.

First, let us discuss the variety of information in Social Media. Social Media data belong to different typologies of data such as contacts, personal details and activities (written opinions, shared links, etc). Moreover, it can exist in different formats (text, image, sound, and video). For these reasons, proper support of storage is necessary.

Accessibility to data is another important problem, since not all data exist in the public domain or are easily accessible. In the worst scenario, the information is private (black hole of information). In addition, all Social Media have two kinds of restrictions: privacy and technical. Both of them protect the user's privacy and the Social Media's business model.

Privacy restrictions are the mechanisms offered to users to restrict access to information. Instead, technical restrictions regard all the mechanisms for controlling the information flow from the Social Media to other applications. Examples of these controls are restrictions based on IP address, rate limit and banning politics. Simultaneously, the action of capturing data from Social Media has legal and ethical issues.

In the next two paragraphs, we will detail two fundamental problems:

- Data Storage;
- Legal and ethical issues of Social Media data.

2.3.2 Data Storage

In previous paragraphs, we explain the importance of Social Media data and the method for capturing it. Obviously, the use of proper storage technology is very important. From our point of view and in accordance with [61], the characteristics of Social Media data are: Volume, Velocity, and Variety.

In order to choose the storage technologies for our task, we must consider the two major classes of Social Media data. The first class is semi-structured data in JSON or XML format. The second class is unstructured contents like documents, images, and videos. The best supports for our purpose are:

- File system;
- Database Management System (DBMS).

Obviously, in the file system, we can store both types of data, but this choice has led to different problems for managing them. For this reason, we study the vast landscape of Database Management Systems. A database is a collection of data managed by a database management system (DBMS). It allows users to insert, to update, and to search data [94]. Historically, the first database was based on the relational data model that was proposed in 1969 by Edgar F. Codd.

In 1999, Jensen and Snodgrass [63] introduced the concept of temporal database management systems. In temporal databases, data are represented by different snapshots, one for every interval of time [10]. Furthermore, the temporal databases introduce two types of attributes: validity time and transaction time. The first is the period when a snapshot is valid with respect to the real world, while the second attribute represents the period when a fact is stored in the database.

Another family is the spatial DBMS. The main feature of this kind of storage is the ability to manipulate objects that represent physical points in space [51]. The spatial databases are important for storing geographical or geometrical data.

In the 2000s, a new generation of DBMSs appeared. These are called NoSQL DBMSs [75]. The common features of a NoSQL DBMSs are non-adherence to the relational data model, horizontal scalability, and easy replication support.[8] Examples of NoSQL DBMSs are Membase, Couchbase, CouchDB, MongoDB, Neo4j, and InfiniteGraph. There are different types of NoSQL DBMS:

Key-value: the simplest NoSQL databases. Every item is a pair consisting of an attribute name (called key) and a value. Examples of key-value DBMS are Redis, Riak, and Voldemort;

Column based: DMBSs like Cassandra and HBase that are optimized for large datasets. Each record spans one or more columns containing the information;

Document databases: DMBSs where each element is a complex structure called document. Documents can contain different key-value pairs, key-array pairs or nested documents. MongoDB belongs to this category;

Graph: DMBSs where each element is a Graph. These systems, like Neo4J or HyperGraphDB, store information about networks.

[8] http://nosql-database.org/

2.3.3 Legal and ethical issues of Social Media data

The Social Media Platforms have introduced different legal and ethical aspects that are research challenges. I particular, the central point is the use of data that is private or semi-public. In fact, content in Social Media, such as Twitter and Facebook, may include personal details (name, date of birth, occupation, etc.) and "it is unclear to what extent the personal information is of a sensitive nature" [72].

Management of sensible data is a worldwide problem and it involves various actors: the user, the Social Media platform and the country's privacy laws. The first actor, the user, has considerable power because he can choose whether to publish something or not. In fact, different studies suggest that users have increasingly adopted more restrictive privacy settings for their personal data. The paper [37] shows that between 2010 and 2011, 1.4 million Facebook users in New York had increased the privacy in their profiles from 12

The Social Media platform has its own *terms of use* that the user must accept during the registration phase. In some platforms, like Twitter, the content is public by default. Other times it can be public or private, this is Facebook's case [91]. The general rule is that if data are available without any authentication action, the data are public. However, this consideration, from a legal point of view, is contentious.

The Social Media business model is the exploitation of a user's activity. In fact, Social Media providers do not share all UGC [74]. As described before, the final actor is the country and the policy law of each country. The paper [43] shows how the concept of personal data changes across different parts of world. Moreover, the studies [42], [18], and [19] studied the concept of sensitive data (sexual orientation, religion, and politics) in relation to the country.

Even though the legal issue is not the goal of this thesis, we introduce some aspects for the management of personal data in the European Union (EU). The problems of management of personal data in the EU are analysed in Directive 95/46/EC of the European Parliament on the protection of individuals with regard to the processing of personal data and on the free movement of such data (1995)[9]. Article 2 claims "'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity".

[9] http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:31995L0046

2.4 Social Media capture via Web APIs

Social Media platforms provide a set of application programming interfaces (APIs), regarding the HTTP protocol, available to third parties. These Web APIs allow third-party companies to implement different services integrated with Social Media [72]. There is a short guide to free Web APIs [98]. In a paper [22], the authors claim the importance for analysis purpose of the enriched set of Social Media data and of the APIs as methodological tool.

The type of data available from an API depends on the Social Media platform itself. Each Web API has zero or more input parameters and it returns data in a certain format (typically JSON). Web APIs allow applications to collect digital footprints and usage patterns (communication, connectivity, etc.) of a profile on a Social Media. Every Web API receives a request from a URL, called endpoint, for making the call. In order to create a classification of Web APIs, we must consider four different aspects: data format, input parameters, response, and type of response.

Regarding the first aspect, most Web APIs use two main formats: XML and JSON. However, nowadays the most frequently format is JSON. This format has various advantages with respect to XML. Although, every modern programming language natively supports both of them, the JSON format is simpler than XML. In addition to this, JSON is less verbose than XML and this reduces bandwidth consumption of Social Media companies. The second aspect is the input parameters. Most Web APIs haves both mandatory and optional parameters. The third aspect is the response parameter. Web API returns a list of objects with a data schema decided by Social Media Platform. The last aspect is the most interesting because it depends on how the software, that captures the data must work. There are different method of Web API interrogation. In fact, sometimes the response has an end while other times, it is a continuous flow of data.

There are three different kinds of response:

REST: most Web APIs belong to this category. An application makes a call to an endpoint and if everything go well, will receive a response;

Streaming: in this case, there is a persistent HTTP connection to the endpoint (the stream). The application must maintain the Streaming connection and it must process the received data rapidly (otherwise the stream will be closed by the Social Media);

Real-time: the inverse version of REST, an application registers a script (called callback) to an endpoint. For each new information, the Social Media platform will notify the callback[10].

The Figure 17[11] shows the difference between Rest (left part) and Streaming API (right part).

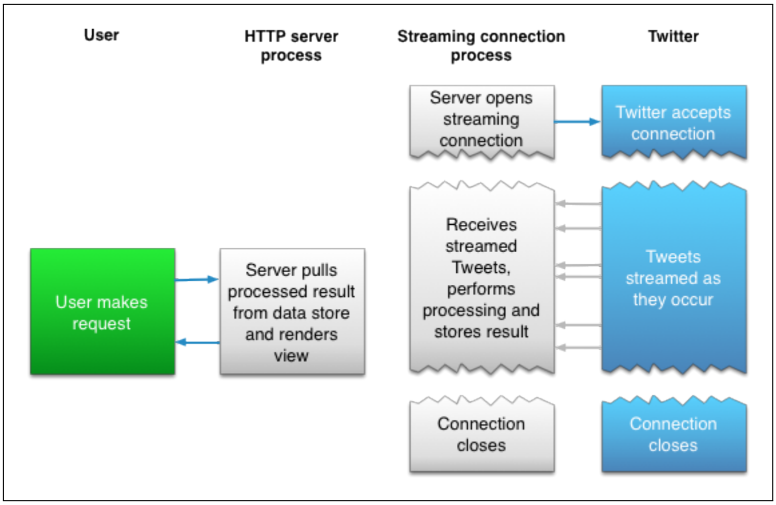

User
HTTP server process
Streaming connection process
Twitter
Server opens streaming connection
Twitter accepts connection
Receives streamed Tweets, performs processing and stores result
Tweets streamed as they occur
User makes request
Server pulls processed result from data store and renders view
Connection closes
Connection closes

Figure 2.4: Difference between REST and Streaming API

Most Web APIs belong to the first class; in case of a very long response, the Social Media implements a mechanism called *pagination*. In this case, the response is in different pages. The application must call the same API with different input parameters to ask the first page, the second page and so on. The following paragraphs will present some details of the APIs of two Social Media platforms: Facebook and Twitter.

2.4.1 Details about Facebook APIs

Facebook provides different Web APIs[12] to respond to the needs of developers, marketing professionals and so on. Moreover, the Social Network provides different software development kits (or SDKs) in different programming languages (PHP,

[10] http://instagram.com/developer/realtime/
[11] https://dev.twitter.com/streaming/overview
[12] https://developers.facebook.com

JAVA, Python, Ruby, etc). In Facebook, every object (user, post, comment, video, event, etc.) has a unique Identification (called ID). The responses of Facebook APIs are in JSON format.

The Facebook API platform is a big platform of possibility, but the company changes the APIs rapidly. For this reason, there are different web pages with the roadmap of these changes. One of these is the Facebook Platform Changelog[13]. In this site, there is a complete roadmap of news about Facebook's APIs platform. This includes the Facebook's server-side APIs, the Facebook SDK for JavaScript, the dialogs, and other services. Facebook Platform Migrations[14] is a web page that covers all the version of APIs. Facebook provides four types of free APIs: Graph, Facebook Query Language (FQL), chat, and ads.

The Graph API[15] allows users to read and write the Facebook social graph to access data of Pages, Users, Events, Groups, and Places, to publish posts, and so on. The base endpoint of graph API is *https://graph.facebook.com/*. There are two kinds of Graph APIs, one for searching data and one for capturing. The search API allow searching a type of object using one or more keywords. Generally, a search call has the form:

https://graph.facebook.com/search?q= QUERY&type=TYPE

where the QUERY is the keyword to search and TYPE is one of the following: user, page, event, group, place, placetopic and ad_*. For example, if the type is *page*, the search call will return all the entities with the keyword in the Page's title.

The second type of Graph API allow capturing data from an ID. The base endpoint is:

https://graph.facebook.com/ID

This call returns all information of an ID. The details available depends on the type of object. For example, the details of a user are different from those of a page or an event. For each object, there are a certain number of *connections* for capturing different information. Examples of a user's connections are *friends*, *feed*, *status*, etc. The connection *friends* returns all the user's friends. For example, the connection *status* returns the user's status.

Facebook Query Language[16], called FQL, is a SQL-style interface that allows a developer to query the data provided by the Graph APIs. The Facebook data are stored in different virtual tables (like a SQL table) and the user searches in them.

[13] https://developers.facebook.com/docs/apps/changelog

[14] https://developers.facebook.com/docs/apps/migrations

[15] https://developers.facebook.com/docs/graph-api/

[16] https://developers.facebook.com/docs/technical-guides/fql/

FQL has many advanced features for interrogation that are not available in the Graph API. For example, these include the possibility of requesting different queries in a single call (multi-query mechanism). For the modification of Facebook policy, the Web APIs version 2.0 are the last version where FQL is available.

A small consideration is necessary. Despite the huge quantity of data available using the Facebook APIs, not all information is available via Graph API or FQL. Moreover, for some call, the quantity of data is less than the information available on the website.

The Facebook Chat API allows integrating Facebook Chat into a Web, desktop or mobile products. The clients must use the Jabber/XMPP protocol. As FQL, this API is also available until the Platform APIs v 2.0. The Ads API (or Marketing API)[17] represents a programmatic access for Facebook's advertising platform inside the Social Network.

In addition to these Web APIs, Facebook provides non free APIs: Public Feed and Keyword Insight. The Public Feed API[18] is a stream of user status updates and page status updates. All the posts are *public*. The stream is not available via HTTP API endpoint and it is restricted to a limited set of media publishers that requires prior approval by Facebook.

The Keyword Insight API[19] is an analysis layer for Facebook posts for querying post with certain terms. This API uses the Facebook Query Language (FQL). For using a Facebook APIs platform, it is necessary to obtain an access token[20] that provides temporary, secure access. An access token is a string that identifies a user, an application, or a page. Every token has a validity time. There are different types of access tokens:

User Access Token: this kind of access token is used to read, modify or write on behalf of a specific user and it has the validity of 2 h[21]. The user access require that a person permitted access to an application obtain one via login dialog;

App Access Token: this kind of access token allow modifying and reading the application settings. It is generated when a user create a new application on the Facebook Developer site[22];

Page Access Token: very similar to a user access tokens, except that it provides permission to APIs that read, write or modify on behalf of a Facebook Page;

[17] https://developers.facebook.com/docs/marketing-apis

[18] https://developers.facebook.com/docs/public_feed

[19] https://developers.facebook.com/docs/keyword_insights

[20] https://developers.facebook.com/docs/facebooklogin/accesstokens

[21] https://developers.facebook.com/docs/facebook-login/access-tokens#long-via-code

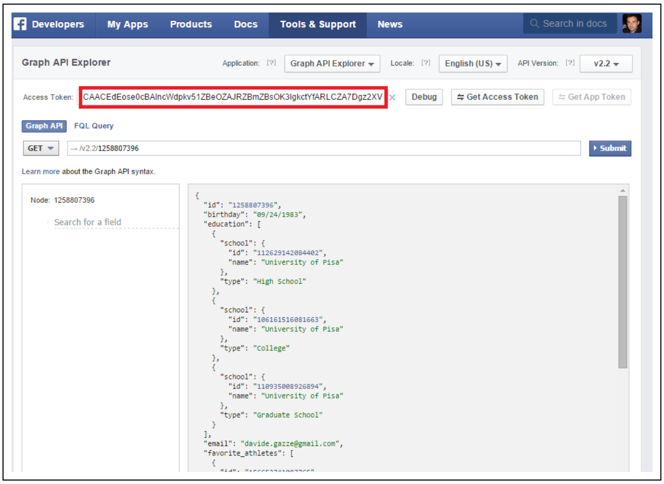
[22] https://developers.facebook.com/

Client Token: an identifier that is embedded into native mobile binaries. The client token is used to access APIs, but only a very limited subset.

If an application uses different types of access token, it may receive different information. The two figures below are a simulation of the response of the following Graph API call:

https://graph.facebook.com/1258807396

where the application uses a different access token. Figure 2.5 shows the information of *Davide Gazzè* when a user's access token is used.



Application: [?] Graph API Explorer

Locale: [?] English (US)

API Version: [?] v2.2

Access Token: CAACEdEose0cBAlncWdpkv51ZBeOZAJRZBmZBsOK3lgkctYfARLCZA7Dgz2XV

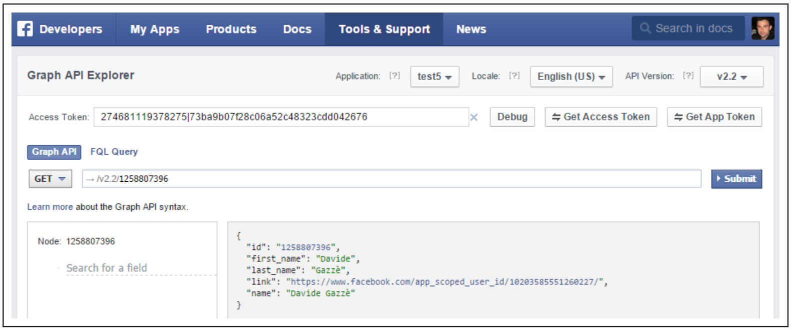
GET → /v2.2/1258807396

Node: 1258807396

```
{
  "id": "1258807396",
  "birthday": "09/24/1983",
  "education": [
    {
      "school": {
        "id": "112629142084402",
        "name": "University of Pisa"
      },
      "type": "High School"
    },
    {
      "school": {
        "id": "106161516081663",
        "name": "University of Pisa"
      },
      "type": "College"
    },
    {
      "school": {
        "id": "110935008926894",
        "name": "University of Pisa"
      },
      "type": "Graduate School"
    }
  ],
  "email": "davide.gazze@gmail.com",
  "favorite_athletes": [
    {
```

Figure 2.5: Information of a Facebook profile using User Access Token

In this case, the response includes birthday, email, and so on. Facebook public information includes the profile's ID, first name, last name, the Facebook profile URL, and the username. This case is shown in Figure 2.6. This request is about the public information of the profile *Davide Gazzè* when a program uses an application's access token.



Application: [?] test5

Locale: [?] English (US)

API Version: [?] v2.2

Access Token: 274681119378275|73ba9b07f28c06a52c48323cdd042676

GET → /v2.2/1258807396

Node: 1258807396

```
{
  "id": "1258807396",
  "first_name": "Davide",
  "last_name": "Gazzè",
  "link": "https://www.facebook.com/app_scoped_user_id/10203585551260227/",
  "name": "Davide Gazzè"
}
```

Figure 2.6: Information of a Facebook profile using App Access Token

2.4.2 Details about Twitter APIs

Like Facebook, Twitter also provides a set of APIs for third party applications[23]. Twitter exposes two different types of APIs, REST and Streaming

The REST APIs[24] provide access to read and to write Twitter data. It allows the developer to publish new Tweets, read author details such as following, follower, and more. The REST API identifies Twitter applications and users using OAuth[25] protocol.

The Streaming APIs give developers low latency access to the stream of tweets. There are three types of Streaming API:

Public stream: Streaming the public data for a specific user, hashtag, keywords, etc;

User Stream: Streaming for capturing all data from a single Twitter user's view;

Site Stream: The multi-user version of user streams.

The responses of Rest and Streaming are available in JSON. Similar to Facebook, Twitter also has a non free API. This is the *Firehose API*, a massive, real-time stream of Tweets. The Firehose API uses a streaming technology based on the protocol XMPP and it overcomes the limitations of Streaming and REST APIs.

2.4.3 Limitations of API Technologies

Despite the vast opportunities that API technology has opened up, various criticisms have been levelled at it. Particular issues are:

[23] https://dev.twitter.com

[24] https://dev.twitter.com/rest/

[25] https://dev.twitter.com/oauth

- Completeness of data;
- Validity of data;
- Reliability of data;
- Absence of transparency;
- Rate limit;
- Banning policy.

Unfortunately, “it is not clear to what extent the APIs of Social Media are actually open for researchers in the sense of offering valid and reliable access points for collecting empirical data” [72].

Bechmann and Lomborg [19] claim that Social Media research might have a critical reflection on users’ roles inside a Social Media. In their article, they analyse the behaviours of the small portion of users that develop applications with respect to the larger groups of end users. Thus, the APIs have big limitations in terms of ability to create a representative group of the entire population because most of contents are created from a small group of *active users* [46]. This is a big problem in certain domain of investigation, for example, voters’ predictions of election outcomes, because the contents created from a few active users are numerically bigger, but unrepresentative.

Both [46] and [48] introduce the concept of *lurkers*. This kind of user is one who observes all the contents of a community, without participating. Obviously, the best, but impractical, strategy for evading this problem is capturing all users of a Social Media. However, this solution is time-consuming and requires an immense server capacity. Herring (2010) [54] argues that the quality and representativeness of a sample is quite impossible to determine, if you do not know the original population.

The authors of the paper [48] have studied the limitations of the Twitter search API. Morstatter, in [76], compares a sample from the Twitter streaming API with the Twitter Firehose and he found that the streaming API misrepresents the volume of hashtags compared to the Firehose.

The issue of incompleteness have two different aspects. First, the developers do not know the APIs work, secondly the information on how the Social Media platforms filter their database is unknown [76]. These issues led to those from social media data, we can know *what* the user do, but we cannot understand the *motivation* of this [73].

Another issue regarding Social Media data is correctness. In fact, as studied in [65], the self-reported data are invalid due to the presence of profile with false information. Moreover, there are fake profiles (persons that register under false name) or programs, called bots, that act like humans [35].

Another problem is the reliability of APIs. The service providers can decide which data are available and the developers have no control over them. Moreover, if a company commercialize on specific sets of information, this data could be removed from the free APIs [48]. The non-negotiable changes in data from APIs introduce the lack of transparency. In fact, it is normal for Social Media platforms to change the behaviours of an API. In this way, an application rapidly becomes obsolete.

Another limitation of the APIs is the maximum number of calls for time that the Social Media allows. Twitter, for example, has a very strict policy. Hence, there is the probability of missing data. In that sense, the API is a fragile and unstable entry point for data collection [72]. Obviously, the above limitations can be overcome after a proper study of domains of research.

2.5 Architectures for capturing Social Media data

The problem of capture data from Social Media is important in scientific literature. Unfortunately, most of the papers present the phase of data acquisition as a necessary task for performing an analysis. Commonly, a Social Media capture module has different requirements similar to a web crawler.

An approach for capturing user profiles from Social Networks is described the work of the paper [44]. The aim of the authors is to develop a framework for collecting events from a user's profile on Facebook. To overcome the privacy limitations, the user must use his own user access token.

Using this token, the framework is able to collect, and periodically store, all the information posted in the user's profile. The term event includes the creation and/or deletion of new relationships (e.g. friendship), the post of a status, photo, video, comment, like, and share between them.

Moreover, the paper shows a simple profile modelling module that presents several user's statistical information of the use of the profile. In this way, the framework enables the creation of statistical models for users' behaviour. The second goal is the creation of normal profile usage. Subsequently, by performing a comparison of the normal model and the most recent profile usage statistics. It is possible to detect deviations that can be indicative an illicit usage of the profile (possibly because it is a case of account hijack).

The authors of [100] present a module based on Javascript for capturing data from Facebook. The authors discuss the challenges of the implementation of the crawler and they provided different suggestions to tackle the challenges. The implemented crawler starts with an initial node and it explores the others iteratively. In each iteration, the crawler visits a node and then discovers the direct neighbours.

For the exploration, two different algorithms are used. The first is Breadth-First-Exploration (BFE) which from the initial node explores all the node's neighbours. The second algorithm is Depth-First-Exploration (DFE). It starts with an initial node and explores each branch. For instance, the crawler captures Facebook profiles of people from Macao. Finally, the authors analyse the social data using some common Social Network Analysis (SNA) techniques and visualize the result using different layout algorithms.

This kind of work is very important, but lack of a generic architecture for capturing. The authors of the paper [59] propose an initial approach. Sponsored by Microsoft, it is a generic architecture, called Social Stream Blog Crawler, able to capture data from blogs and other Social Media platforms. The main issue in Social Media Capture is the temporal constraints of the content. For this reason, the application is scalable.

The characteristics of the proposed architecture are: Real-time, Coverage, Scale, and Data Quality. The real-time constraint is respected because the information of blogs is time-sensitive and it is already old a short time after its publication.

The coverage feature is necessary to fetch the entire blogosphere. Moreover, the system must scale with the number of blogs. The last feature, *Data Quality*, means the capability of the architecture to capture uncorrupted data. The Social Stream Blog Crawler is a project at Microsoft's Live Labs.

In paper [29], the authors analyse the topic of scalability. In this work, this problem is solved with parallel crawlers. The authors exploit eBay and the user's profile. A characteristic of eBay's profile is that data are well structured and each user has a unique identifier.

The solution proposed use a master machine, while the crawling of a user's profile is distributed across different agents. Each agent requests the master for the next available user to crawl, and returns the crawled data. The master maintains global consistency and it ensures that a user is crawled one time. The system uses a MySQL database for storing information.

As a result, the system captured 66,130 of 11,716,588 users from 10/10/2007 to 02/11/2017. The results of this study confirm that the bottleneck of this task is the download-rate of Internet connection.

The above papers do not present any theoretical model able to describe Social Media. An approach of this sort is found in paper [81]. The model is the Artefact-Actor-Networks (ANNs). The term artefact network indicates the use of the information objects and their connections. This model is an approach to semantically interconnecting social networks with artefact networks. The relations in the network are between objects with the semantic context of *isAuthor* or *isRightHolder*. The architecture of Artefact-Actor-Networks is composed of a backend and several frontends. Backend has the functionalities of storing and processing the data and it uses the OSGi Service Platform for communication. The applications that execute the data capture are the CrawlerManager and the Crawler.

The Crawler provides low-level functions to process tasks (URIs). The CrawlerManager exposes services used by the Crawler. When a Crawler receives tasks for exacting URI, the CrawlerManager handle complex jobs. In this way, the module can analyse the structure of a page using the HTML hyperlinks.

In [32], the authors take into account the Twitter platforms to improve the volume of data. The authors compare different sampling methods for the selection of nodes. In particular, the authors study the use of attributes (location and activity) and topology (forest fire). The results suggest that the use of attributes can improve the volume of data by 15-20

In [87], the authors describe the architecture and a partial implementation of a system designed for monitoring and analysis of communities on Social Media. The main contribution of this paper is an architecture able to monitor diverse Social Media platforms. It consists of three main modules, the crawler, the repository, and the analyser. The first module, based on ontology, can be adapted to crawl different Social Media platforms.

The problem of collecting of massive quantities of data from Facebook is the topic of the paper [28]. Authors approach to the capture step with different approaches. They develop different crawlers to capture two large samples (with millions of connections). The data are stored as an undirected graph with anonymised nodes. The purpose of this study is the evaluation of different sampling methodologies. The first method is the Breadth-first-search (BFS). This is well-known graph traversal algorithm for unweighted and undirected graphs. The second method is Uniform sampling (Gjoka et al. [47]).

2.6 Introduction to Social Media Analytics

Social Media platforms are an effective, sophisticated, and powerful way to capture preferences and activities of groups of users [84]. For this reason, Social Media data have become very important in different areas of analysis [68].

This field is Social Media Analytics. It exploits User Generated Content (UGC) to achieve a specific goal [57]. According to Zeng et all [101], “Social Media analytics is concerned with developing and evaluating informatics tools and frameworks to collect, monitor, summarize, and visualize Social Media data, usually driven by specific requirements from a target application". There are various techniques in Social Media Analytics: Social Network Analysis, Sentiment Analysis/Opinion Mining, Insight Mining, Trend Analysis, Topic Modelling, Influence Analysis, etc.

Obviously, we will not introduce each field, but we provide an overview of some of them. In particular, we introduce the following research domains: Online Reputation, Social Network Analysis, Sentiment Analysis, and Early Warning of disasters. Moreover, we will present some topics for the exploitation of Social Media in business.

2.6.1 Online Reputation

An important field of Social Media Analytics is the analysis of time series to classify or predict the popularity of an entity. The growth of Social Media platforms allows capturing a user's content. Different websites collect, elaborate, and visualize statistics about these contents.

Starcount[26] provides the rank of the most popular people. Starcount collects data from 11 popular Social Media (Facebook, Twitter, YouTube, Google+, etc) and it ranks users. In order to build rankings, Starcount assigns a score to each entity. The Starcount score is the weighted sum of the popularity that the entity has on every Social Media. PageData is a service for tracking statistics about Facebook Pages discussed in [78] and [25]. For each entity, it shows trends on new likes per day.

Social Media Ranking [26] is an Austrian website that shows the top scored Austrian entities of the week. In order to build the ranking, it takes data from Facebook, Twitter, Google+, and Foursquare and it combines them to build its own score.

Twitalyzer[27] and Twitaholic[28] are web applications that provide statistics about single entities. Twitalyzer use Twitter data and it covers numerous metrics, such as impact, engagement, influence, klout and generosity. Socialbakers[29] is a website that exploits Facebook, Twitter, YouTube, Google+, and LinkedIn. It builds a

[26] http://www.Starcount.com/

[27] http://www.twitalyzer.com/5/index.asp

[28] http://twitaholic.com/

[29] http://www.socialbakers.com/

rank of entities, grouped by category. Blogmeter[30] and Reputation Manager[31] are proprietary systems providing business solutions for companies.

2.6.2 Social Network Analysis

The social graph is the representation of individuals in a certain context. The nodes of the graph represent the users, while the edges (also called ties) are social relations between nodes. This field of research, called Social Network Analysis (SNA), investigates the relationships among people and the information flow in groups. The SNA uses different metrics such as centrality, betweenness, and closeness.

There are various books and papers with the purpose of studying the structural part of social network analysis; of them we cite [16], [49], and [50]. Papers such as [41] and [62] study Facebook and Twitter respectively. There are many free tools for SNA, such as Cytoscape[32], NodeXL[33] [89], and Gephi[34].

In relation to our knowledge, with SNA it is possible to study a single person (ego) or a community.

Ego networks are a kind of social network in which the relationships are between one person (called ego) and all the individuals directly connected to ego (alters). These networks are representations of human personal social networks and are largely studied in the anthropology literature [39], [38], [50], and [82]. Given two individuals, their tie strength is a numerical representation of the importance of their relationship. In [50], the authors study the frequency of contact as a predictor for social tie strength.

The diffusion of social networks is fostering the availability of social interactions between people. Several conjectures made by sociologists, on social networks, have been confirmed by the use of data obtained from the Internet [36], [55], and [71]. In [11], the authors address this problem by analysing results from a measurement study on Facebook users. The goal of this study is to investigate whether the structure of ego networks formed on Facebook is similar to the structure of ego networks observed in real human networks.

The paper takes into account two classes of variables, the socio-demographic variables, such as the size of the ego network of the users and various factors possibly impacting on it (e.g., age, gender, etc...), and the relational variables, which are related to the strength of the social relationships between each ego and her

[30] http://www.blogmeter.it/

[31] http://www.reputazioneonline.it/

[32] http://www.cytoscape.org/

[33] http://nodexl.codeplex.com/

[34] http://gephi.github.io/

alters. Moreover, the paper presents an initial correlation analysis, showing how the relational variables correlates with the real tie strength of social relationships.

In [99], the authors use the interactions graph to investigate whether Social Links are good indicator of real user interactions.

The above papers characterize an ego network; other papers study the evolution of a network. For example, the topic of paper [96] is the evolution of activity between users on Facebook. Instead, in [15], there is the exploitation of friendship links and community membership on LiveJournal, and co-authorship and conference publications in DBLP for characterizing the evolution of these networks.

2.6.3 Social Media Intelligence

Due to the high availability of personal information on users, authorities, such as Law Enforcement Agencies (LEAs), exploit a Social Media platform to find criminals and terrorists. This field is the Open Source Intelligence (OSINT). It is defined as the techniques for gathering and analysing the information from public sources, namely Open Source Information (OSINF). The term Social Media Intelligence (SOCMINT) indicates the exploitation of information from one or more Social Media to capture the *wisdom of crowds* [79].

The data of criminal networks have implications in anti-terrorism campaigns and, in general, in investigation actions. In the paper [31], the authors identify three possible categories of criminal groups that exploit information and communications technologies (ICT) in illegal activities. These groups are the traditional organized criminal groups that use ICT to enforce their terrestrial criminal activities, the organized cybercriminal groups that only work online and the politically motivated groups that use ICT to facilitate their criminal activities.

Terrorist groups and sympathizers use virtual communities like Facebook, MySpace, Second Life, and the Arabic Social Media. Recent analysis suggests that since the late 2000s activity has increased in Social Media platforms. According to Aaron Zelin “it is only a matter of time before terrorists use Twitter and Instagram as part of ongoing operations”[35].

All over the world, there are various projects in OSINT. For example, the European Project CAPER is a Collaborative Platform for Open and Closed Information Acquisition, Processing and Linking. The goal of this project is to prevent organized crime by exploiting and sharing different information sources [7] and [5]. The aim of the American Project IARPA[36] (Intelligence Advanced Research Projects

[35] http://www.washingtoninstitute.org/policy-analysis/view/the-state-of-global-jihad-online

[36] http://www.iarpa.gov

Activity) is the continuous and automatic monitoring of public data for predicting events.

2.6.4 Sentiment analysis and Social Media

As previously mentioned, Social Media data allow us to acquire the wisdom of crowds [79]. This fact is very important for a decision support system. Sentiment analysis is a type of natural language processing (NLP) for monitoring the opinion of a product, a topic, a person, a place, etc. Sentiment analysis is involved in different studies for examining the opinions from different sources like blog posts, comments, and reviews [95].

The paper [77] is a study carried out on over than 1000 Facebook posts about newscasts, comparing the sentiment for two Italian public broadcasting services: Rai (the national service) and La7 (a private company). The result highlights that Facebook is a good platform for online marketing. Moreover, with Facebook it is possible to measure customers' interests and their feeling about products or brands.

The paper [27] presents an analysis of Italian Twitter users during the period of national political elections. The aim of the work to monitor the volume of the leaders' tweets. Furthermore, the authors compare the tweets with the results of elections. Statistical analysis of the data does not predict the election outcome, but it provides an approximation.

The public opinion on important topics, such as nuclear power plants is a relevant application of monitoring of citizens' opinions. This paper [67] proposes an approach to monitoring public sentiments on nuclear topic on Twitter. The process consists of different steps: (1) crawling tweets, (2) text preparation, (3) sentiment dictionary construction, and (4) sentimental scoring. Based on an experiment with the nuclear related tweets in Korean between 2009 and 2013, the authors verify the usefulness of their approach and confirm that the changes in national opinion on nuclear generation depends on critical events such as the Fukushima Daiichi nuclear disaster.

2.6.5 Early warning via Social Media

Social Media platforms can be sources of information about situations and facts related to the social environment. In this research field, every user is like a social sensor. This paradigm has the name of Human as a Sensor (HaaS) paradigm and contains terms such social sensing, crowd-sourcing, citizen sensing. The focus of this field is the exploitation of the participation of citizens for social warning [102].

Emergency Management is a field for Social Sensing that exploits Social Media contents created in case of an emergency or a disaster. This critical information can help create decision support systems for emergency services and notify government authorities. The advantage of this methodology is the spontaneous participation of the users and the contribution of citizens without any pressure or influence. The term Early Warning System (henceforth EWS) indicates an information system for detecting dangerous events of social concern [97].

As there are different dangerous events, there are also different projects, but the central point is the platform to exploit. Twitter is a microblog platform that counts more than 645 million active users and 58 million messages shared every day. Moreover, Twitter has adopted a policy and a message format that encourages users to make their messages public by default. On Twitter, users share more specific content due to the limitation of 140 characters per tweet.

The global spread of the Twitter phenomenon enabled a new wave of experimentation and research. In fact, different studies, such as [69], claim that Twitter is a News Media. For the above considerations, Twitter is a good source for obtaining details of an event's impact and it is good enough as a base for social sensing platforms.

The paper [85] is a study about the real-time detection of tornadoes in Japan. The authors had created a EWS based on Bayesian statistics. In this study, the data acquisition module exploits the Twitter Search APIs. Unfortunately, the APIs can access only a portion of all tweets, so some events are undetected.

An innovative approach is the Italian project SOS (Social Sensing) [13], [12]. In this project, the people are *social sensors*. The authors analyse what users write on the most popular Social Media to identify particular events such as earthquakes, floods, civil unrest or other emergencies. The goal is to build a decision support system for emergency management that can analyse in real time the content on Social Media. In [13], the platform EARS (Earthquake Alert and Report System) is described. The platform exploits the Twitter Streaming API to provide a decision support system for INGV (National Institute of Geophysics and Volcanology) in case of earthquakes.

2.6.6 Social Media in Business

In [68], the authors introduce Business Social Media Analytics as: "Business SMA refers to all activities related to gathering relevant Social Media data, analysing the gathered data, and disseminating findings as appropriate to support business activities such as intelligence gathering, insight generation, sense making, problem recognition/opportunity detection, problem solution/opportunity exploitation, and/or decision making undertaken in response to sensed business needs”. The paper [57] outlines the following benefits of this field of study:

Improve Marketing Strategy: Customer-generated content is a valuable source of information for a product or a service [68] [58];

Better Customer Engagement: SMA may identify new channels for two-way communication [4];

Better Customer Service: SMA can provide better customer service [56];

Reputation Management: Social Media platforms offer information for monitoring the reputation related to brands, products, services, etc. [8];

New Business Opportunity: SMA may identify new potential customers or new untapped business opportunities [33].

For example, i [90], Sterne uses the measure of volume of UGC about a product or service for predicting the relative impact on sales.

2.7 The CAPER project

The EU FP7 Project CAPER provides a collaborative platform for the detection and prevention of organised crime in which the Internet and Social Media are used (e.g. cybercrime, terrorism, and counterfeit). The end users of this platform are the European Law Enforcement Agencies (LEAs). The CAPER platform allows the integration with legacy systems and it provides modules for gathering and analysing data. CAPER comprises multiple modules that use a service-oriented architecture (SOA) for interoperability. The publicly available data are used for different kinds of analysis; textual analysis for the identification of named entities, biometric analysis for the detection of people in images, audio analysis for speech recognition, etc. Two important sources of information are the Web and the Social Media platforms. For the latter one, we exploit the User Generated Content (UGC) offered by Facebook to analyse interactions between people in terms of strength, frequency and duration.

2.8 The OpeNER project

OpeNER (Open Polarity Enhanced Name Entity Recognition) is a project funded under the 7^{th} Framework Program of the European Commission. Its main objective is to implement a pipeline to process natural language. More specifically, OpeNER focuses on building a linguistic pipeline supporting six languages (English, Spanish, German, French, Italian, and Dutch) that enables the identification and disambiguation of named entities and the analysis of sentiment of content. Within OpeNER, we have developed Tour-pedia project, a Web application that exploits the linguistic pipeline to extract the sentiment of places in tourism domain. In particular, we analyse places belonging to different typologies: accommodation, attraction, point of interest, and restaurant. In detail, each place is associated whit zero or more reviews extracted from SM. Each review is processed by the OpeNER pipeline to extract its sentiment. The sentiment of a place is calculated as a function of all the sentiments of the reviews on that place. As a result, Tour-pedia shows all the sentiments of all places on a map. Tour-pedia guides the user to choose the most suitable solution for his or her needs. In addition, it helps the user to overcome the common problems of tourism platform such as the fragmentation of reviews in different sources. In fact, users generally need to explore several pages on the web to extract information, but in Tour-pedia we exploit the Web API of four Social Media Platforms: Facebook, Foursquare, Google Places, and Booking.com.

3

Social Media Data Capture

In this chapter, we will introduce a generic architecture for capturing data from Social Media. As described in the previous chapter, one of the most promising techniques is the Web API. However, this modality has the following limitations:

- Completeness of data;
- Validity of data;
- Reliability of data;
- Absence of transparency;
- Rate limit;
- Banning policy.

Obviously, the above criticisms depend on the Social Media platform considered. For example, on Twitter the Rate Limit Policy is more stringent than on Facebook. However, we can mitigate some of above criticisms with a proper software architecture. In order to create an architecture for capturing data from Social Media, several considerations are necessary. First, every Social Media platform manages different types of elements. For example, the elements inside Facebook are: user, page, group, event, and place. Sometimes there is only one type of element, for example on Twitter there is only the user's account. This variety creates the need for architecture allowing different data schemas. Moreover, the architecture must guarantee that a change in an entity's data schema does not cause the lose of data. Henceforth, we will use the following concepts.

An *entity* is an abstract concept that represents something in the world (e.g. person, event, group, organization, brand, etc.). Every entity has a set of attributes (ID, name, country, etc.).

A *channel* is the Social Media where each entity can perform one or more of the following operations [23]:

- Build its profile;
- Share its profile with other individuals;
- Communicate with other entities.

A *source* is a web place of an entity on a specific channel. Every source has a unique identification (called ID) that the channel provides. The ID is the declaration that an entity is registered on a Social Media. With the term *social metrics*, we denote all the statistical information about a source. An example of a social metric on Facebook is the number of fans, which is the number of people who like the source associated with that entity. The verb *Crawling* represents the action of capture data from a Social Media. A software module, called *crawler* or *sampler*, performs this task.

3.1 Generic architecture for Social Media data capture

In this paragraph, we present a generic architecture for capturing data from Social Media platforms. In accordance with [88] and [9], the capture architecture must satisfy the following properties:

Scalability: the architecture must work with large volumes of data;

Configurable: the crawler may observe some parameters such as refresh time or maximum number of parallel crawlers;

High Performance: the system must able to run on different machines;

High variance of data: the system must be able to store and to manage different kind of contents: Text, Image, and Video;

Resilience: the system must overcome problems such as absence of Internet connection or unavailable response;

Adaptability: ability to exploit particular features of Social Media. For example, Facebook provides search API and YouTube allows retrieving metadata for a single video.

From the above considerations, we divide the requirements of the architecture into functional and non functional. The functional requirements are:

1. Ability to let end user describe the source that will be captured;
2. Ability to capture all available data of a Social Media;
3. Ability to store the temporal information of moment of capture (when a source is captured);
4. Ability to refresh the stored information of a source;

5. Ability to implement some algorithms such as the snowball-algorithmic [93] and the Breadth-first-search (BFS) [28];
6. Ability to capture data from the unique identification of a source;
7. Ability to implement search functionalities (if a search API is available).

Figure 3.1 shows the proposed architecture.

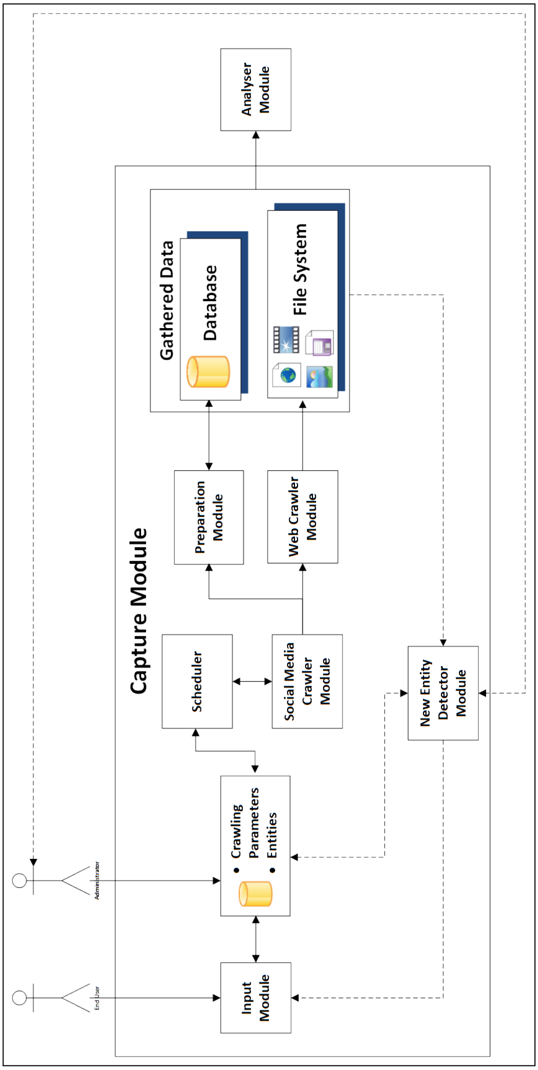


Figure 3.1: Architecture of a generic Social Media Crawler

The non-functional requirements are:

1. The system must allow the creation of parallel instance of crawler;
2. The system must be robust with regards to change of data schema;
3. The system must be able to store the data and some crawling statistics (the number of calls to Web API performed, the successful and failed calls);
4. The system must help end user and administrator to establish whether an entity can be crawled;
5. The system must provide statistical information of all sources captured.

The actors of the system are:

End User: uses the platform;
Administrator: modifies the system parameters.

Now we will detail the functionalities of each module. The Input module is the interface that allows end-users and administrator to:

- Change global parameters of architecture;
- Insert new sources to capture.

Moreover, the input module must help the user to:

- Select the source to capture;
- Indicate what information is necessary.

The input module can be implemented as a web page, a web service or a mobile application. The Scheduled is the module that every n_activation seconds (where n_activation is a global parameter) performs the following actions:

- Start a new instance of Crawler with a specific source;
- Check if a Crawler ends with an normal status;
- Restart the failed Crawler.

The Social Media Crawler is a module able to capture one or more sources. The module exploits the Social Media's API and it stores the raw data in a Database. The Social Media Crawler can use some scraping techniques if necessary.

The Preparation Module is an optional application that allows transforming the raw data to another format (for example conversion from JSON to XML).

The Web Crawler Module is an optional application that exploits HTTP calls to download web pages, images, documents, videos, etc.

From the captured data, we can infer new sources to crawl. This task is performed by the New Entity Detector Module. This is an optical application that enables one to discover new sources to capture from the original raw data. This module can work in supervised, semi-supervised or unsupervised modalities. In our architecture, this module, in unsupervised modality, can implement the Snowballs algorithm [93], the Breadth-first-search (BFS) [28] or Uniform sampling [47].

The Analyser Module analyses the captured data. The implementation of this module depends on the type of study performed. In fact, a module that analyses the interaction between users, is conceptually different from the one that analyses the users' opinion on a brand. Some examples of analyser modules will be presented in the next chapter. For storing the data, we chose three different technologies:

- The file system for multimedia documents (Documents, Images, Videos, websites);
- A Relation database (like MySQL) for storing the system parameters;
- A NoSQL database (like MongoDB) for storing the raw data.

Each technology responds to specific requirements. In particular, the file system is chosen for its capability to store a huge volume of unstructured data (in particular videos). The database MongoDB is necessary for storing semi-structured data, like JSON, and, all information without any fixed schema.

As detailed in 2.3.2, in the file system, we can store every type of data, but this choice led to different management problems.

In the next paragraphs, we will present a simplified implementation of the proposed architecture to satisfy the requirements for three different domains.

3.2 Online Reputation: model

Nowadays most politicians, singers, journals, public people and companies have an account on a Social Media platform. With these tools, the entity attempts to promote its reputation. For example, a politician will improve his own reputation or a company will try to attract more consumers. Obviously, presence in Social Media is not an assurance of success, thus, the field of study *Reputation Management* on Social Media became very important.

Historically, this concept is related to public relations with mass media. Today, Reputation Management on the Internet and on Social Media is the state of the art. Without loss of generality, it is possible to define the reputation of an entity as the popularity of a profile on one or more Social Media platforms (Facebook, Twitter, YouTube, etc.).

In this field of study, we use the previous abstract model. On Twitter and YouTube, the source is called an *account*. A source can generate content, which includes all the activity inside the channel. These statistics are the indicators of the entity's reputation on the Social Media. We model a channel as directed graph G(N;A), where sources S and content C represent nodes N and the actions between a source and content or between two sources are the arcs A. An arc from a source to a content exists when the source generates that content, while an arc from a source to another one exists when the first expresses an interest in the second.

We define the indegree deg$^-$(N) and the outdegree deg$^+$(N) of a node N as the number of arcs pointing to N and outgoing from N, respectively.

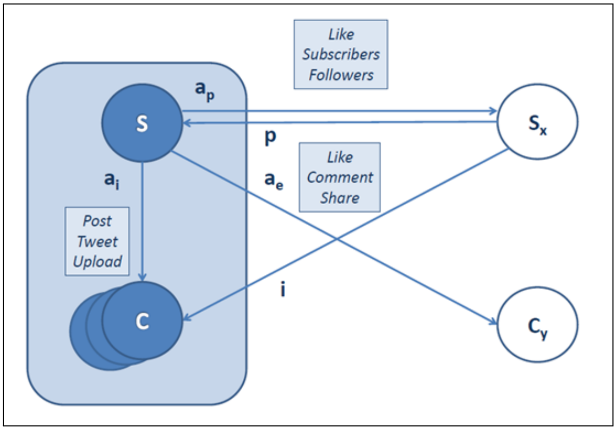


Figure 3.2: Relationship between two entities

Figure 3.2 shows how two sources S and S_x are connected. In order to simplify the diagram, only the relationships to S are shown. By C and C_x we denote content produced by S and S_x. The arc pointing from S_x to S, namely p, represents the interest that S_x has in S. The arc from S to C, namely a_i, represents the action performed by S when it produces its own content, while a_e represents an action performed by S on the content produced by S_x. Finally, the arc from S_x to C, namely i represents an action performed by S_x on the content produced by S.

The *popularity* of a source on a channel is the level of attention it receives from other sources [83]. More formally, the popularity $\mathcal{P}_i$ of the source S_i is the indegree of S_i:

$$\mathcal{P}_i = deg^-(S_i) = |p| \tag{3.1}$$

where $| * |$ represents the cardinality.

The *activity* of a source on a channel is defined as the frequency of contents publication. More formally, we define the *activity* $\mathcal{A}_i$ of a source S_i as the outdegree of S_i:

$$\mathcal{A}_i = deg^{+}(S_i) = |a_i| + |a_e| + |a_p| \tag{3.2}$$

The *influence* of a source on a channel is the feedback that it receives on its generated content. We define the *influence* $\mathcal{I}_i$ of a source S_i as the sum of the indegrees of all the contents C produced by S_i

$$\mathcal{I}_i = \sum_{C \in \{C_i\}} deg^{-}(C) = \sum_{C \in \{C_i\}} |i_i| \tag{3.3}$$

where C_i is the set with all the contents produced by S_i.

The Reputation of an Entity is time variant, so the architecture must store the value of every metric for every period. In fact, the value of a metric in the range of time from

$$t_i$$

to

$$t_{i+1}$$

is lost if the system samples it in that interval. Moreover, it is important to emphasize that some entities can belong to more than one category. An example of an entity belonging to only one category is the creator of Facebook, Mark Zuckerberg. He belongs to the CEO (Chief Executive Officer) category. Instead, Silvio Berlusconi belongs at least to two categories: CEO and Politician. In our proposal, we compare only entities of the same category.

3.2.1 Social Media data on SocialTrends

As described in the previous paragraph, the requirements for monitoring the entity's web reputation are the following:

- Ability yo capture the popularity, activity, and influence of a source;
- Ability to refresh source's metrics every period.

From these requirements, we implement the project SocialTrends, a web application that collects, elaborates, and visualizes data from Social Media. SocialTrends is focused on the match of entities that belongs to the same category. As previously described, the proposed model is composed of four elements: Entity, Channel, Source, and Metrics (see Figure 3.3).

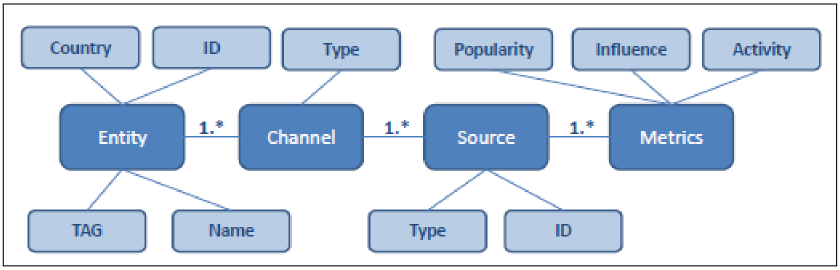


Figure 3.3: SocialTrends' Model

In SocialTrends, we exploit the metadata from three different platforms: Facebook, Twitter, and YouTube. Then we map the metadata in the three groups: popularity, activity, and influence. For every Social Media, we select only the official account of the entity. A certain consideration is necessary regarding Facebook. In this platform, a generic user can have a personal profile (with a maximum of 5,000 friends) and a public page (with millions of Fans). We considered only the official page of entity because it promotes the entity's reputation. Table 3.1 shows the mapping of Social Trend in the generic architecture model.

Social Media	Popularity	Activity	Influence
Facebook	Fan	Post	Like, Comment, Share
Twitter	Follower	Tweet	Reply, Retweet, Mention
YouTube	Subscribed	Video	Visualization, Like, Dislike, Comment

Table 3.1: Example of Facebook page with related actions

In our implementation, SocialTrends analyses only the numerical values of these statistics without taking in account the textual contents, the photos, and the videos. Figure 3.4 shows SocialTrends architecture. It is composed of five modules: the Sampler Module (SM), the Social Analyser Module (SAM), the Data Visualization Module (DVM), the Administrator Module (AM), and the Database (DB).

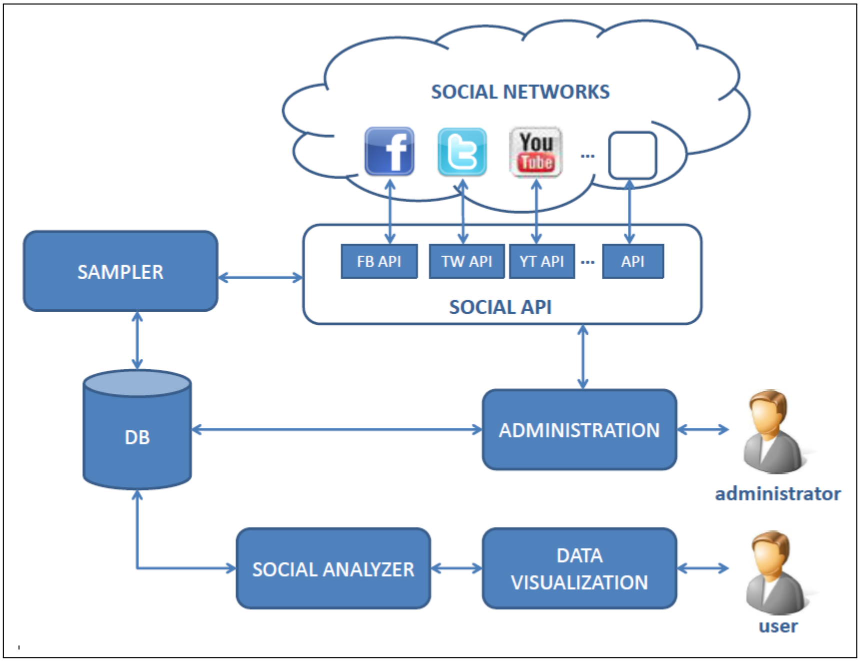


Figure 3.4: SocialTrends Architecture

We do not save all raw data on MongoDB, but we saved only the metrics in a MySQL database. In fact for analysis purposes, only a small amount of information is necessary.

Let us now discuss some implementation details. The Sampler Module (SM) and the Social Analyser Module (SAM) are developed with the PHP language. Instead, the Data Visualization Module (DVM) and the Administrator Module (AM) use HTML, CSS, and Javascript. Moreover, we use the libraries Jquery[1] and Hightcharts[2].

To refresh the information on a source, the Sampler Module repeats the capture of the source metrics every hour. This feature is implemented using the Linux CRON daemon.

In SocialTrends, we propose a simplified implementation of our generic architecture. In particular, we merged several modules to simplify the project and we

[1] http://jquery.com/
[2] http://www.highcharts.com/

added the module DVM, which represents the Home Page of SocialTrends. Table 3.2 shows the mapping of the SocialTrends module and the proposed architecture.

General Architecture	SocialTrends
Input Module	Administrator Module
Scheduler Social Media Crawler	Sampler Module
Web Crawler Module	NOT PRESENT
Preparation Module	NOT PRESENT
New Entity Detector Module	NOT PRESENT
Analyser Module	Social Analyser Module

Table 3.2: Mapping of Generic Architecture module in SocialTrends

3.3 Social Media Intelligence: Considerations

The final goal of this work is the development of useful applications for Law Enforcement Agencies (LEAs) in intelligence activities based on Social Media. The contributions of our work covers the aspects of capture and analyse. First, we create a module, called Social Media Capture (SMC), able to capture data from Facebook. Secondly, a module, called Social Media Analyser (SMA), processes the capture data. Figure 3.5 shows the simplified version of our work.

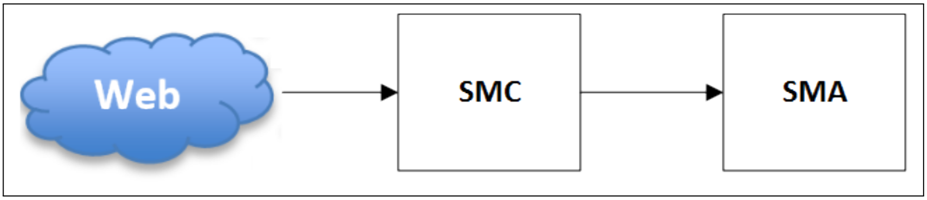


Figure 3.5: SMC and SMA

In this paragraph, we will present Social Media Capture (SMC). In the CAPER project, this module is responsible for capturing data from Social Media.

In the Social Media Intelligence field, the concept of a capture module is different from Online Reputation. In fact, the LEAs are interested in contents (users' name, posts, contents, etc.) and interactions of a Social Media. This is the main difference with respect to the Online Reputation field.

The SMC captures every information of a source. From a conceptual point of view, the SMC allows an end user to make a *virtual photo* of a source. With this modality, the volume and variety of raw data are bigger with respect to SocialTrends. For these reasons, we choose MongoDB as storage technology (in accordance with the described architecture).

For correct design of a Social Media capture, we must consider all the issues. The first is the technical restrictions imposed by Social Media. In fact, this question is connected to the user's privacy. On all Social Media, the user can choose to show information only for a group (for example the user's friends). On Facebook, a generic user can see the photo of a friend, but he might not see the photo of a non-friend (except profile and cover photos). Moreover, there are particular places where the discussion is public, such as the Facebook Pages. For this consideration, the SMC allows using the user credentials of the Social Media.

Second, the SMS must allow crawling of different Social Media's sources. (data with different schemas and images). For this, SMC uses both MongoDB for data and metadata and the file system for the images.

Another issue is that the operation of "capture" can be long. For this, the SMC allows multiple instance of samplers.

The question of what Social Media platform is useful for the LEAs is also very important. The answer to this question is not trivial because it is necessary to take into account the following requirements:

- Number of users in the Social Media;
- Growth rate of users in the Social Media;
- Geographical distribution of the users;
- Sociological factors of the users;
- Availableness of Web APIs.

From this point of view, we started with the statistics of active users on a Social Media. For this, we studied the distribution of Social Media all over the world in the years 2011 and 2012.

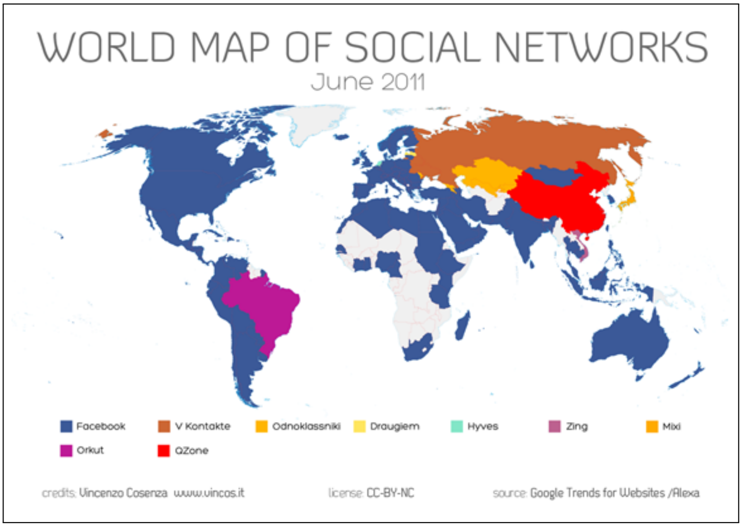

WORLD MAP OF SOCIAL NETWORKS
June 2011
Facebook
V Kontakte
Odnoklassniki
Draugiem
Hyves
Zing
Mixi
Orkut
QZone
credits: Vincenzo Cosenza www.vincos.it
license: CC-BY-NC
source: Google Trends for Websites /Alexa

Figure 3.6: Diffusion of Social Networks platforms in June 2011 (source: http://vincos.it)

Figure 3.7 shows the diffusion of Social Network in 2012.

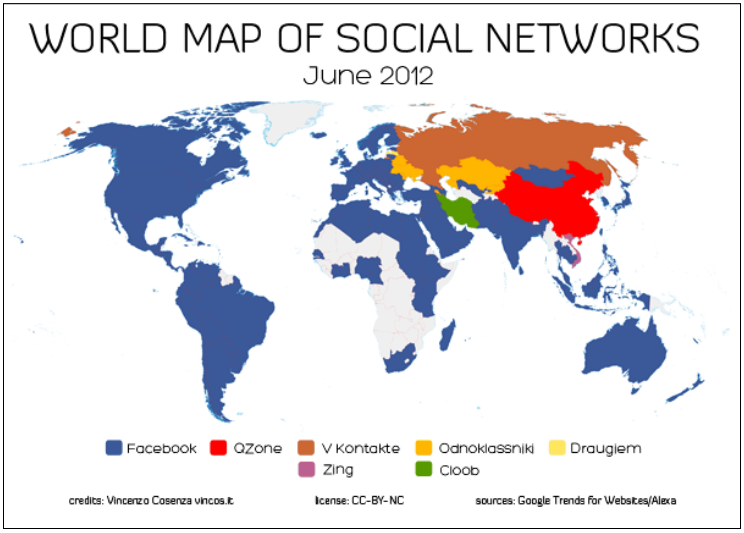

WORLD MAP OF SOCIAL NETWORKS
June 2012
Facebook
QZone
V Kontakte
Odnoklassniki
Draugiem
Zing
Cloob
credits: Vincenzo Cosenza vincos.it
license: CC-BY-NC
sources: Google Trends for Websites/Alexa

Figure 3.7: Diffusion of Social Networks platforms in June 2012 (source: http://vincos.it)

From Figs 3.6 and 3.7, we can see that the best candidate is Facebook. Unfortunately, Russia is not well covered (though obviously many Russians have Facebook account).
The report for 2014 is available at the URL[3].

3.3.1 Social Media Capture: the Facebook case study

Facebook is a huge virtual world that contains more than 1 billion of active users[4] and it has five different types of elements:

User: profile of a person;
Page: page used for promoting a brand, an association, a corporate or a public person;
Group: page used for discussion about a limited number of topics (like a Forum);
Event: page used for organizing real events (disco, flash mob, party);
Place: page with geographic information, used for marketing goals.

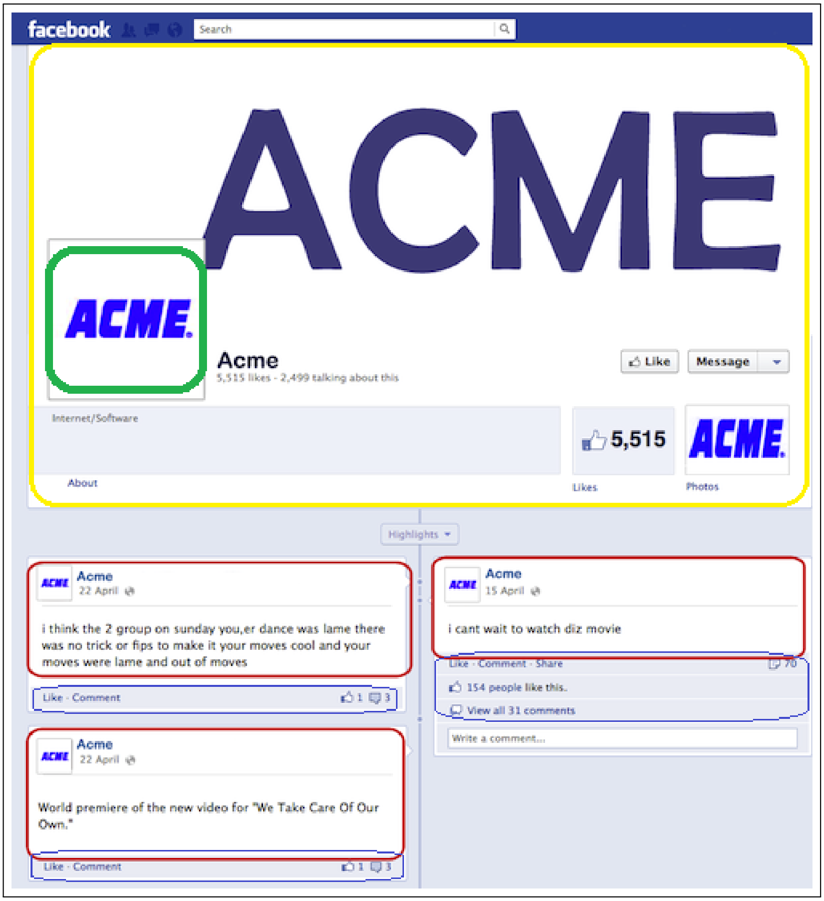
The SMC crawls pages, events, groups, and users. The places are useless for this task. The simplest Facebook's entity is the page. As shown in Fig 3.8, It is composed of four different parts:

- Yellow area represents the page information;
- Red area represents posts with the related message (text);
- Blue area represents the comments, likes, and shares related to a post;
- Green area is the page picture.

The total number of posts reported on the page represents the *feed*. In addition, a Facebook's Page contains different images in the info area (profile and cover) and within the posts.

Group and Event entities have different layouts but similar data. For a group, it is possible to retrieve the list of members (with the information of the administrator). Instead, for an event, it is possible to capture the people invited to the event and their intention regarding attendance (to go, not go or may be go). For every user, the SMC can capture different information: text, image, and the friend list (in same case).

[3] http://vincos.it/2015/02/04/la-mappa-dei-social-network-nel-mondo-dicembre-2014/
[4] http://newsroom.fb.com/company-info/



i think the 2 group on sunday you,er dance was lame there was no trick or fips to make it your moves cool and your moves were lame and out of moves

i cant wait to watch diz movie

World premiere of the new video for "We Take Care Of Our Own."

Figure 3.8: Example of Facebook Page

The current version of the SMC performs two types of capture operations:

- By Facebook ID;
- By keyword(s).

For each capture operation, the SMC captures the following kind of data:

- Profile information of the source (fan count, subscriber);
- Images of the source's profile, source's cover, and posts;
- Geographic information of source (in particular for event);
- Corpora (post, comment);
- Post metadata (number of likes, users who likes the post, mention);

- Comment metadata (number of likes, users who like the comment, mention, users who respond to a comment);
- Creation time of post and comment;
- Update time of post and comment.

3.3.2 Social Media Capture Implementation

As previously described, Social Media Capture (SMC) is a module able to crawl Page, Group, Event, and User from Facebook using the ID or some search terms. Unfortunately, some entities are private. Examples of these are the user and some groups. Table 3.3 shows the privacy setting of entities on Facebook.

Type	Only public	Only Private	Either Cases
User		X	
Page	X		
Group			X
Event	X		
Place	X		

Table 3.3: Privacy of Facebook's entity

Similarly to the SocialTrends architecture, Table 3.4 shows the mapping from the module of generic architecture and the SMC. The Social Media Analyse (SMA) will be presented in the next chapter.

General Architecture	SMC
Input Module	Web Services
Scheduler	Sampler Module
Social Media Crawler	
Web Crawler Module	
Preparation Module	Output Module
New Entity Detector Module	NOT PRESENT
Analyser Module	Social Analyser Module

Table 3.4: Mapping of modules of the proposed Architecture and SMC

The implemented version of SMC uses the Graph API (Application Programming Interface) and FQL (Facebook Query Language). Details about these technologies are in paragraph 2.4.1. SMC is developed in PHP and it uses the PHP Facebook SDK[5]. Figure 3.9 shows the SMC architecture.

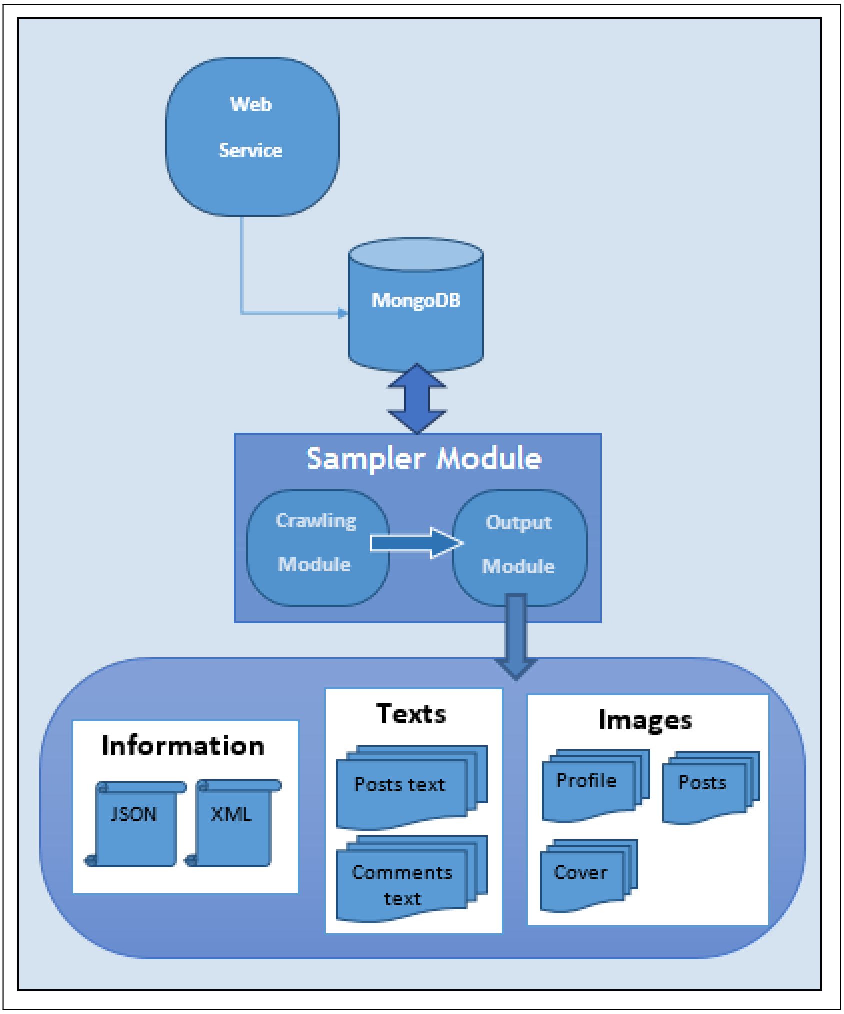


Figure 3.9: The architecture of SMC

[5] https://developers.facebook.com/docs/reference/php/

3.3.3 Input Module

As previously described, the Input module allows users to register entities to crawl. The module provides a Web service, in accordance with CAPER specifications. The web service allows the invocation of *registration* operations:

- RegistryAuth;
- SearchEntityAuth.

The registryAuth operation performs the input for a single entity (page, group, event or user). The Facebook APIs allow SMC to capture all the source's post without any temporal limitation. To avoid the presence of very old posts, the SMC allows the specification of a numeric parameter that represents the numbers of days for crawling posts.

The searchEntityAuth operation performs the search inside Facebook. For this, the user must specify at least a search term. Some terms can produce different entities, so the specification of the maximum number of entities to crawl is necessary. Like the registryAuth, the searchEntityAuth requires the number of days for crawling posts. Commonly the registration operations need the specification of the entity's type (page, group, event, and user).

Unfortunately, the Facebook Search APIs have a high rate of false positives. For this, for some keywords, the precision of SMC is not high.

For example, if you search *Messi*, it is possible to find the pages of:

- Football player;
- Italian city of Messina;
- country Mexico.

To overcome this limitation, the end user must specify more keywords.

The capture operations can be time-consuming. In order to maintain consistency, the *Sampler* changes different statuses. We will explain these statuses in the next paragraph. The web service expose the operation *check* to check the current status of Sampler.

When the end user registers an entity, the status is *registry*, this status becomes *starting* or *stopping* using the operations *start* and *stop*. During crawling, the associated status is *downloading* and when is finished the status is *complete* or *error*. The web service returns an *empty search* status if a search operation does not provide any results.

3.3.4 Sampler module

The Sampler module is the core of SMC and it performs all capture operations. At the time of writing, SMC performs two kinds of capture operations:

- Capture of single Pages/Groups/Events/User;
- Search and Capture of Pages/Groups/Events by using keyword(s).

The Sampler module is a Finite-state machine that takes the entity from the list of sources and starts the required capture operation.

The action of crawling an entity from Facebook is a long operation and it can require several days, in some cases. For this reason, the Sampler is a server that every M minutes starts, takes an entity to crawl and get data. To avoid overloading the physical server, the Sampler module checks if there are less of N parallel servers (where N is a global parameter).

Figure 3.10 shows the flow of different states on Sampler module.

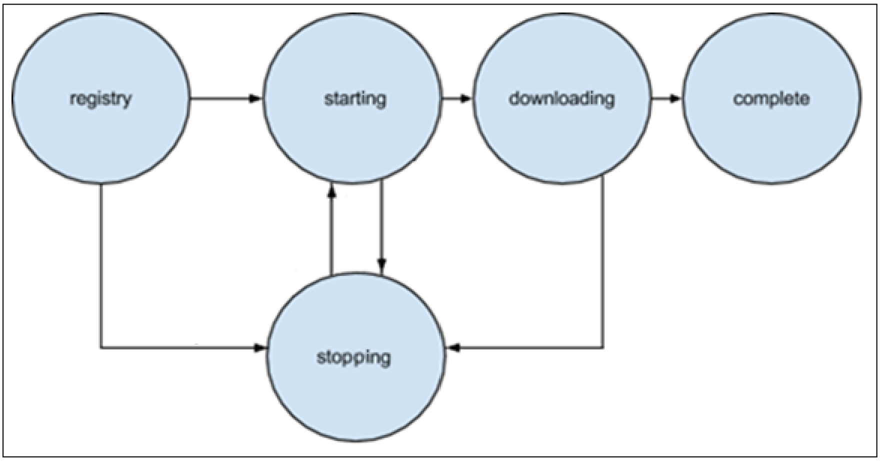


Figure 3.10: Flow of operation of sampler module

Therefore, the Sampler Module takes an entity in *starting* state, after the entity pass in *downloading* state. From this state, the entity can pass to the *stopping* state using the *stop* operation by the web service or in *complete* state when the module finishes.

For simplicity, in Fig 3.10 only the *complete* status is inserted, but in case of issues during crawling the status is *error* or *empty search* if, during the search of keywords, the query provides any results.

The Sampler uses a greedy algorithms in order to retrieve as much data as possible. The module stores corpora and metadata in a MongoDB collection and images in the file system. The type of information captured depends on the source. In fact, the Page's data are different from the Group's data.

For each type of entity, the data captured is:

Page: Information, Entity's Picture, Entity's Cover, and Feed;
Group: Information, Entity's Picture, Feed, and Members of the group;
Event: Information, Entity's Picture, Feed, and Invited people;
User: Information, Entity's Picture, Entity's Cover, Feed, and Friends.

One problem, that the SMC overcomes, is pagination. In fact, a group can have more than 1000 members. The Facebook API returns the first 25 users (by default). For this reason, the SMC recalls the same Web API adding the parameter of last retrieve data using a mechanism called pagination[6]. For each post, the SMC captures:

- Likes of a post;
- Comments on a post;
- Mentions on a post;
- Likes on a comment;
- Comments on a comment;
- Mentions on a comment.

The algorithm for capturing the data from a single source is:

```
Get the source
Capture all source's information
Obtain all source's posts
For each post:
    Capture information of post
    Take number of post's likes
    Capture users that liked the post
    Take number of post's comments
    Capture users that commented on the post
    For each comment:
        Capture information of comment
        Capture users that liked the comment
        Capture users that have commented the comment
```

Algorithm 1: Capture of Facebook source by ID

[6] https://developers.facebook.com/docs/graph-api/using-graph-api/v2.2

The search operation is very similar to the previous one. The main difference is that the first step is a search call to obtain the list of entities. For the search operation, the previous algorithms become:

```
Get the terms to search
Start with search
Obtain the list of entities to capture
For every entity :
    Capture all source's information
    Obtain all source's posts
    For each post:
        Capture information of post
        Take number of post's likes
        Capture users that liked the post
        Take number of post's comments
        Capture users that commented on the post
        For each comment:
            Capture information of comment
            Capture users that liked the comment
            Capture users that commented on the comment
```

Algorithm 2: Capture of Facebook source by Keywords

As introduced in paragraph 2.4.1, the Facebook APIs require the use of an access token. Table 3.5 shows the type of access token necessary for capturing a source.

Type of Entity	Type of Access Token for Id	Type of Access Token for search
Page	Application	Application
Event	Application or User	User
Group	Application or User	User
User	Application or User	Not Available

Table 3.5: Access Token necessary for every Facebook's entity

In the table, the label *Application or User* means that the source can be public or private. The Application Access Token is necessary for the first case. In the second case, a User Access Token is necessary.

3.3.5 Output Module

The data for each source is stored in a MongoDB collection and in the file system (for images). In accordance with CAPER specifications, the data must be transformed in XML format, pre-processed and finally upload over Original and Normalize CAPER Repositories. These Repositories are two instances of MongoDB with the following purposes:

Original: Store JSON data of a Source
Normalize: Store XML data and Images of a Source

The output module performs the operations:

1. Prepare the information in two different formats (JSON and XML);
2. Convert the gif image into jpeg;
3. Upload corpora and images in accordance with CAPER specifications.

In particular the upload specifications are:

- Upload JSON file over Caper Original Repository;
- Upload XML file over Caper Normalize Repository;
- Filter the image with size lower than 10 KB because they are not valuable for the project;
- Upload all image over Normalize Caper Repository;
- Upload post's message over Normalizer Caper Repository;
- Upload each post's comments over Normalizer Caper Repository.

3.4 Data Capture for Tourism domain

Another field in Social Media Analytics is the exploitation of user's reviews in tourist places for marketing goals. In fact, as described in the previous chapter, some Social Media platforms, like Booking.com[7] and TripAdvisor[8] , are billionaire businesses. This particular domain is similar to Web Reputation with the difference that the entities are physical places.

In details, we focus on places that exist in a specific geographic area. Moreover, the places belong to a particular category (Hotel, Restaurant, etc.). We consider these two requirements to answer a question like “What is the best hotel in Rome?”. Obviously, this is the core business of TripAdvisor, but other Social Media

[7] http://www.booking.com
[8] http://www.TripAdvisor.it/

platforms provide similar features. In our study, we take the places' information and reviews from different platforms. Moreover, we provide a mechanism for integrating the data from different sources.

We consider four Social Media platforms: Facebook, Foursquare, Google Places, and Booking.com. We chose these platforms both for their popularity and the availability of data.

Foursquare[9] is a location-based social network for mobile devices. Users can *checkin* a place (called venue) using the Foursquare's mobile application. This application provides a list of venues located near the user. Location is largely based on GPS. Foursquare provides a mechanism that allows user to leave a review (called tip) to venues. A review can express appreciation, criticism or a suggestion.

As previosly described, Facebook[10] is the biggest Social Network in the world. In addition to the communication's features, Facebook has introduced different mechanisms for Tourist Places. In detail, on Facebook there is a type of entity called *place*. A place has the same layout as a normal Facebook Page, but it has the geographical information on where the place is located in the world. Moreover, Facebook provides features such as *checkins*, reviews and category.

Google Places[11] is the location service of Google. In this Social Media are stored over 95 million of companies and points of interested. Moreover, the information about a place can be added only by the its owner. In fact, it is necessary information like: phone number, address and website. Each Place has both geographic information (latitude/longitude coordinates) and a category (Hotel, restaurant, etc.).

Booking.com[12] is an online booking Social Media founded at Enschede in 1996. This platform offers the booking service for over 550,000 accommodation structures worldwide.

The main difference between Google Places, Facebook, Foursquare, and Booking.com consists in the user's behaviours. In fact, on Facebook and Foursquare any common user can generate a page about a business activity, although he/she is not the activity owner. In Google Places only the business owner can register a page about the activity. Booking.com is a commercial product, so it is the most complete, but the platform lacks a set of free Web APIs. Due to this fact, the crawler module extracts the information directly from the web page.

[9] http://www.foursquare.com/

[10] http://www.facebook.com

[11] https://plus.google.com/u/0/local

[12] http://www.booking.com

3.4.1 Tour-pedia Architecture

Figure 3.11 illustrates the Tour-pedia architecture. The Data Extraction module consists of four ad-hoc crawlers, which extract data from Facebook, Foursquare, Google Places, and Booking.com. We choose these Social Media first because they are very popular and secondly because they provide an easy way to extract data.

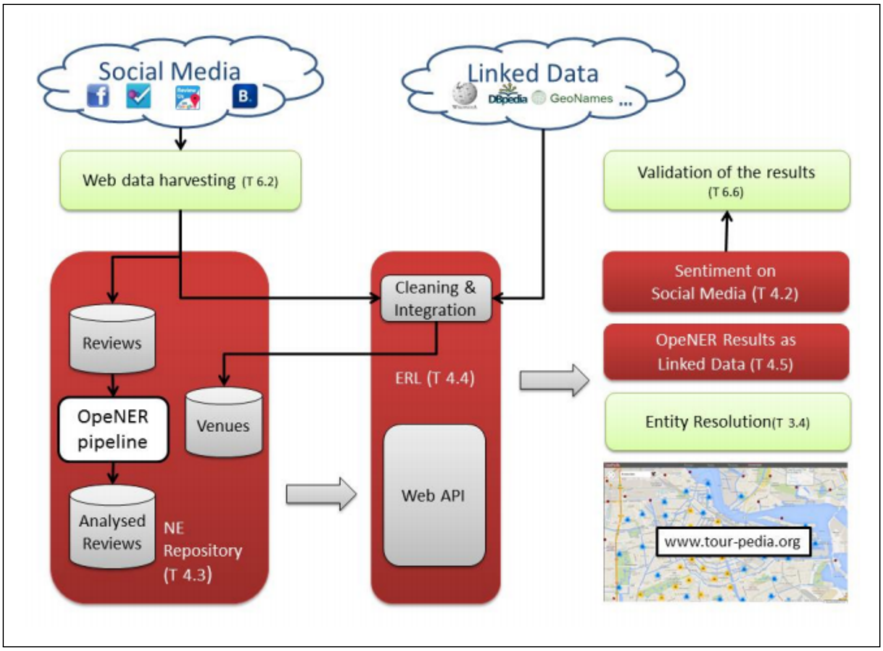


Figure 3.11: The architecture of Tour-pedia

The Named Entity repository contains two main datasets, which belong to the specific domain of tourism: Places and Reviews. The dataset of Places contains more than 500.000 places in Europe divided in four categories: accommodations, restaurants, points of interest, and attractions[13]. At the time of writing, Tour-pedia covers several cities: Amsterdam, Barcelona, Berlin, Dubai, London, Paris and Rome.

The places were elaborated and integrated through the Data Integration module to build a unique database. A merging algorithm, based on distance and string

[13] http://tour-pedia.org/about/statistics.html

similarity, performs the Data Integration of the same place on different Social Media platforms. This algorithm will be presented in paragraph 4.3.1. The dataset of Reviews contains over 600.000 reviews. Reviews were analysed through the OpeNER pipeline to extract their sentiment.

3.4.2 Tour-pedia Data Acquisition

Figure 3.12 shows the conceptual schema behind the crawlers. For each Social Media there is present a Data Extractor module that collects data. The Data Extractors of Facebook, Google Places, and Foursquare exploit the Web APIs that each Social Media provides. Instead, Data Extractor for Booking.com is a scraper that extracts information from each accommodation page using the XPath language[14].

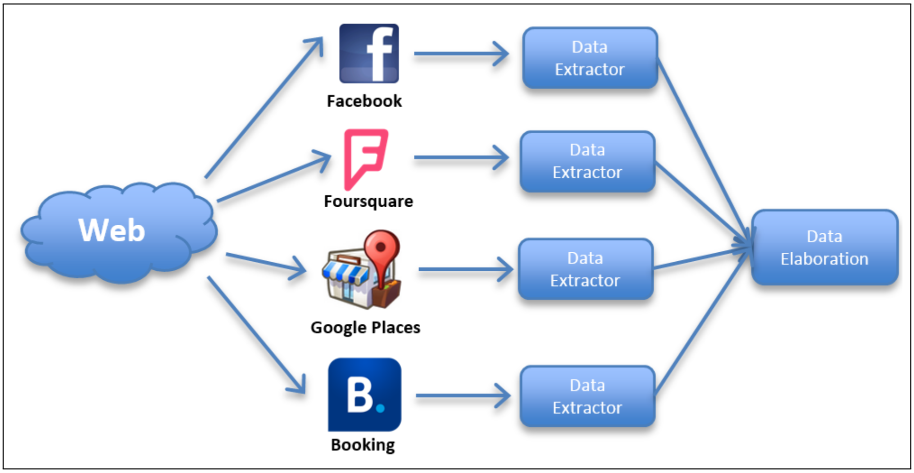


Figure 3.12: The architecture of Tour-pedia Data Acquisition

Each Data Extractor is conceptually different with respect to SocialTrends or SMC. The aim of this study is to evaluate the sentiment about a place in a geographic area. Therefore, we can define the first big difference in this crawler: the geographic information. In fact, the Data Extractor module uses some parameters that represent an area in the world (a city name, a geographic coordinate, etc).

[14] http://www.w3.org/TR/xpath20/

The second difference is that a place can belong to different categories. In fact, some places are useful for this project (restaurant, hotel, bar, etc) and others, such as Universities, are useless.

For this reason, every data extractor can derive the place's category from the raw data. The details are on the deliverable [2] and [3]. For the purposes of the project, the categories are the follows: accommodation, attraction, point of interest, and restaurant. The next two paragraphs will introduce our solution to the problems of coordinate and category discovery.

3.4.3 Coordinate discovery

The problem of finding all the places and reviews of a given location has different solutions. Our way takes into account the features of geographical search APIs of each Social Media.

On Booking.com, the data extractor is a scraper that uses an HTTP connection for capturing the Place's Page. The search of all places for a certain city or a region is one of Booking.com features. The scraper starts from the first response page of the search query and then extracts all information. Often, all the accommodations of a city can be divided into different response pages. In this case, the scraper extracts the information from all the web pages.

For Facebook and Foursquare, we use the search APIs that allow geographic search using latitude, longitude, and radius (see Figure 3.13).

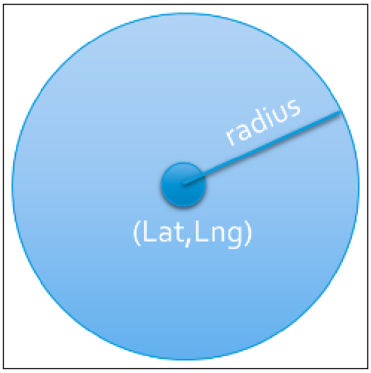


Figure 3.13: Crawling area given latitude, longitude, and radius

Obviously, the zone, which covers a city, is not a regular circle, but has irregular boundaries (see Figure 3.14).

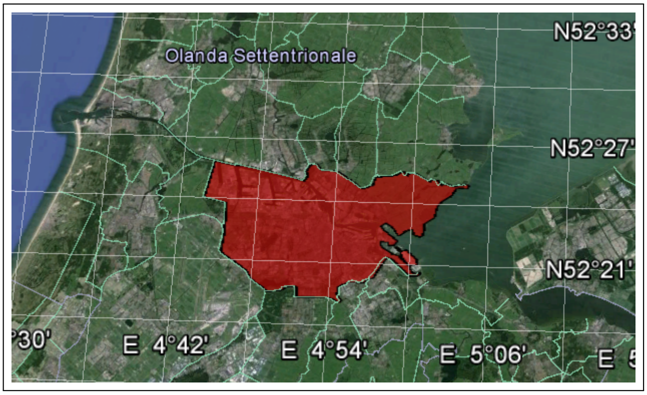


Figure 3.14: Geographic zone of Amsterdam (KML from www.gadm.org)

For this reason, we implement an algorithm that generates a list of coordinates with the information of latitude and longitude from the KML that specifies that geographic zone. The software module that performs this task, follows the steps:

1. Looking for a bounding box of the geographic zone from www.gadm.org;
2. Start generation of coordinates near the zone using a given step-distance (increment) between coordinates;
3. Apply the Jordan Curve theorem [20] in order to see if a coordinate pair falls inside the area;
4. Save the pair if it is inside the geographic area.

The result of this task is the list of coordinates, with fixed radius of 1 km. The coordinates cover the whole zone (see Figure 3.15).

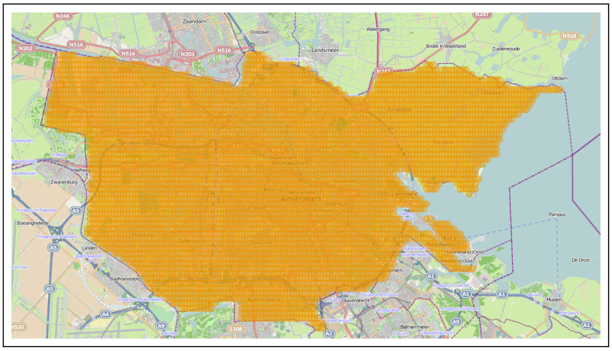


Figure 3.15: Coordinates coverage within the geographic area

A specification for Google Places is necessary. For this Social Media, we use a more efficient algorithm that takes into account the Google Places search limitations. In fact, every search call respond with a maximum of 25 elements. Therefore, if there are more than 25 places, we cannot capture them. In this case, we must use a smaller radius.

The implemented algorithm is recursive. It starts from a geographical area d, the area is considered as a circle, with a central point and a radius.

The central point is described by two coordinates, latitude and longitude, while the radius is a number. In this exemplification, the circle circumscribes the area. Then, the algorithm divides this circle in four parts. For each part, a search call is performed. If the response has less than 25 places, the part is correctly captured, otherwise the part is divided into four parts and is recursively analysed. This algorithm produces circles with different radius and has the advantage that it produces fewer circles that the one for Facebook and Fousquare. Figure 3.16 shows an example of how the algorithms works.

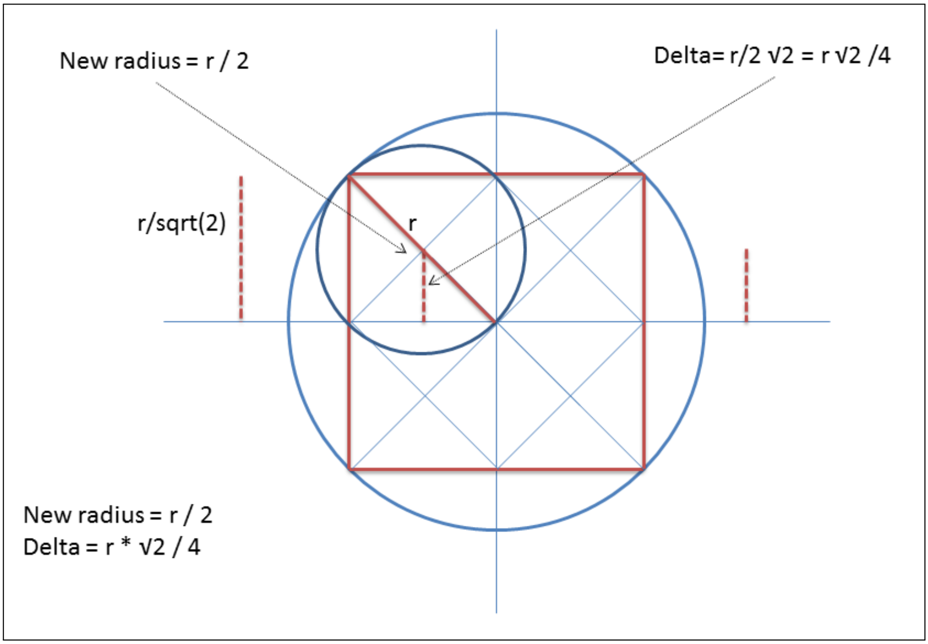


Figure 3.16: Google Place coordinates discoverer

3.4.4 Category discovery

At the time of writing, Tour-pedia contains data on seven cities (Amsterdam, Barcelona, Berlin, Dubai, London, Paris, and Rome). Every place belongs to different categories (Hotel, Restaurant, etc.). Obviously, the end user would compare place that belong to the same category. Unfortunately, every Social Media platform has its own internal classification and this can generate confusion in the end user. For this reason, the places inside Tour-pedia are divided into four categories: accommodations, restaurants, points of interest, and attractions.

In table 3.6, we provide a definition of each category.

Category	Definition
accommodation	place where it is possible to sleep
restaurant	place where it is possible to eat and drink
point of interest	place where people can have public services such as airport, railway station, etc
attraction	place of entertainment, both for cultural purposes (museum, theatre, cinema), and for sport or recreation (night club, swimming pool, golf, gym)

Table 3.6: Description of Tour-pedia Categories

The categorization of places in Booking.com is very simple, because all the places are hotel. For others Social Media, this task is more difficult. In fact, every Social Media provides an internal classification of Places.

Usually the end user chooses the category of a place. Due to this fact, many places belong to different categories. For example, a Hotel can be classified as accommodation for some users and as a restaurant for others. For this, we decided that a place can belong to only one category.

We choose an order of importance of categories in this way every place assumes only the most important category. We imposed the following category priority: accommodation, restaurant, point of interest, and attraction.

To perform this task, we start from the Social Media categorization. Then, we create a table of Social Media categories and Tour-pedia categories 3.7.

TourPedia Categories	Facebook Categories	Foursquare Categories	Google Places Categories
accommodation	Hotel Bed & Breakfast	Hotel Bed & Breakfast Boarding House Hostel Hotel Pool Motel Resort Roof Deck	lodging campground
restaurant	Fast Food Restaurant Italian Restaurant Restaurant Bar Caffè Seafood Restaurant Wine Bar Beer Garden	Food	bar cafe food restaurant
attraction	Arts & Entertainment Historical Place Landmark Pub Night Club Dance Club Auditorium Theatre Monument Museum/Art Gallery Museum Public Places & Attractions Church	Arts and Entertainment Nightlife Sport Outdoors and Recreation Spiritual Center Auditorium	Amusement park aquarium art gallery casino church hindu temple movie theatre mosque museum night club park spa stadium synagogue zoo

poi	Sports Center College & University Fitness Center Pool & Billiards Piscina Club Community Center Airport Terminal Airport Shuttle Airport Lounge Dentist Dermatologist Event Venue Government Organization Home Improvement Neighborhood Convention Center	Airport Bike Rental Bike Share Bus Station Embassy / Consulate Ferry General Travel Light Rail Moving Target Rental Car Location Rest Area Road Subway Taxi Tourist Information Center Train Station Travel Lounge Convention Centre Government building Library Medical Centre Parking	airport bus station city hall cemetery courthouse embassy gym health hospital library local government office parking pharmacy physiotherapist place of worship police post office park school subway station train station taxi stand university veterinary care

Table 3.7: Mapping of Social Media's categories inside Tour-pedia

The Data Extractor looking for the category for every place, the steps are in the algorithm below.

```
Read the place categories
For each category:
    if category is allowed
        save the place
    else
        discard place
```

Algorithm 3: Category Match

3.4.5 General details about the Data Extractors

The four Data Extractors have different implementation, but they have similar architecture. Moreover, the modules are developed in PHP and they store the raw JSON data. The Data Extractor for Booking.com is an exception, because it is a scraper and it uses XPath query. Every Data Extractor stores the data in a MongoDB database called with the name of Social Media captured. In particular, the places are stored in a collection called **[SocialMedia]_places** and the reviews in a collection called **[SocialMedia]_reviews**, where [SocialMedia] is the platform. The Table 3.8 shows the mapping from the module of generic architecture and the Data Extractor.

General Architecture	Data Extractors
Input Module	Coordinates/City List
Scheduler	Crawler
Social Media Crawler	
Preparation Module	
Web Crawler Module	NOT PRESENT
New Entity Detector Module	NOT PRESENT
Analyser Module	Social Analyser Module

Table 3.8: Mapping of modules of the proposed Architecture and the Data Extractors

Like the SMC, in this implementation as well the scheduler and the Social Media Crawler are merged in a module called Sampler. Moreover, there are no Modules: Web Crawler, Preparation Module, and New Entity Detector. Every Crawler has small differences with respect to the others, the details will be illustrated in the next paragraphs. The result of the crawling phase is showed in Table 3.9

Social Media	Dubai	London	Paris	Berlin	Amsterdam	Rome	Barcelona
Facebook	1052	4893	832	2084	583	4465	455
Foursquare	14469	47148	6545	21875	7735	16913	6339
Google Places	7301	121723	51665	39765	13635	31455	18499
Booking.com	440	1489	2035	1114	729	2668	1602

Table 3.9: Number of places captured from January to March 2014

3.4.6 Logical architecture of Foursquare Data Extractor

The logical architecture of the Foursquare Data Extractor is in Figure 3.17. The Places crawler reads the list of coordinates. Two adjacent coordinates always have the same distance (that represent the radius of capturing). The places are stored on MongoDB. The Places crawler performs the filtering of categories. Moreover, the Reviews Crawler reads the unique identification of places and it captures all the reviews. The details about Foursquare's API are available at this URL[15].

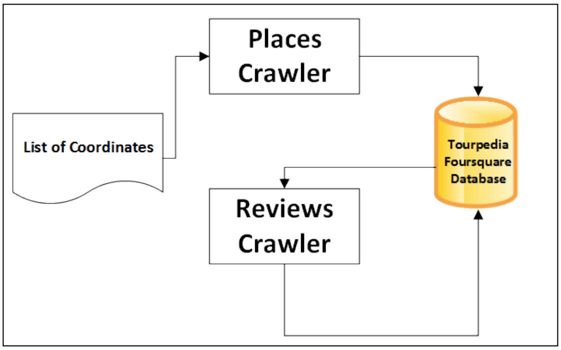


Figure 3.17: Logical architecture of Foursquare Data Extractor

Figure 3.18 shows the places captured from Foursquare Data Extractor in Amsterdam.

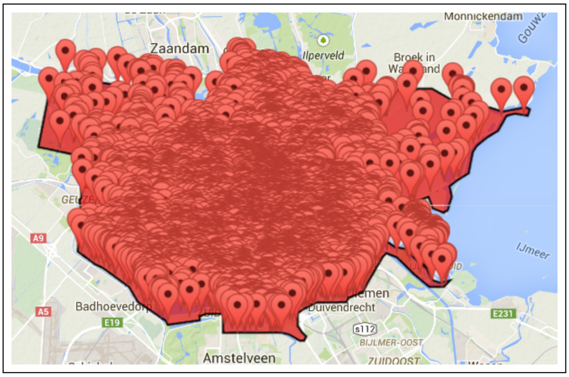


Figure 3.18: Places Captured from Foursquare Data Extractor in Amsterdam

[15] https://developer.foursquare.com/

3.4.7 Logical architecture of Facebook Data Extractor

This implementation is different from the previous one, due to the presence of a single crawler that capture both the places and reviews. This module uses the coordinates and it saves places and reviews on MongoDB. The Data Extractor uses the graph API (for search places from a coordinate and for reviews) and FQL (for reviews details). Moreover, the module performs the category filtering. Figure 3.19 shows the logical architecture of the Facebook Data Extractor.

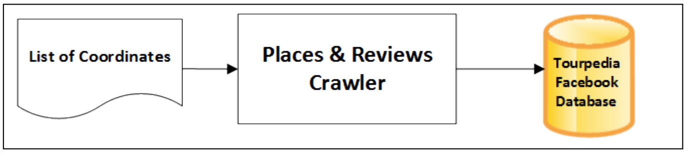


Figure 3.19: Logical architecture of Facebook Data Extractor

The Figure 3.20 shows the places captured from Facebook Data Extractor in Amsterdam.

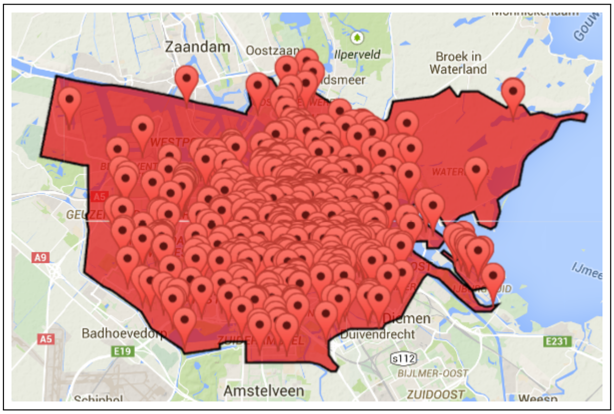


Figure 3.20: Places Captured from Facebook Data Extractor in Amsterdam

3.4.8 Logical architecture of Google Places Data Extractor

The data extractor captures both the places and reviews and it performs the category filtering. Figure 3.21 shows the logical architecture of Google Places Data Extractor. Details about Google Places APIs are available at this URL[16].

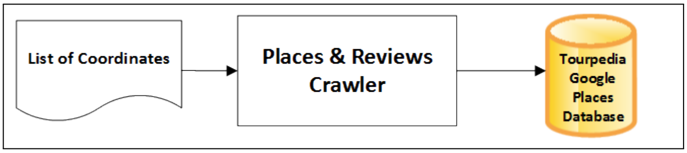


Figure 3.21: Logical architecture of Google Places Data Extractor

Figure 3.22 shows the places captured from Google Places Data Extractor in Amsterdam.

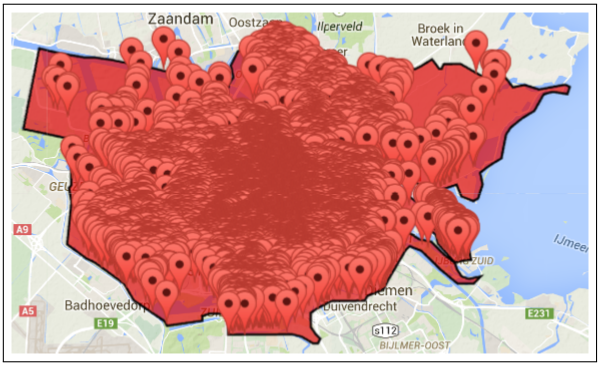


Figure 3.22: Places Captured from Google Places Data Extractor in Amsterdam

3.4.9 Logical architecture of Booking.com Data Extractor

As previously describe, this module does not use Web API, but it scrapes the hotel's page and extracts all the information about the page and reviews. This module uses the XPath queries for extraction. Figure 3.23 shows the logical architecture of the Booking.com Data Extractor.

[16] https://developers.google.com/places/documentation/

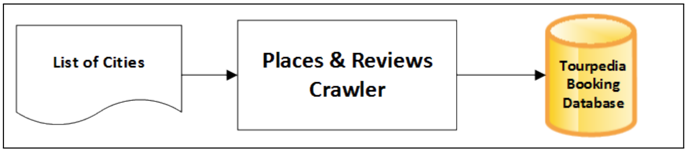


Figure 3.23: Logical architecture of Booking.com Data Extractor

Booking.com provides a Web API platform, but it is available only for affiliate partners, more detail are at this URL[17]

The Figure 3.24 shows the places captured from Booking.com Data Extractor in Amsterdam.

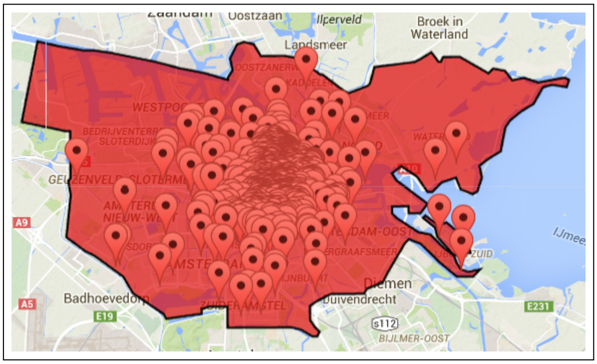


Figure 3.24: Places Captured from Booking.com Data Extractor in Amsterdam

[17] https://www.bookingsync.com/en/documentation/api

4

Social Media Data Analysis

In the previous chapter, we introduced a generic architecture and its partial implementation to capture data in three different domains of investigation. Likewise, in this chapter, we will explain some particular method of analysis of Social Media data.

4.1 SocialTrends: Monitoring Online Reputation

SocialTrends analyses data collected from Facebook, Twitter, and YouTube. The datasets are captured from Facebook pages, Twitter accounts and YouTube channels. In particular, we captured 304 entities, organized in 14 categories. In these paragraphs, we will present the result of metrics from April 20, 2012 to May 20, 2012. As introduced on the paragraph 3.2, for each entity, we monitor the metrics of popularity, activity, and influence day by day. In this way, we can calculate the increment of each metrics.

Increments belong to two categories: a) percentage increase and b) absolute increase. The percentage increase and the absolute increase of a metric M indicates how much the value of the metric is calculated in percentage and in absolute, respectively. They are calculated as follows: assume that mt and m_{t+1} are the values that a metric M assumes at time t and t +1 respectively for a given entity. The percentage increase PI is defined through the following formula:

$$PI = \frac{m_{t+1} - m_t}{m_t} \tag{4.1}$$

while the absolute increase AI is calculated through the following formula:

$$AI = m_{t+1} - m_t \tag{4.2}$$

The strength of SocialTrends resides in the comparison of entities that belong to the same category. In particular, the match of entities is achieved through visual rankings, grouped in accordance with the metrics and to the channel.

For each entity, it shows the distance to the other entities of the same category. By distance between entity i and entity j we denote the difference between the value of the metrics of i and that of j.

SocialTrends can efficiently monitor entities from three different point of views. For now, we will give some example for Popularity in the year 2012. Figure 4.1 shows the classification of the most popular American CEOs on Facebook on May 2012. It is interesting that in this classification the people belong to the ICT field.

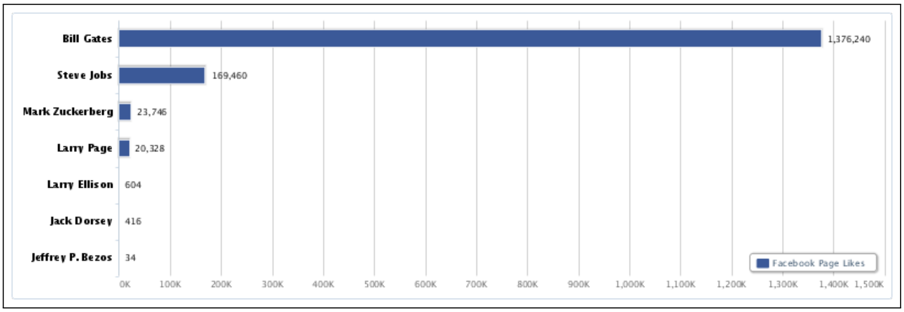


Figure 4.1: American CEOs' popularity on Facebook

Figures 4.3 and 4.2 show the absolute increase and the percentage increase of popularity of the CEOs on Facebook. As introduced in the previous chapter, on Facebook the increase in popularity corresponds with an increase in the number of fans.

It is interesting to note that the histograms give a visual idea of the distance between two entities. Furthermore, Mark Zuckerberg, who is not the most popular CEO on Facebook, has a percentage increase with a peak of 9.10

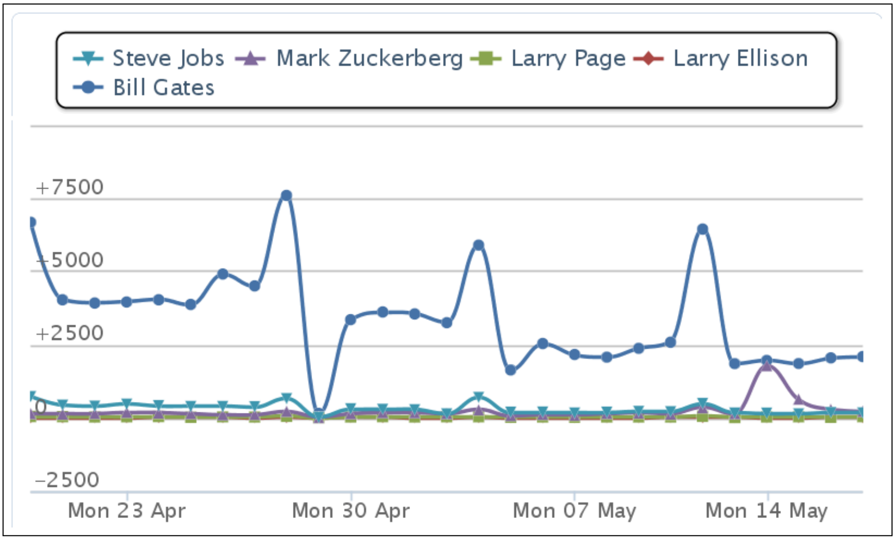


Figure 4.2: Absolute increase of American CEOs' popularity on Facebook

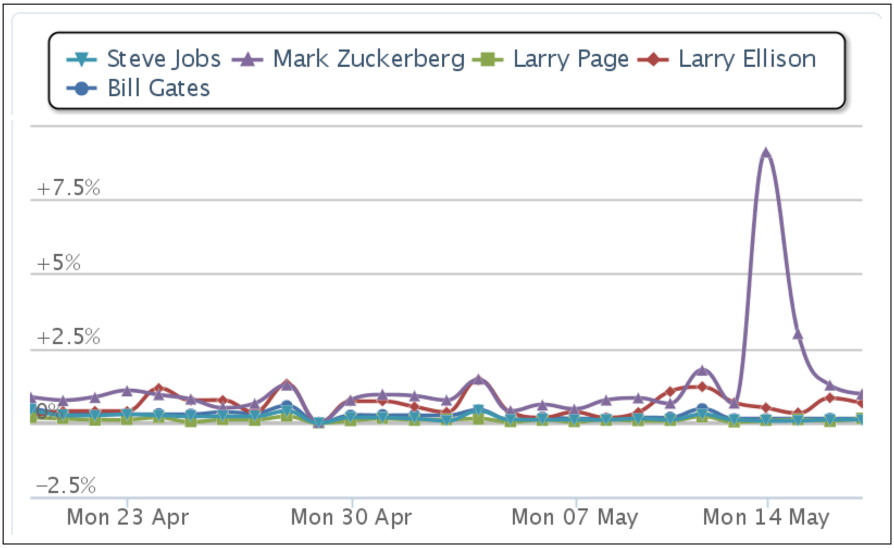


Figure 4.3: Percentage increase of American CEOs' popularity on Facebook

Figure 4.4 shows the classification of the most popular Italian politicians on Twitter in May 2012.

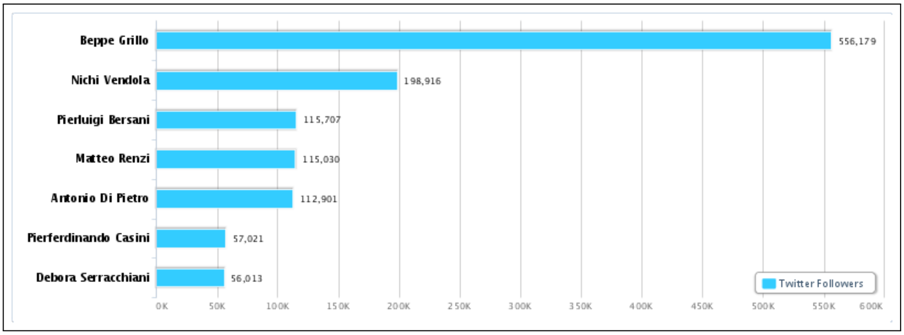


Figure 4.4: Italian politicians' popularity from Twitter.

Figures 4.5 and 4.6 show the percentage increase and the absolute increase in popularity of the most popular Italian politicians on Twitter.

Sometimes, the peaks are the results of real events. For example, we note that Beppe Grillo is the most popular politician with 553.166 followers, followed by Nichi Vendola, with 198.117 and then all the other politicians.

Beppe Grillo's percentage increase presents some peaks in correspondence to April 28th, May 7th, 8th and 12th of 0.42%, 0.64%, 0.69%, and 0.50%, respectively.

All these peaks correspond to real events that occurred. On April 28th and May 4th 2012, Beppe Grillo presented his political party, called *Movimento Cinque Stelle* in Sarego and then in Milan. On May 7th 2012, there were political elections for Mayor in some cities, while on May 8th 2012, Beppe Grillo criticized the President of the Italian Republic on his blog. Then on May 12th 2012, the Time Magazine talked about Beppe Grillo. The percentage increase directly maps to the absolute increase, as shown in Figure 4.6.

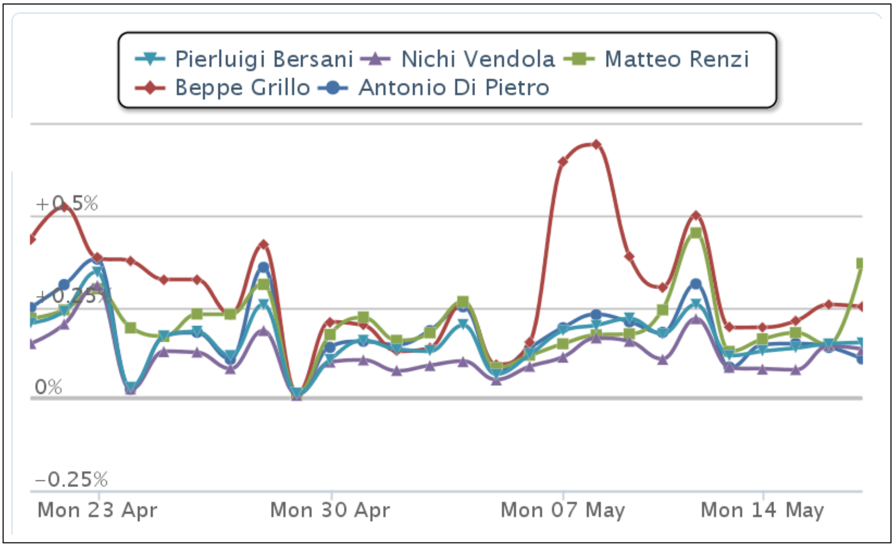


Figure 4.5: Percentage increase of Italian politicians' popularity on Twitter

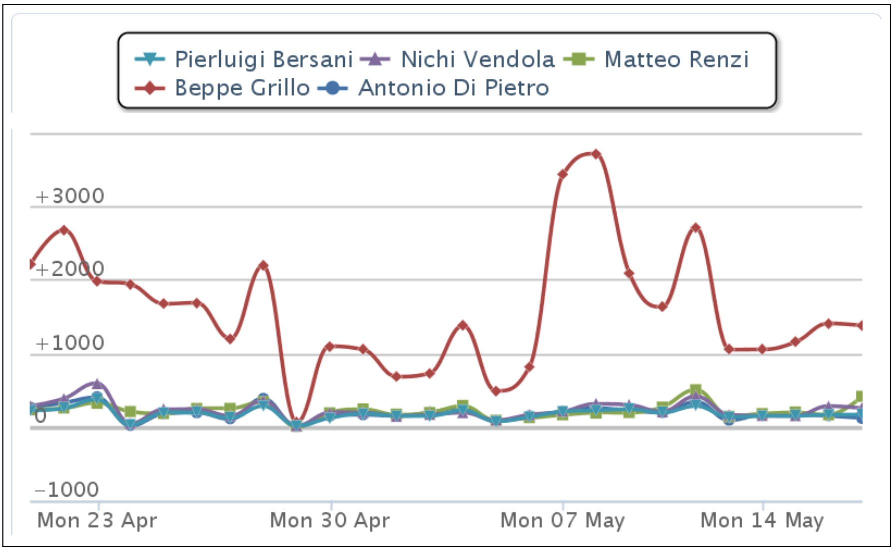


Figure 4.6: Absolute increase in Italian politicians' popularity from Twitter

4.2 Analysis of User Interaction

Within the project CAPER, we develop the SMA (Social Media Analyser) a tool able to analyse a Facebook Page, Group, Event or User to create a weighted interactions graph. The results of this work were previously introduced in [70].

Our work is based on the concept of *Social interactions*, all the interactions that are not regulated by the price mechanism [86]. The interaction graph, introduced by [99], represents a subset of the social graph parametrized by two parameters n and t. The n-parameter is the minimum number of interaction events, while the t-parameter represents a time window during which interactions must have occurred.

In [70], the concept of *Interaction network* is introduced. This is a graph built with users' interactions in public spaces of Social Media. Examples of interactions on Facebook are the likes or comments on a post. The exploitation of public spaces of Social Media eliminates the limitations due to the privacy and ethical reasons. Moreover, we do not use the friendship relation for two reasons. First, in the paper [99], the authors claim that the friendship relation online does not guarantee any implication of real interaction. Second, this relation depends on the user's privacy settings.

Algorithm 4 details how we compute the interactions between users. We started from the list of different entities (for example the Facebook's pages or groups). Next, we considered the list of all users involved in the entity and the relative interactions.

For each interaction, we define the source user as one who performs the action, while the target user is involved in the action itself. We take into account two type of interaction: *single* and *multiple* or *double*.

An interaction is single if only two users are involved. It is multiple if it is possible to identify a single source user and different target users.

It is important to underline that this algorithm is independent from the Social Media and the entity considered. In this way, Algorithm 4 represents a conceptual model of analysis of interaction. This algorithm is entities and Social Media independent.

This feature is especially important for intelligence activities. In fact, LEAs involved in fighting organised crime, have to deal with groups of people that interact with each other exploiting, for instance, more than one Facebook page. For a complete study of Algorithm 4, we suggest [70].

```
for all entity e involved in the analysis do
    getEntityUsers
    getListOfInteractions
    getAllAction
    for all user u on the entity do
        for all action a performed by u do
            getRelatedInteraction i
            if i is single then
                get the target user q
                set a link u −> q
            else
                get other user j involved in a
                for all other user j do
                    set the link u <−> q
                end for
            end if
        end for
    end for
end for
```

Algorithm 4: Determine users interactions

4.2.1 The Analysis of Facebook interactions

Our study case is the implementation of Algorithms 4 on Facebook. In this platform, users can interact in different ways (e.g., messaging, applications, photo uploads, chat, etc). On a Facebook Page, the users interact in different ways, for examples with the action of like/comment to post.

We call these interactions respectively *like a post* and *comment a post*. Instead on a blog the main action is *comment a post* or, on WordPress, *like a post*.

For our analysis, we distinguish two objects involved in interactions: posts and comments. A post is an UGC, published by the administrator (the user who manages the page) or by a generic user of the page, which can contain text, photos, video or link. A comment is similar to a post but with the difference that it is a reply to a post. Users can interact with an object through actions such as like, comment, mention, and share. Unfortunately, in our analysis, we do not include the share action. This due to the fact that Facebook provides, via Web APIs, only the total number of shares on a post and not the information of who shares a post. Every action needs a different treatment due to the associated metadata. The user's like on a post is an easy way to let someone know that you enjoy it, without leaving a comment. A comment is a more explicit way to interact with the author of the post.

A mention is a particular type of tag that is performed on a post or a comment. The mention allow users to create a link between a user and a content (post or comment). As an example, Figure 4.7 shows a Facebook page with related objects and actions.

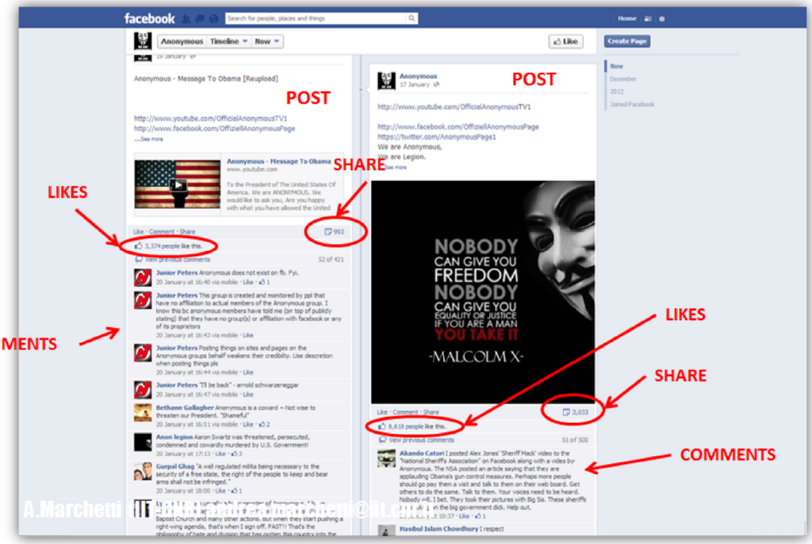

facebook
POST
POST
SHARE
LIKES
LIKES
SHARE
COMMENTS
COMMENTS
NOBODY CAN GIVE YOU FREEDOM NOBODY CAN GIVE YOU EQUALITY OR JUSTICE IF YOU ARE A MAN YOU TAKE IT
-MALCOLM X-
A.Marchetti

Figure 4.7: Example of possible action on a Facebook Page

Table 4.1 summarizes all the possible actions on a Facebook page.

Action	Post	Comment
Like	like a post	like a comment
Comment	comment on a post	comment on a comment
Mention	mention in a post	mention in a comment

Table 4.1: Actions available on posts and comments

From the actions on posts and comments, we described all the interactions between two users (see the table 4.2).

Description	Type	FB object	Action
u likes a post published by q	single	Post	like
u comments a post published by q	single	Post	comment
u is mentioned by q in her post	single	Post	mention
u and q comment the same post	double	Post	comment
u and q are mentioned in the same post	double	Post	mention
u likes a comment published by q	single	Comment	like
u comments a comment published by q	single	Comment	comment
u is mentioned by q in her comment	single	Comment	mention
u and q comment the same comment	double	Comment	comment
u and q are mentioned in the same comment	double	Comment	mention
u likes a comment on a post published by q	single	Post	like
u comment a comment on a post published by q	single	Post	like

Table 4.2: Possible interactions on a Facebook page

From the table, we can extract various considerations. The first is that the interactions are similar. In fact, for each interaction on a post, there is the dual on a comment. One key point is the type of interaction. In fact, there are *single* and *double* interactions. In the first case, there is a relationship between the first user, who creates the content, and the second user, who performs the action. For the interaction of the second type, there is a relation between a single source user and more than one target user. In terms of computation, the single interactions grow linearly instead, the double ones grow quadratically. In accordance with these considerations, the algorithm to build the interaction graph of a Facebook page is the following:

```
getPageUsers
getListOfPossibleInteractions
getAllActionOfThePage
for all user u on the page do
    for all action a performed by u do
        if a is monodirectional then
            get the target user q
            set a link u −> q
        else
            get other user j involved in a
            for all other user j do
                set the link u <−> q
            end for
        end if
    end for
end for
```

Algorithm 5: Building interactions graph of a Facebook page

The same algorithm can used for the analysis of a Facebook group or event or even a user.

4.2.2 Weighted Interactions Graph

In order to improve our study on the interactions graph on Social Media, we consider that different interactions are characterized by a different level of strength. For instance, we can considerer two users who interact on the same Facebook page. If user *A* mentions a user *B* in a comment on a post, both *A* and *B* know each other (maybe virtually) between A and B a strong direct interaction occurs.

The action *mention* is stronger than *like*. Due to this aspect, we built a weighting system for the interactions graph of a Facebook's entity. The theoretical reason of our analysis is that all the interactions between two generic users have different strength. In intelligence activity, some interactions are more interesting than others. Moreover, there are meaningless interactions that create noise in the output. The idea at the base of the weighting system is related to the noise concept and to the fact that rare information are more important than recurring information. The problem has been approached from two different perspectives: we conducted a frequency study of different interactions to determine the rarest.

For the frequency study, we used a pool of Facebook pages suggested by LEA involved in the CAPER project. To obtain these pages, LEA operators suggested some keywords that are meaningful for an intelligence analysis, the SMC capture the related page and posts.

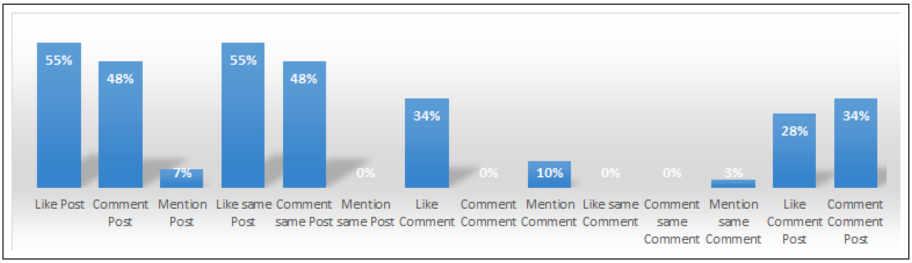


Figure 4.8: Percentage of interactions in the sample

In Figure 4.8, we reported the percentage of different types of interactions in our sample of Facebook pages. It is possible to notice that, comment_comment interactions, related to a situation in which a user *A* comments a comment published by another user *B*, is not present; this because the feature was not available at the time of the experiment. Then, we studied how much every single interaction was repeated in our dataset.

In Figure 4.9, we plotted the frequency of every single interaction. As shown in the graph, the most frequent interactions are those related to actions performed on the same Facebook object. For example, likes on the same post (like_same_post) and comments on the same post (comment_same_post) are the most frequent.

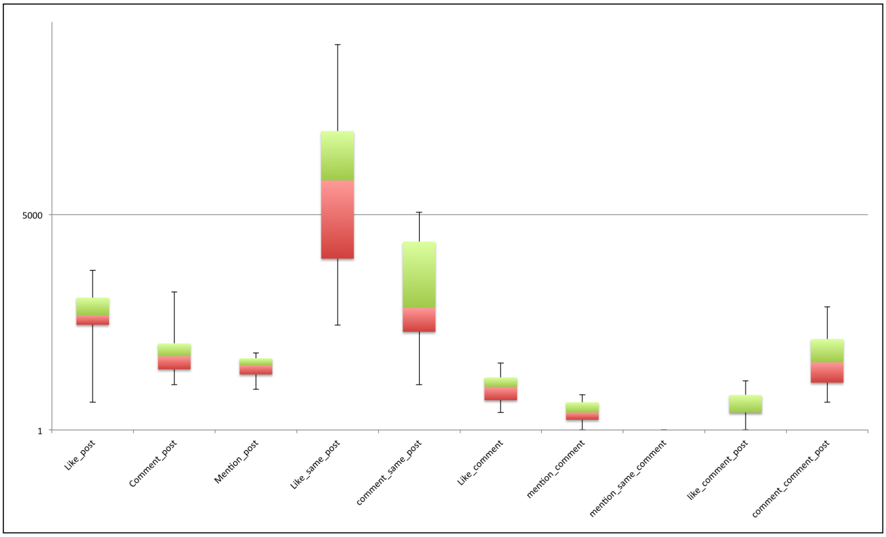


Figure 4.9: Distribution of the frequencies of different interactions

Moreover, these interactions are less significant due to the fact that a huge number of users can like or comment on the same post. In fact, these interactions do not add any important information because the probability that two users who like the same post are interacting is very low. In contrast, more direct interactions, like mention_comment or mention_post are less frequent [70]. To overcome this problem, we decided to weight each interaction considering two aspects:

- Mean frequency of every single interaction;
- Number of pages under analysis.

To allow a multi-entity analysis, we consider how many entities (pages, groups, and events) are analysed. Then for each entity, we considered the list of interactions. Next we count the frequency of every single interaction for each entity. Given $F_i(1)$ as the number of times in which the interaction *i* appears in entity 1, and considering a dataset composed of n entities, we have

$$F_{m_i} = \frac{F_i(1) + F_i(2) + ... + F_i(n)}{n} \tag{4.3}$$

We then calculate the sum of each contribution on the entity. Assuming that the interactions we consider are m:

$$F_tot = F_{m_1} + F_{m_2} + ... + F_{m_m} \tag{4.4}$$

Finally, we calculate the value of every single interaction, by using the following formula:

$$W_i = 1 - \frac{F_{m_i}}{Ftot} \tag{4.5}$$

Once the weight of every single interaction is calculated, we use this information in the interactions graph. To do this, “we summarize all the interactions occurring between each pairs of users and, consequently, we assign a weight to the aggregate interactions calculated as the sum of the weights of the all interactions” [70]. Algorithm 6 describes the procedure. The links between two users will represent all the interactions occurring during the observation period.

```
get a pair of users in a page
interactionWeight = 0
for all interactions between the pair do
    get the corresponding weight
    sum the weight to interactionWeight
end for
```

Algorithm 6: Determine aggregate interactions

The SMA provides a simple graphic interface based on the d3 library[1]. The interface shows the weighted graph and enables users to hide the page’s administrator. For the theoretically enormous number of interactions, the GUI shows only the top 1000 interactions. Figures 4.10 and 4.11 show the result of the analysis of the Anonymous Page crawled by SMC on date 02/12/2014 for 30 days. The analysis is performed using 4843 posts and 47206 comments Figure 4.10 shows the resulting graph with the administrator of the page.

[1] http://d3js.org/

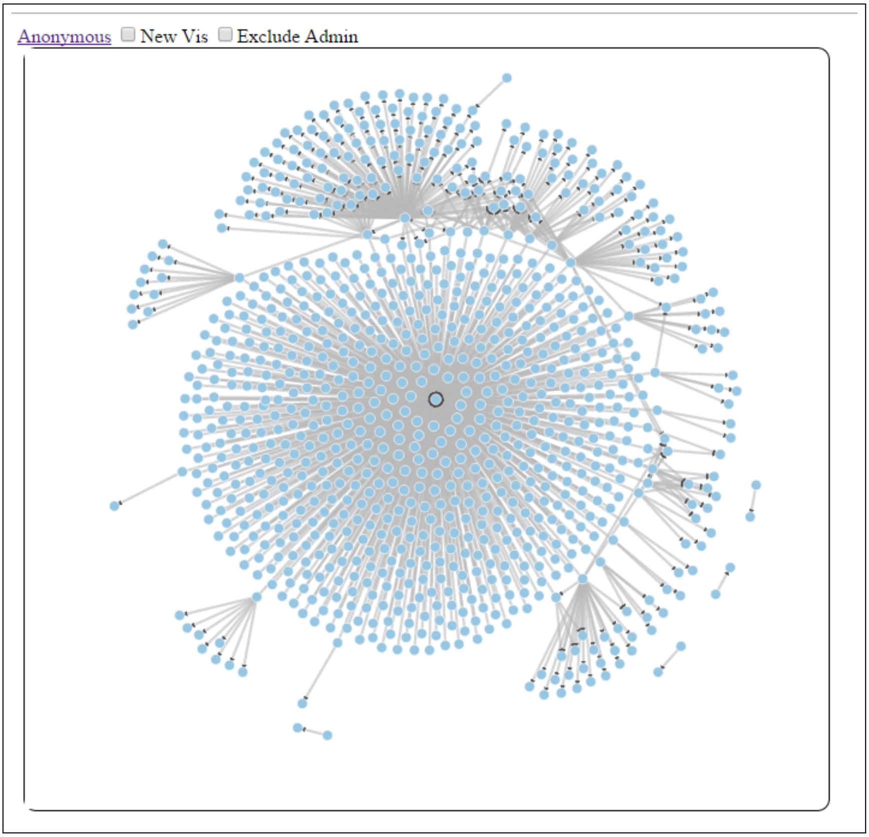


Figure 4.10: Visualization of Anonymous Graph with Admin

From Figure 4.10, we note that the administrator of the page is the most important node of the graph. Figure 4.11 shows the result of the graph without the administrator of the page.

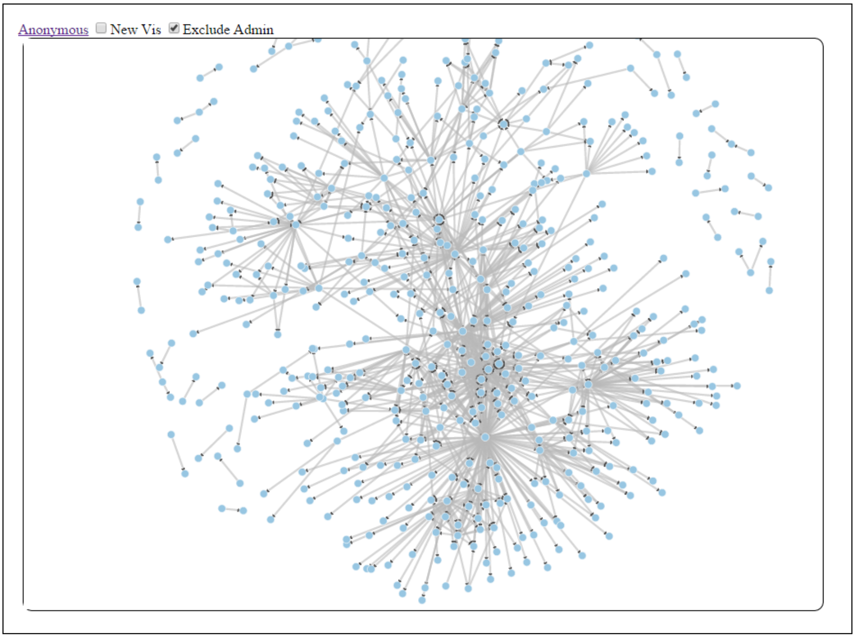


Figure 4.11: Visualization of Anonymous Graph without Admin

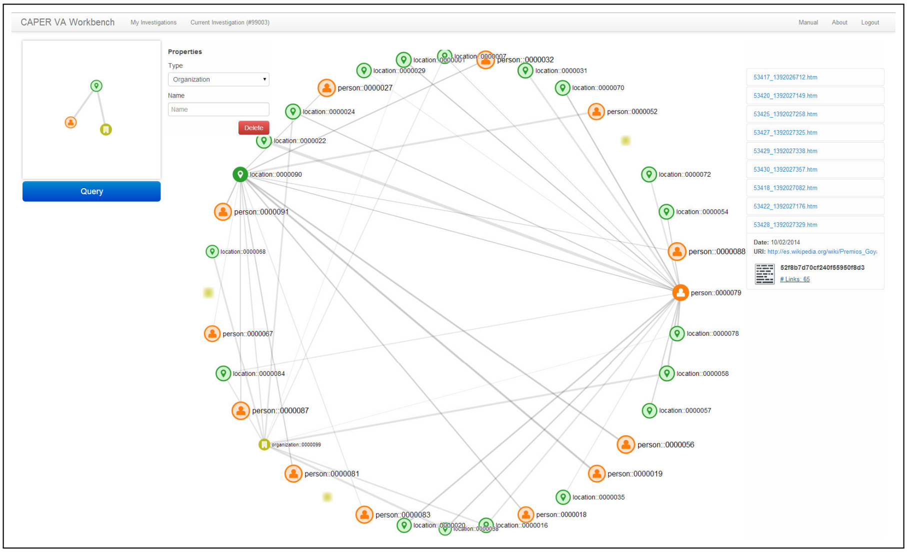
Figure 4.10 is more dense than 4.11, but both of them may be useful for LEA purposes.

4.2.3 Integration of Social Interaction analysis and textual analysis

The method introduced in the previous paragraphs is innovative, at the best of our knowledge, but take in account only the interactions between users. However, on Facebook, also the text of posts and comments is very important. In our work [6], we approach to this challenge. In this case, we use two different types of analysis: interaction networks and named entity networks. These two networks are constructed separately and then merged into the final network that is shown to the end user.

The name entity network is created by the Synthema Text Mining platform (TM). This module extracts all the entities from the Facebook's content. The TM process produces a Knowledge Annotation Format (KAF) file as output. The pipeline executes the following steps: Linguistic Analysis, Multiword, Word Sense Disambiguation, and Name Entity Recognition.

Named entity relationships (retrieved from the linguistic analysis of user generated textual content) and user interaction relationships (retrieved from the analysis of Facebook's social graph) are added to the KAF Entity Relationship layer. The CAPER Social Network Analysis takes as input the KAF Entity Relationship layer and builds a network graph based on the relationships between users and cited entities. The CAPER Visual Analytics (VA) Application allows end users to display and search for specific patterns inside the data. For this use case, Synthema provides a visual query interface to define graph patterns, which can search inside the database. Search results are then visualised using a circular graph layout, showing all entities (Facebook's users and cited entities) as part of the specified pattern (Fig. 4.12).



Properties

Type: Organization

Name: Name

53417_1392026712.htm
53420_1392027149.htm
53425_1392027258.htm
53427_1392027325.htm
53429_1392027338.htm
53430_1392027357.htm
53418_1392027082.htm
53422_1392027176.htm
53428_1392027329.htm

Date: 10/02/2014
URI: http://es.wikipedia.org/wiki/Premios_Goy

52f8b7d70cf240f55950f8d3
Links: 65

Figure 4.12: VA Application - Visualisation of entity relationships

4.3 TOUR-PEDIA: Social Media Analysis for Tourism domain

As introduced in the paragraph 3.4, a common tourist question is “What is the best Hotel/Restaurant in Rome?” Various tourism websites (like TripAdvisor) allow the users to search hotels, restaurants, and so on. Moreover, these platforms allow end users to leave a review for that place. In this way, it is possible to create a ranking of the best hotel or restaurant in a city.

Unfortunately, most users generally need to explore several pages on the web to extract the required information. Moreover, most of the tourism websites analyse and show the reviews from their own source. This tend to create the problem that a single dataset of reviews may be unrepresentative. Within the project OpeNER, we have developed Tour-pedia, a Web application available at[2] that exploits the OpeNER pipeline in order to display the sentiment of the reviews in tourist places. Tour-pedia uses the reviews extracted from Social Media (i.e. Facebook, Foursquare, Google Places, and Booking.com). Each review is processed by the OpeNER pipeline and is rated, to extract its specific sentiment (positive, negative or neutral). For each place a global sentiment is associated that depends on the sentiments of each place's reviews.

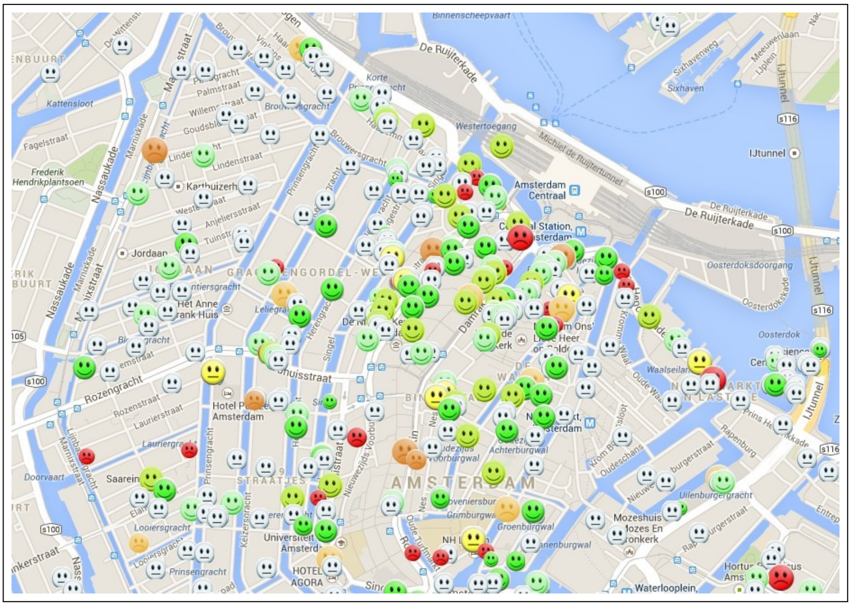


Figure 4.13: A snapshot of Tour-pedia

Tour-pedia shows all the sentiments of all places on a map (see Figure 4.13). This view allows a user to locate the best places with little effort. In practice, Tour-pedia guides the user in choosing the most suitable solution for his needs. There are many initiatives having almost the same purpose as Tour-pedia such as (Dun-

[2] http://www.tour-pedia.org/

lop et al., 2004) [40] and (Kenteris et al., 2009) [66]. Dunlop et al. designed and implemented a tourism information software program called Taeneb City Guide, for Personal Digital Assistant (PDA) and handheld computer. It is limited to the city of Taeneb. The main features of the application are the dynamic map interface, the dynamic information content and the community review system.

Kenteris et al. described the issues connected to a Mobile tourism application both in terms of networking capabilities of mobile and in terms of User Experience of the application (design, usability, portability). In addition, they implemented a prototype named myMytileneCity Guide.

The web platforms, like The Hotel Map and Google Hotel Finder, are examples of similar projects. The Hotel Map[3] is a web application, which shows hotels on a world map and allows the user to obtain some information, such as address, website, and reviews from Travel Now[4] about the selected entity. However, the graphic seems very old and the website seems not so rich in information. Google Hotel Finder[5] is a Google service for room booking. After the specifying the dates of the holiday, the user can see a list of available hotels with their regarding address, website, services, reviews, and price.

The above mentioned websites offer information about accommodations but focus only on booking rooms or beds.

The projects Hotel Map and Google Hotel Finder are devoted only to accommodations, while also Tour-pedia includes POIs, attractions, and restaurants. Finally, Google Hotel Finder provides links to many external Social Media that allows users to book a room, while the others do not. However, Tour-pedia provides links to other kinds of Social Media platforms, i.e. Facebook, Foursquare, Google Places, and Booking.com. Details about the data extraction of Tour-pedia datasets are available on paragraph 3.4.2.

Table 1 shows a comparison between the four described initiatives plus Tour-pedia. B stands for Booking, E for Expedia, FS for Foursquare, FB for Facebook, GP for Google Places, TA for TripAdvisor, and I for Instagram. The Hotel Map and Google Hotel Finder cover all over the world, Tour-pedia only a subset of Europe and the Taeneb City Guide and myMytilenCity Guide only a city. However, the number of places hosted by Tour-pedia is greater than those hosted by The Hotel Map. This means that Tour-pedia contains more places than The Hotel Map for the same location.

[3] http://www.thehotelmap.net/

[4] http://travel.ian.com/

[5] https://www.google.com/hotels

Table 4.3 summarizes the most important information regarding the above projects.

Name	**Coverage**	**Categories**	**Num. places**	**Social Media**	**Reviews**
Taeneb City Guide	Taeneb	tourism	n.a.	NO	YES
myMytileneCity Guide	Lancaster, UK	Accommodation, Restaurant, Attraction	n.a.	NO	NO
The Hotel Map	all the World	Accommodation	$\geq$ 200.000	TA	NO
Google Hotel Finder	all the World	Accommodation	n.a.	B, E	YES
Tour-pedia	Some parts of Europe	Accommodation, POI, Attraction, Restaurant	$\geq$ 500.000	FS, FB, B GP, I	YES

Table 4.3: Comparison between existing initiatives

4.3.1 Merging Places

As previously mentioned, Tour-pedia shows on a single web page the places captured from the following Social Media platforms: Facebook, Foursquare, Google Places, and Booking. To accomplish this, various issues must be overcome. The first problem is the data integration of the places captured from the four Social Media platforms. This operation is known as Holistic Data Sources2.2. In details, we perform the merging operation of two geographical datasets. In particular, we assume that the entries, of the two database, are characterized at least by two attributes:

- Geographical coordinates (latitude and longitude);
- Text information called r-string (name and address of the place).

Moreover, we assume that the same resource 1) could not be present in both datasets, 2) could have different values of attributes in the two datasets, 3) may have different (or similar) r-strings in the two datasets. Given these hypotheses, we define a Merger algorithm, which integrates entries belonging to two different datasets. The algorithm works in two steps: a) geographical search, b) string similarity. The geographical search lists all the pairs (e_a, e_b) such as the geographical distance between e_a and e_b is less than a given threshold tt.

```
g = empty;
for every e_a do
    g[a] = empty;
    for every e_b do
        if distance(ca, cb) <= x then
            push(e_b, g[a]);
        end
    end
end
return g
```

Algorithm 7: Merger Algorithm

In detail, the distance between c_a and c_b is calculated as follows:

$$\Delta(c_a, c_b) = \arccos\{\sin\phi_a \sin\phi_b + \cos\phi_a \cos\phi_b \cos\Delta\lambda\}R \tag{4.6}$$

where $\Delta\lambda = \lambda_b - \lambda_a$ and R is the radius of Earth.

The geographical match returns a value for each pair of items, which represents the distance between them. If this value is less than a threshold (x), then the two items are considered matched. When all the items are compared, the algorithm performs the string similarity between all the two r-strings of two items of the previous step. If the string similarity is greater than a string threshold (ss), the two items have a match. We use the Merger Algorithms on the places captured by the four Data Extractors (see Table 3.9). The results of the Merger algorithm are publicly available with this Web API[6]. Table 4.4 shows the number of places for every City.

City	Accommodation	Attraction	Restaurant	Point of Interest
Amsterdam	1393	3185	5241	10441
Paris	3397	4351	21854	26927
London	4372	20727	80510	50930
Barcelona	2450	2390	10778	8411
Rome	5207	7317	20881	15249
Berlin	2887	9660	28888	14484
Dubay	1515	5038	7105	7342

Table 4.4: Detail of Merged places for Cities

4.3.2 Exploitation of OpeNER Linguistic Pipeline for sentiment analysis of Tour-pedia reviews

The OpeNER project provides a set of ready-to-use modules for the processing of natural language. OpeNER focuses on building a linguistic pipeline that supports six European languages: English, Spanish, German, French, Italian, and Dutch, in order to enable the identification and disambiguation of named entities and the analysis of sentiment in opinionated texts.

Tour-pedia exploits the OpeNER pipeline. A dedicated module elaborates the text of each review, exploiting the following modules of the OpeNER pipeline: language identifier, tokenizer, polarity tagger, pos-tagger, and opinion detector. The language identifier extracts the language of the review. Then, the tokenizer extracts tokens from the text of the review. After that, the pos-tagger extracts the parts of speech for each term in the review. The polarity tagger extracts the polarity of each term. The opinion-detector, eventually, extracts the opinion.

[6] http://tour-pedia.org/api/getPlacesStatistics

Once analysed, we aggregate the polarity of all reviews about the same place. In this way, we perform a sentiment score about a place[34]. Finally, the sentiment of every place is shown on a map, see Figure 4.14. As previously stated, Tour-pedia shows the result of sentiment analysis on reviews extracted from Facebook, Foursquare, and Google Places. For this reason, sentiments resulting from the OpeNER analysis reflect real users' sentiments. A mechanism for continuously analysing reviews should be implemented to automatize new capture sessions. However, the overall opinion about a place does not change frequently, unless the place itself changes something (adding new features or solving issues reported in past reviews). For this reason, it is quite reasonable that the analysis is an offline task. Tour-pedia is a practical example of exploitation of the OpeNER pipeline. Tour-pedia exposes the analysed data as linked data node and provides SPARQL endpoint [14]. The service uses a D2R server[7]. For each place, the VCARD [60] and DBpedia OWL[8] ontologies are used to represent the generic properties. Instead, Acco [53], Ontology [30], and GoodRelations [52] are used for domain specific properties.

Tour-pedia provides a Restful API to access places and statistics. The output of a request can be in JSON, CSV or XML format. An example of search is available at this URL:

http://tour-pedia.org/api/getPlaces?parameters

The parameters must be at least one of the following: location (the location of the places), category (the type of the places such as accommodation, attraction, restaurant, poi), and name (the keyword to be searched).

4.3.3 Tour-pedia Web Application

Recent research showed that the APIs provided by Google Maps are very flexible (Pan et al., 2007) [80]. For this reason, Tour-pedia exploits them and it emulates the navigation style of Google Maps[9], leveraging and enhancing its fundamental characteristics: a map that occupies the whole page; a simple menu placed over the map; a search bar embedded inside the map itself.

In order to draw users' attention on the map (the focal point of the interface) all info-windows appear over the map, without subtracting too much space. Figure 2 shows an example of info-window for "The Monk Amsterdam Apartments".

[7] http://d2rq.org/

[8] http://wiki.dbpedia.org/Ontology

[9] https://maps.google.com/

Places appear on the map as smileys: colour and mood of the icons are determined by aggregating the sentiment extracted from the reviews for that entity. If there are more positive reviews compared with the negative ones, on the map the place will represented by a green smiley. Otherwise, there is a red smile. Different colours express intermediate ranges. White locations have no reviews available for evaluation. In addition, the size of the emoticon is proportional to the number of reviews for that entity, so big smileys mean many reviews and small smileys mean few reviews.

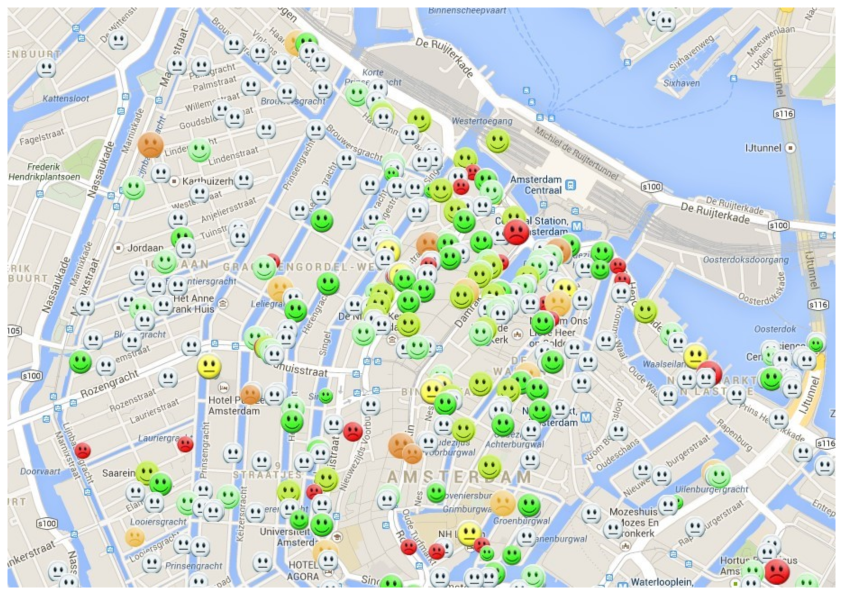


Figure 4.14: Tour-pedia Web Interface

5 Conclusions

In recent years, Social Media platforms have become very important tools to enable communication between users, and to share opinions, photos, videos, and others. In addition, many companies are very interested in Social Media data, especially personal data, for marketing purposes. From a scientific point of view, a growing number of researchers have been using this information for various investigations.

This thesis provides a contribution to Social Media monitoring and analysis by answering the following questions:

i) What specific capabilities on the Social Media platform are required to enable data capture by a structured tool?
ii) Given such enabling means, how can the data capture tool deal with the related limitations?
iii) What is the impact of flexibility and robustness requirements in data capture on the tool architecture? What are the required components?
iv) Is it possible to derive a simplified, lean version of the proposed general architecture to implement efficient tools to address specific domains?
v) What kind of analyses can be carried out on the captured data?

To answer these questions, we used a bottom-up approach: from the study of particular domains of application of Social Media data, we approached issues related to the collection and management of enormous quantities of data. Then, we studied the techniques for capturing data from Social Media. The main result of this phase was the identification of the Web APIs, provided by Social Media platforms, as one of the most promising methodology.

Next, we proposed a generic theoretical architecture for capturing data from a platform. Since some requirements were relaxed in a specific domain of investigation, we implemented a simplified version of the proposed architecture in three domains of investigation: Online Reputation, Social Media Intelligence (SOCMINT), and Opinion Mining in tourism. For each domain, we detailed the advantages and disadvantages of our implementation.

Finally, we presented the analysis of captured data in the aforementioned domains. In the Online Reputation field, we presented SocialTrends a web application that collects, elaborates, and visualizes Social Media data from Facebook, Twitter, and YouTube. In the SOCMINT field, we presented the European Project CAPER. Within this project, we developed two tools called Social Media Capture (SMC) and Social Media Analyser (SMA) for capturing and analysing data from Facebook. The SMA, in particular, employed an innovative methodology for weighing the interactions between users in terms of strength, frequency, and duration.

In the Opinion Mining field, we presented the European Project OpeNER. Within this project, we have developed Tour-pedia, a Web application that exploits the OpeNER's linguistic pipeline to show the sentiment associated with tourist places. In particular, we analysed places belonging to different typologies: accommodation, attractions, points of interest, and restaurants. A strong point of our work is that the implemented solutions were developed within the context of two European Projects CAPER and OpeNER. Moreover, SocialTrends and Tour-pedia are publicly available.

Summarizing the main contributions of our work: i) Creation of a generic architecture for capturing data from different Social Media platforms ii) Presentation of a simplified version of the proposed architecture for specific domain of investigation iii) Exploitation of captured data for examples of analysis in above mentioned domains. This thesis opens up various possible scenarios. First, we could capture all the nodes of a network. Then, we could implement several algorithms, such as Snowballs, to capture all public users from a Social Media.

We can enrich SocialTrends in different ways. For example, we could add new platforms, to capture more values for understanding of the Online Reputation of an entity. Moreover, we could study different modalities for creating a unique score of the entity's Online Reputation in all Social Media. Examples of commercial services that provide this value are Klout Score[1] and Wevo[2]'s SocialScore in TvZap[3].

We can analyse particular events that produce a high increment of popularity of an entity (peaks of popularity). These peaks might be studied with Anomaly Detection techniques. This research could be useful for understanding how the real

[1] https://klout.com/corp/score

[2] http://www.wevo.it/

[3] http://tvzap.kataweb.it/chi-siamo/

and virtual worlds are connected. Lastly, we could use some sentiment analysis tools to establish whether there is a correlation between sentiment and fan count.

We developed the Social Media Capture (SMC) and Social Media Analyse (SMA) for the exploitation of Facebook UGCs. We could generalize this experience to create a complete approach to different Social Media platforms.

In fact, we studied only Facebook because it was the best platform for LEA purposes but we could extend this result to other domains. For example, we could exploit this analysis in political studies, where knowledge of the support' group of a candidate is necessary. A problem related to Facebook search API is the high rate of false positives. We could develop some modules to overcome this issue. In this way, we could create a search engine for Social Media. In this moment, the SMA offers a complete analysis of users' interactions. We could add the difference in time between interactions in the computation of edges' weight. Moreover, we could add the sentiment analysis of posts and comments to edges' weight. For this, we could conduct a study to establish whether there is a correlation between the sentiment of the content and users' interactions.

Inside Tour-pedia, we could add different features. For example, we may add a module for named entity recognition to show on the map all the entities (people, monuments, services) cited in the reviews. One of the most important results of Tour-pedia is that the Social Media platforms have enormous quantities of data about real places. Unfortunately, it is not simple to establish the quality of data.

This fact is due to duplication, incompleteness, and errors of User Generated Content. Moreover, there is not official ground truth of all the places in the world. For this reason, we could perform a qualitative comparison between places extracted from Social Media and places extracted from open datasets released by government agencies.

Another important field of research is the analysis of a topic's diffusion over Social Media. In this field, we could start from the RSS Feeds of different online newspapers for monitoring the diffusion of news on Social Media. Moreover, the aggregation of the information from two different platforms (RSS feed and Social Network) permits greater knowledge of information diffusion in society in terms of news categories (politics vs sports).

Unfortunately, because of the marketing implications of Social Media data, the platforms are increasingly restricting access to information. However, commercial solutions, such as the reseller GNIP or DataSift, are useful, although two issues arise. The first issue is the presence of another intermediary that could filter the data. Second, this service is not free and this is a big problem especially for research groups. It is important that researchers continue to access Social Media data. Otherwise, this field of research would become an opportunity only for government agencies, big companies and a limited number of universities.

References

1. Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions. 2014.
2. Stefano Abbate, Andrea Marchetti, and Davide Gazzè. Wp6 - entities and opinions 6.21 - named entity and opinion mining. Confidential deliverable, The OPENER Project (FP7- ICT -2011.4.1), 2013.
3. Stefano Abbate, Andrea Marchetti, and Davide Gazzè. Wp6 - entities and opinions 6.22 - named entity and opinion mining. Confidential deliverable, The OPENER Project (FP7- ICT -2011.4.1), 2013.
4. Alan S Abrahams, Jian Jiao, Weiguo Fan, G Alan Wang, and Zhongju Zhang. What's buzzing in the blizzard of buzz? automotive component isolation in social media postings. *Decision Support Systems*, 55(4):871–882, 2013.
5. Carlo Aliprandi, Giulia Di Pietro, Ercole De luca, Matteo Raffaelli, Maurizio Tesconi, Gazzè Davide, and Rubio Aitor Rodriguez. *The CAPER Project – capitolo 4 – Data Acquisition, Springer Books, in press*. 2014.
6. Carlo Aliprandi, Antonio Ercole De Luca, Giulia Di Pietro, Matteo Raffaelli, Davide Gazzè, Mariantonietta Noemi La Polla, Andrea Marchetti, and Maurizio Tesconi. Caper: Crawling and analysing facebook for intelligence purposes. In *ASONAM*, pages 665–669, 2014.
7. Carlo Aliprandi and Andrea Marchetti. Introducing caper, a collaborative platform for open and closed information acquisition, processing and linking. In Constantine Stephanidis, editor, *HCI International 2011 – Posters' Extended Abstracts*, volume 173 of *Communications in Computer and Information Science*, pages 481–485. Springer Berlin Heidelberg, 2011.
8. Naveen Amblee and Tung Bui. Harnessing the influence of social proof in online shopping: The effect of electronic word of mouth on sales of digital microproducts. *International Journal of Electronic Commerce*, 16(2):91–114, 2011.
9. Mark Andrejevic. Big data, big questions| the big data divide. *International Journal of Communication*, 8(0), 2014.
10. Gad Ariav. A temporally oriented data model. *ACM Transactions on Database Systems (TODS)*, 11(4):499–527, 1986.
11. Valerio Arnaboldi, Andrea Passarella, Maurizio Tesconi, and Davide Gazzè. Towards a characterization of egocentric networks in online social networks. In *On the Move to*

Meaningful Internet Systems: OTM 2011 Workshops, pages 524–533. Springer Berlin Heidelberg, 2011.

12. M. Avvenuti, S. Cresci, M.N. La Polla, A. Marchetti, and M. Tesconi. Earthquake emergency management by social sensing. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, pages 587–592, March 2014.
13. Marco Avvenuti, Stefano Cresci, Andrea Marchetti, Carlo Meletti, and Maurizio Tesconi. Ears (earthquake alert and report system): A real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1749–1758, New York, NY, USA, 2014. ACM.
14. Clara Bacciu, Angelica Lo Duca, andrea Marchetti, and Maurizio Tesconi. Accommodations in tuscany as linked data. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).
15. Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM, 2006.
16. Albert-László Barabási. *Linked: The New Science Of Networks*. Basic Books, 2002.
17. Bogdan Batrinca and PhilipC. Treleaven. Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY*, pages 1–28, 2014.
18. Anja Bechmann. Managing the interoperable self. *Nordmedia2013*.
19. Anja Bechmann and Stine Lomborg. Mapping actor roles in social media: Different perspectives on value creation in theories of user participation. *New media & society*, 15(5):765–781, 2013.
20. G. Berg, W. Julian, R. Mines, and F. Richman. The constructive jordan curve theorem. *Rocky Mountain J. Math.*, 5(2):225–236, 06 1975.
21. David Bollier. The promise and peril of big data. Technical report, The Aspen Institute, 2010.
22. Stephen P Borgatti. Centrality and network flow. *Social networks*, 27(1):55–71, 2005.
23. danah m. boyd and Nicole B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.
24. Peter Buneman. Semistructured data. In *Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 117–121. ACM, 1997.
25. Moira Burke, Robert Kraut, and Cameron Marlow. Social capital on facebook: differentiating uses and users. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 571–580, New York, NY, USA, 2011. ACM.
26. Moira Burke, Cameron Marlow, and Thomas Lento. Social network activity and social well-being. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1909–1912, New York, NY, USA, 2010. ACM.
27. Guido Caldarelli, Alessandro Chessa, Fabio Pammolli, Gabriele Pompa, Michelangelo Puliga, Massimo Riccaboni, and Gianni Riotta. A multi-level geographical study of italian political elections from twitter data. *PloS one*, 9(5):e95809, 2014.

28. Salvatore A. Catanese, Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. Crawling facebook for social network analysis purposes. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 52:1–52:8, New York, NY, USA, 2011. ACM.
29. Duen Horng Chau, Shashank Pandit, Samuel Wang, and Christos Faloutsos. Parallel crawling for online social networks. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 1283–1284, New York, NY, USA, 2007. ACM.
30. Marcirio Silveira Chaves, Larissa A. de Freitas, and Renata Vieira. Hontology: A multilingual ontology for the accommodation sector in the tourism industry. In Joaquim Filipe and Jan L. G. Dietz, editors, *KEOD*, pages 149–154. SciTePress, 2012.
31. Kim-Kwang Raymond Choo. Organised crime groups in cyberspace: a typology. *Trends in organized crime*, 11(3):270–295, 2008.
32. Munmun De Choudhury, Yu-Ru Lin, Hari Sundaram, K. Selçuk Candan, Lexing Xie, and Aisling Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? In *ICWSM'10*, pages –1–1, 2010.
33. Richard Colbaugh and Kristin Glass. Detecting emerging topics and trends via social media analytics. In *Proceedings of the 2011 IADIS International Conference e-Commerce*, pages 51–51, 2011.
34. Stefano Cresci, Andrea D'Errico, Davide Gazzè, Angelica Lo Duca, Andrea Marchetti, and Maurizio Tesconi. Towards a dbpedia of tourism: the case of tourpedia. In *International Semantic Web Conference (Posters & Demos)*, pages 129–132, 2014.
35. Stefano Cresci, Marinella Petrocchi, Angelo Spognardi, Maurizio Tesconi, and Roberto Di Pietro. A criticism to society (as seen by twitter analytics). In *Distributed Computing Systems Workshops (ICDCSW), 2014 IEEE 34th International Conference on*, pages 194–200. IEEE, 2014.
36. Aron Culotta, Ron Bekkerman, and Andrew Mccallum. Extracting social networks and contact information from email and the Web. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, number d, 2004.
37. Ratan Dey, Zubin Jelveh, and Keith Ross. Facebook users have become much more private: A large-scale study. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, pages 346–352. IEEE, 2012.
38. R. I. M. Dunbar. The social brain hypothesis and its implications for social evolution. *Annals of human biology*, 36(5):562–72, 1998.
39. R. I. M. Dunbar and S.G.B. Roberts. Communication in Social Networks: Effects of Kinship, Network Size and Emotional Closeness. *Personal Relationships*, 2010.
40. MarkD. Dunlop, Piotr Ptasinski, Alison Morrison, Stephen McCallum, Chris Risbey, and Fraser Stewart. Design and development of Taeneb City Guide: From Paper Maps and Guidebooks to Electronic Guides. In AndrewJ. Frew, editor, *Information and Communication Technologies in Tourism 2004*, pages 58–64. Springer Vienna, 2004.
41. Nicole B Ellison, Charles Steinfield, and Cliff Lampe. The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.
42. Charles Ess. Ethical decision-making and internet research: Recommendations from the aoir ethics working committee. 2002.

43. Charles Ess. *Digital media ethics*. Polity, 2013.
44. H. Fonseca, E. Rocha, P. Salvador, A. Nogueira, and D. Gomes. A facebook event collector framework for profile monitoring purposes. In *Computers and Communication (ISCC), 2014 IEEE Symposium on*, pages 1–6, June 2014.
45. Lise Getoor and Christopher P. Diehl. Link mining: A survey. *SIGKDD Explor. Newsl.*, 7(2):3–12, December 2005.
46. Fabio Giglietto, Luca Rossi, and Davide Bennato. The open laboratory: Limits and possibilities of using facebook, twitter, and youtube as a research data source. *roceedings of the 19th international conference Linking Geospatial Data*, 2014.
47. M. Gjoka, M. Kurant, C.T. Butts, and A. Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9, March 2010.
48. Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. Assessing the bias in communication networks sampled from twitter. *arXiv preprint arXiv:1212.1684*, 2012.
49. Mark S Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.
50. Mark S. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, December 1973.
51. Ralf Hartmut Güting. An introduction to spatial database systems. *The VLDB Journal—The International Journal on Very Large Data Bases*, 3(4):357–399, 1994.
52. Martin Hepp. Goodrelations language reference. Technical report, Hepp Research GmbH, Innsbruck, 2011.
53. Martin Hepp. Accommodation ontology language reference. Technical report, Hepp Research GmbH, Innsbruck, 2013.
54. Susan C Herring. Web content analysis: Expanding the paradigm. In *International handbook of Internet research*, pages 233–249. Springer, 2010.
55. R. A. Hill and R. I. M. Dunbar. Social network size in humans. *Human Nature*, 14(1):53–72, March 2003.
56. Shawndra Hill and Noah Ready-Campbell. Expert stock picker: the wisdom of (experts in) crowds. *International Journal of Electronic Commerce*, 15(3):73–102, 2011.
57. Clyde W. Holsapple, Shih-Hui Hsiao, and Ramakrishnan Pakath. Business social media analytics: Definition, benefits, and challenges. In *20th Americas Conference on Information Systems, AMCIS 2014, Savannah, Georgia, USA, August 7-9, 2014*. Association for Information Systems, 2014.
58. Nan Hu, Noi Sian Koh, and Srinivas K. Reddy. Ratings lead you to the product, reviews help you clinch it? the mediating role of online review sentiments on product sales. *Decision Support Systems*, 57(0):42 – 53, 2014.
59. M. Hurst and A. Maykov. Social streams blog crawler. In *Data Engineering, 2009. ICDE '09. IEEE 25th International Conference on*, pages 1615–1618, March 2009.
60. R. Iannella and J. McKinney. VCARD ontology. Available at: http://www.w3.org/TR/vcard-rdf/. Technical report, 2013.
61. IBM, Paul Zikopoulos, and Chris Eaton. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media, 1st edition, 2011.
62. Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD*

and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 56–65. ACM, 2007.

63. Christian S Jensen and Richard T Snodgrass. Temporal data management. *Knowledge and Data Engineering, IEEE Transactions on*, 11(1):36–44, 1999.
64. Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.
65. David Karpf. Social science research methods in internet time. *Information, Communication & Society*, 15(5):639–661, 2012.
66. Michael Kenteris, Damianos Gavalas, and Daphne Economou. An innovative mobile electronic tourist guide application. *Personal and Ubiquitous Computing*, 13(2):103–118, 2009.
67. DongSung Kim and Jong Woo Kim. Public opinion mining on social media: A case study of twitter opinion on nuclear power. *Proceeding of CES-CUBE*, 2014, 2014.
68. Kurniawati Kurniawati, Graeme G. Shanks, and Nargiza Bekmamedova. The business impact of social media analytics. In *ECIS*, page 48, 2013.
69. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.
70. MARIANTONIETTA NOEMI LA POLLA. Social media analytics and open source intelligence: the role of social media in intelligence activities. 2013/2014.
71. Jure Leskovec and Eric Horvitz. Planetary-scale views on an instant-messaging network. Technical report, 2007.
72. Stine Lomborg and Anja Bechmann. Using apis for data collection on social media. *The Information Society*, 30(4):256–265, 2014.
73. Merja Mahrt and Michael Scharkow. The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, 57(1):20–33, 2013.
74. Frank McCown and Michael L Nelson. What happens when facebook is gone? In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 251–254. ACM, 2009.
75. Erik Meijer and Gavin Bierman. A co-relational model of data for large shared data banks. *Commun. ACM*, 54(4):49–58, April 2011.
76. Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv:1306.5204*, 2013.
77. Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, and Tomas By. Sentiment analysis on social media. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM '12, pages 919–926, Washington, DC, USA, 2012. IEEE Computer Society.
78. Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. In *Conference on Weblogs and Social Media (ICWSM)*, pages 570–573, July 2011.
79. David Omand, Jamie Bartlett, and Carl Miller. Introducing social media intelligence (socmint). *Intelligence and National Security*, 27(6):801–823, 2012.
80. Bing Pan, JohnC. Crotts, and Brian Muller. Developing Web-Based Tourist Information Tools Using Google Map. In Marianna Sigala, Luisa Mich, and Jamie Murphy, editors, *Information and Communication Technologies in Tourism 2007*, pages 503–512. Springer Vienna, 2007.

81. Wolfgang Reinhardt, Tobias Varlemann, Matthias Moi, and Adrian Wilke. Modeling, obtaining and storing data from social media tools with artefact-actor-networks. In *Proceedings of the 18th Intl. Workshop on Personalization and Recommendation on the Web and Beyond*, 2010.
82. Sam G.B. Roberts, Robin I.M. Dunbar, Thomas V. Pollet, and Toon Kuppens. Exploring variation in active network size: Constraints and ego characteristics. *Social Networks*, 31(2):138–146, May 2009.
83. Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. Influence and passivity in social media. In *ECML/PKDD (3)*, pages 18–33, 2011.
84. A. Rosi, M. Mamei, F. Zambonelli, Simon Dobson, G. Stevenson, and Juan Ye. Social sensors and pervasive services: Approaches and perspectives. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on*, pages 525–530, March 2011.
85. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *Knowledge and Data Engineering, IEEE Transactions on*, 25(4):919–931, 2013.
86. Jose A Scheinkman. Social interactions. *The New Palgrave Dictionary of Economics*, 2, 2008.
87. A. Semenov, J. Veijalainen, and A. Boukhanovsky. A generic architecture for a social network monitoring and analysis system. In *Network-Based Information Systems (NBiS), 2011 14th International Conference on*, pages 178–185, Sept 2011.
88. A. Semenov, J. Veijalainen, and A. Boukhanovsky. A generic architecture for a social network monitoring and analysis system. In *Network-Based Information Systems (NBiS), 2011 14th International Conference on*, pages 178–185, Sept 2011.
89. M.A. Smith. Nodexl: Simple network analysis for social media. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, pages 89–93, May 2013.
90. J. Sterne and D.M. Scott. *Social Media Metrics: How to Measure and Optimize Your Marketing Investment*. New Rules Social Media Series. Wiley, 2010.
91. Fred Stutzman, Ralph Gross, and Alessandro Acquisti. Silent listeners: The evolution of privacy and disclosure on facebook. *Journal of Privacy and Confidentiality*, 4(2):2, 2013.
92. Lei Tang, Huan Liu, Jianping Zhang, and Zohreh Nazeri. Community evolution in dynamic multi-mode networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’08, pages 677–685, New York, NY, USA, 2008. ACM.
93. M. Tsvetovat and A. Kouznetsov. *Social Network Analysis for Startups: Finding Connections on the Social Web*. Real Time Bks. O’Reilly Media, 2011.
94. Jeffrey D. Ullman and Jennifer Widom. *A First Course in Database Systems*. Prentice-Hall, 1997.
95. G Vinodhini and RM Chandrasekaran. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6), 2012.
96. Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42. ACM, 2009.
97. Nuwan Waidyanatha. Towards a typology of integrated functional early warning systems. *International journal of critical infrastructures*, 6(1):31–51, 2010.

98. Pete Warden. Data source handbook.

99. Christo Wilson, Bryce Boe, Alessandra Sala, Krishna PN Puttaswamy, and Ben Y Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. Acm, 2009.

100. Chi-In Wong, Kin-Yeung Wong, Kuong-Wai Ng, Wei Fan, and Kai-Hau Yeung. Design of a crawler for online social networks analysis. *WSEAS Transactions on Communications*, 13, 2014.

101. Daniel Zeng, Hsinchun Chen, Robert Lusch, and Shu-Hsing Li. Social media analytics and intelligence. *Intelligent Systems, IEEE*, 25(6):13–16, 2010.

102. Aoying Zhou, Weining Qian, and Haixin Ma. Social media data analysis for revealing collective behaviors. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1402–1402, New York, NY, USA, 2012. ACM.

Acknowledgments

Foremost, I would like to thank my supervisors Maurizio Tesconi, Alessio Bechini and Andrea Marchetti for their support and their valuable advice that helped me during this threes years.

I would like to express my gratitude to my Research group, the Web Application for the Future Internet at IIT-CNR, which gave me a constant support to perform this work.

Besides my advisor, I would also like to express my most gratitude to my colleagues working on the CAPER FP7 and OpeNER FP7 projects. These experiences gave me the opportunity to conduct my research activity and know the great world of European researchers.

I would also express my most sincere gratitude to Maria Claudia Buzzi and Marina Buzzi for the support during the projects ABCD and Mediterranean Diet.

I would thank also my colleagues (in alphabetic order): Matteo Abrate, Clara Bacciu, Sergio Bianchi, Stefano Cresci, Fabio Del Vigna, Andrea D'Errico, Mariantonietta Noemi La Polla, Angelica Lo Duca, Fabio Valsecchi, Francesca Sacchini and Caterina Senette. A special thank goes to Fabio Del Vigna for his help on the revision phase.

A thank goes to all the other people at IIT-CNR with whom I shared many good moments during these three years. Moreover, I would like to thank to the "ing. secretariat" group for all answers.

Last but not least, I would like to thank my family and my closest friends and all the people who have been part last three years of my life. My last thank is direct to a person that is not in this dimension anymore. I miss you very much and I will never forget you.