POLITECNICO DI MILANO

School of Industrial and Information Engineering

Master of Science in Management Engineering

POLITECNICO
MILANO 1863

A survey on data analytics techniques to analyse social media for companies

Supervisor

Prof. Barbara Pernici

Master thesis of:

Davide Sala, 877130

Academic year 2017/2018

Abstract

One of the main trends that in the recent years is affecting, not only the information and technology environment, but each area of the everyday life is Big Data. With the advent of internet of things and smartphones data are generated with a continuously increasing volume and a high variety at an incredible rate. Another huge source of data are social media, nowadays the majority of persons, with the availability of an internet connection, create and share data and contents with the rest of the world. This data represents an incredible resource of value and has several areas of application; from e-government and politics 2.0 to smart health and wellbeing passing through science and technology, security and public safety and e-commerce and market intelligence. The goal of this thesis is precisely the individualization of the possible applications of social media data and mainly the description of the particular techniques and practices that a company should take into consideration to exploit the value enclosed within this data. In this direction the paper is divided into three main parts, in the first one a description of the main characteristics and aspects of Big Data is presented with an introduction to the environment of social media. The second one is instead subdivided into three sub sections that take into consideration three different types of analytics related to social media data; respectively structure based analytics, content based analytics and descriptive analytics. Finally the last chapter, the conclusion, provides an overall vision of the whole work.

Table of Contents

1.Introduction

1.1 Definition of Big Data

The objective of this paper is to analyse the role of Big Data in the field of social media. That of Big Data is now-a-days one of the most prominent issue not only in the IT industries, but also in any kind of industries. Indeed, in such a digital world that of gathering and exploit data is becoming one of the important issues that a firm must face. Reasoning on Big Data the first aspect that jump to mind is related to the quantity of data due to the huge number of sources; with the establishment of the internet of things, sensor networks, open data from the web paradigms and the continuous increasing of the devices connected to internet, such as smartphones, personal computers and tablets, the number of data created every day is not even comparable to the one of ten years ago (according to the news provided by IBM, (web site 1) 90% of the data available in December 2017 has been created only in 2016 and 2017) . But the volume is not the only one feature that characterized Big Data; in the last years multiple definitions of the term Big data were born. In the table below introduced by (Cavanillas, J.M., Curry, H., 2016) are collected some of them.

Big Data definition	Source
"Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization"	Laney (2001), Manyika et al. (2011)
"When the size of the data itself becomes part of the problem and traditional techniques for working with data run out of stream"	Loukides (2010)
Big Data is "data whose size forces us to look beyond the tried-and-true methods that are prevalent at the time"	Jacobs (2009)
"Big Data technologies [are] a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis"	IDC (2011)
"The term for a collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications"	Wikipedia (2014)
"A collection of large and complex data sets which can be processed only with difficulty by using on-hand database management tools"	Mike 2.0 (2014)
"Big Data is a term encompassing the use of techniques to capture, process, analyse and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies." By extension, the platform, tools and software used for this purpose are collectively called "Big Data technologies"	NESSI (2012)
"Big data can mean big volume, big velocity, or big variety"	Stonebraker (2012)

Table 1 - Table of Big Data definitions (Cavanillas and Curry, 2016)

From the above definitions emerged, with volume, other two features to describe Big Data: velocity and variety. With the passing of years new Vs are added creating a robust model to describe Big data in a complete way. This V's model is treated by (Patgiri, R., Ahmed, A., 2016) and below there is a brief presentation of it.

1.2 5Vs Model

The terminologies that characterized Big Data are multiple and each one emphasizes a different quality of it. Below there is the list of the Vs used to describe Big data; the first five are the most frequently in use, the others are more recently introduced to better point out the picture.

- **Volume.**
  The volume refers to the huge amount of data that every day is created, the quantity grows exponentially without any bound. The volume is calculated with zettabytes, yottabytes and probably even more in the next future. The volume increases because of the growing number of sources of data and the evolution in the technology infrastructures that permit the gathering of bigger and bigger amount of data. Exist three more Vs to better define volume: voluminosity; the greatness of volume, vacuum; refers to the need of empty space to store the data, and vitality; the power of enduring, some data are actively used today, some other could be useful in the future.
- **Velocity.**
  The velocity is strictly related to the volume, it refers to the rapidity whereby the data is generated and needed to be analysed. Velocity concerns with the speed of grow; due to internet users, IOT, cloud computing and web sites the data grows faster and faster, and with the speed of transfer, it is very complex to transfer and elaborate so a huge amount of data with low latency and at an incredible rate.
- **Variety.**
  Big Data are characterized by high variety of data due to the different types of data that every day an organization gather; structured, semi-structured and unstructured data exist, but this topic will be deeper developed below. The variety of the data is related to the variety of the sources, any source can create a specific type of data but also the same source can create two different types of data. For this reason, to manage Big Data, firstly it is necessary to understand what type of data is going to be analysed.
- **Veracity.**
  Veracity is linked to accuracy, precision and meaningfulness of the data. Wrong data bring to wrong insight and so to wrong decision making. For this reason, check the veracity of the data is a main issue in data management, but due to high volume, velocity and variety it is not simple and sometimes even possible to control the accuracy of the data and of the sources.

- **Validity.**
  The validity refers to worthiness, and so to the utility of the data. Sometimes data can become obsolete and so useless for the organization that gathered it. It is essential to exploit data utility until it is valid. Some types of data need to be elaborated and analysed in real time because of the their really low utility life.
- **Value.**
  The only reason to gather and manage Big Data is to obtain value from it. Without any process of extraction of knowledge or analysis, data is useless. The generation of value starting from data is a complex process and it will be deeper described in the next paragraph.
- **Visibility/Visualization.**
  The visibility is the possibility to be seen, instead the visualization is the capacity to see the insights and the hidden knowledge present in the data. For good management and exploitation of Big Data is not enough a good analysis but a good presentation of the results of the analysis, that permits to any sort of organizations to make the right decision at the right time, is also essential.
- **Virtual.**
  The term virtual describe clearly the nature of the data and of the processes related to data management. For example, the cloud computing paradigm, that is fundamental for Big Data, is based on virtualization.
- **Variability.**
  Variability refers to the nature of changing of Big Data; as already said, data could become obsolete or change for user modification. Data is not constant in time and needs to be updated continuously to not lose its significance and validity.
- **Vendee.**
  The term vendee refers to the users/client paradigm that moves and regulate the Big Data world.
- **Vase.**
  With the term vase the model refers to a requirement that need to be satisfied to manage Big Data, for example to the infrastructures to gather and manage this high volume, velocity and variety data.

1.3 Different types of data

As anticipated talking about the variety of Big Data, different types of Big Data exist. The simplest and most ready to be used data are structured data; this type of data is organized in a formatted repository, usually a database, and it comprises all data that can be stored in SQL databases, tables with columns and rows. Structured data are marked with relational key, for this reason structured query are very effective ad efficiently in this area. Moreover, these data are very robust; an example are relational data about the sales of a company that could be represented as a table where there is information about the details of the order (who/when/what buy, with what type of payment). Another type of data is semi-structured

one; this type of data is not represented by a relational database, but it has some organizational characteristics that permit to analyse it in an easier way. For some kinds of semi-structured data, after required process, it is also possible to store it in a relational database and so to transform them in structured one. An example of semi-structured data is XML language. The last and most complex to be analysed data type is unstructured data; there is no structural organization and pre-defined data model. Unstructured data are also the most frequent and used data, they include text data, like word and PDF format, audio data and video data. These types of data are so diffused because they are mainly users content created, for example in social networks the amount of text data that every day are created like posts, comments or tags is huge and could be exploit by organizations and profit companies to obtain value.

1.4 Big Data as a source of value

As the resource-based view affirms to get and maintain a competitive advantage on the competitors a firm should manage and exploit its tangible and intangible assets, capabilities and resources in the best way possible. Moreover, to obtain value from a resource is not sufficient to own it but it is essential to apply some practices and processes that permit the resource to generate value. According to this perspective Big Data is a key resource for organizations and companies that can be exploited to generate value. Big Data, after being transformed and analysed, generates two types of value; the first one, more common with respect to a business perspective, is the economic value. In the bound of a firm the data could be used in many areas and functions to get business insights and to take knowledge-driven decisions; starting from the recruitment of personal and raw material to the field of marketing where, for example, the data are useful to know the customers in a better way and so to cluster and target them in a more effective way. The other type of value, sometimes seen as a consequence of the first one, is the social value; it refers to the value created for the customers and in general for the citizens. For example, in field like health and transport the utilization of Big Data can bring improvement to the personal condition of many users and citizens. But as already said, owing Big Data is not sufficient (Zeng, J., Glaister, K.W., 2017) but a company must adopt some organizational practices and execute activities to extract value.

1.5 Big Data Value Reference Model

The management of Big Data, as a key source of value, is a complex and articulate issue, in order to have an overall vision of the problem is now presented a reference framework introduced by Big Data Value Association (BDVA) to take into consideration the main concerns, aspects and priorities for Big Data Value systems.

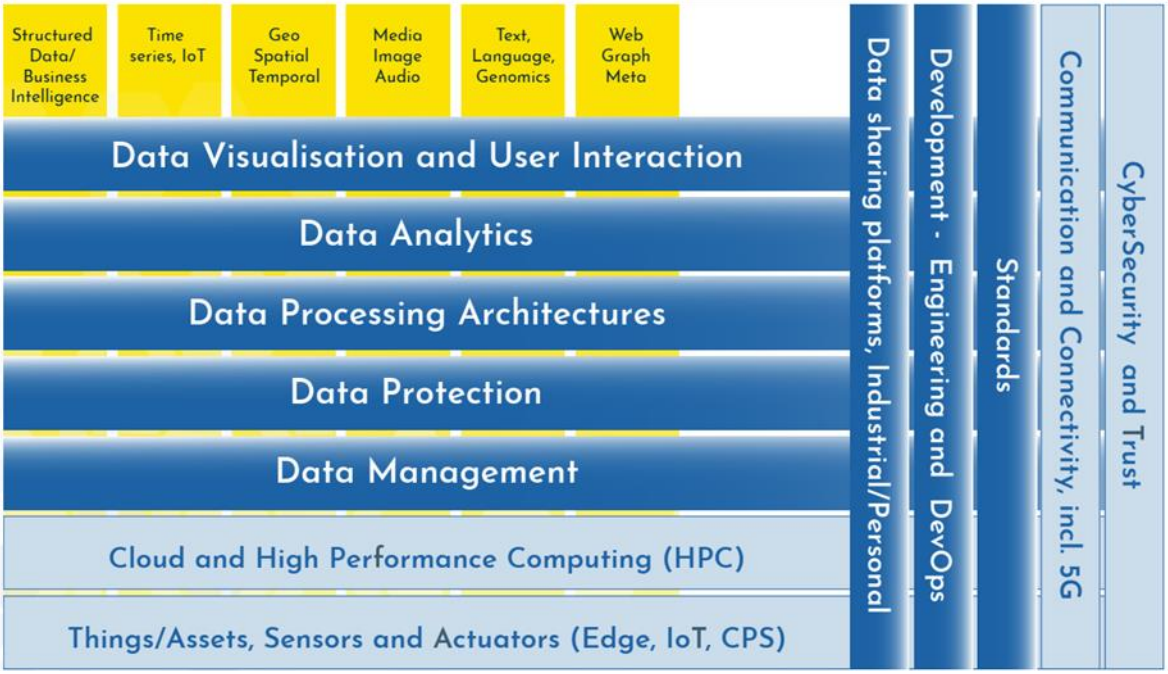

Figure 1- BDV Reference Model (BDVA, 2017)

In the above figure the yellow boxes represent the different data types included in the Big Data environment; from, as specified in the previous section, structured data to unstructured ones like geospatial or temporal data. The reference model is structured in horizontal and vertical concerns.

- **Horizontal concerns** cover specific technical priorities along the data processing chain, starting from data collection phase to data visualisation one. Horizontal concerns refer to the main macro practices that a company should focus on to perform an effective Big Data management.
- **Vertical concerns** refer instead to issues transversal to the technical ones that however can have a big influence on the results of data exploitation as a source of value.

Below there is a brief description of the first five technical concerns that are essential areas to be analysed in order to understand in which technical directions an organization should go to achieve effective results from the exploitation of the Big Data.

Data Management

The amount of data available is increasing day by day, one of the reasons for this huge enlargement are the several sources that in this digital era are available. According to this perspective the problem is that there is no an alignment between these data sources because a support language, appropriate resources and a semantic interoperability layer lack

(BDVA, 2016). In this sense data management is the ability of clearly define, interoperate, openly share, access, transform, link, syndicate, and manage data. In the past years there were some attempts to develop vertical processes for data management, but a consistent data lifecycle management still does not exist.

Data Protection

Data protection and anonymisation is another main priority in the field of Big Data and data analytics. In the huge amount of data that every day is collected there are person-specific and sensitive information from disparate sources such as social networking sites, mobile phone applications and electronic medical record systems. Gathering this data is a big opportunity, to create value for the firm, but also a risk. Recent works have demonstrated that simple approaches (BDVA, 2016), such as the removal or masking of the direct identifiers in a data set (e.g., names, social security numbers, etc.) are insufficient to guarantee privacy. Indeed, these protection strategies could be deceived by attackers that know specific details about the data subject. For this reason, the development of privacy models and techniques such as differential privacy, private information retrieval, syntactic anonymity, homomorphic encryption, secure search encryption, and secure multiparty computation, among others is a priority in the Big Data Value systems.

Data Processing Architectures

Today there is not only the need to process data-at-rest, inactive data physically stored, but, due to the high number of sensor data streams, is raised the need to process also data-at-motion, data not stored that could be represented as a continuous flow. The data must be processed in real-time with a low latency. Today is possible to realize integrated procession of data-at-rest and data-in-motion, but design of generic, decentralised and scalable architectural solution is required. Moreover, the capabilities of the existent systems to process data-in-motion and to reply to queries in real-time and for thousands of concurrent users are limited. The problem to achieve an effective and efficient system processing of stream of data is far from a solution.

Data Analytics

The term data analytics includes all the practices and methodologies to turn data into knowledge, from which it is possible to create value for the firm. The direction of the development of these methods is not only the transformation into knowledge of the data, but also the attempt to make these practices available to a wider audience. Regarding firm perspective, data analytics no more as a centralized company function, but as the possibility for each company function to exploit these potentialities by its own. The huge challenge of data analytics in these years and in the next ones is to deal with a great amount of information, from several sources, with different features, levels of trust and frequency of updating; all of these in a cost-effective and in an economically sustainable way. The focus of this survey will be on this specific priority applied to the field of social media environment.

Data Visualisation and User Interaction

Data visualisation has a key role for an effective exploration and exploitation of Big Data and the knowledge extracted from it. As data analytics, data visualization is a cornerstone to make Big Data exploitable to a wider audience, so that the value creation is worth both for the firm both for the society and the citizens. Visual analytics is the science of analytical reasoning assisted by interactive interfaces. The objective of data visualisation is to present to end users via traditional and innovative multi devices reports varying forms of media; ranging from text and charts, to dynamic, 3D and possibly augmented reality visualisations.

1.6 Big Data Value Chain

In addition to the practices described above, particular activities and processes are necessary to manage and extract value from Big Data; from the gathering of the data to the use of the knowledge extracted from them. All these activities are grouped and represented by the Big Data value chain, a value chain in the management field is used as a decision support tool that permit an organization to focus on its overall structure, it includes processes needed to obtain value from a product, service or resource. This permits to highlight which are the value-added activities fundamental to transform raw data into utilizable knowledge.

Figure 2-Big Data Value Chain (Miller and Mork, 2013)

The image above shows a Big Data value chain (Miller H.G., Mork, P., 2013) in which the activities are divided in three macro steps; data discovery, data integration and data exploitation; this permits an easily understanding of how the organization is structured and how the data management system operates. At an organizational level could be very useful to set a Big Data value chain to point out responsibilities and tasks needed for each step because it is not simple to coordinate and manage the overall flow of data in a firm. Now there is a brief description of each phase of the chain.

Data Discovery

The first macro phase is the preparation for the analysis of the data, in the discovery phase there is the understanding of which sources are needed and of what kind of knowledge is essential to manage those ones. Sometimes having too much low-quality sources slows the process and does not permit to focus on the more valuable ones. Another essential feature is the capacity to relate, in an efficient way, with high value sources and to prepare them to be used. Data discovery is composed by **collect and annotate**; the scope is to create a list of sources whose quality is described in terms of validity, consistency, completeness, accuracy and timeliness by metadata, only with the help of valid metadata the transformation of unstructured data in structured data is possible. Two techniques are

utilized to develop collect and annotate phase; the Dublin Core, which uses methods regarding web-based metadata to support metadata vocabulary terms, the second one is Department of Defense Discovery Metadata Specification, focused both on developing a metadata taxonomy and both on using that taxonomy to discover quality sources. The second activity that composed data discovery phase is **prepare**; in which the links with the sources is created, the access to the sources is made by coping them in a shared system. When this process is performed, for reason of privacy and security, it is essential to create rules and a language for access control. The possibility to use the data and the knowledge hidden into it must be strictly controlled and permitted only to those with the needed requirements. The last activity of the first macro phase is called **organize**; before integrating the data, to organize the data according to the choices taken by the source developer is necessary, for doing this it is possible to utilize both schemata both metadata repository. If the analysts share details about internal data organization with providers' environment and consumers' environment, they could benefit with the creation of a seamless upstream and downstream integration.

Data Integration

After the organization activity the data are ready to be represented in a common and integrated way, as later on will be explain each technique need a representation, so it is important to know which the final scope the data are devoted for is. To suit with the representation data coming from different sources and so with different organizations need to be elaborated, metadata are indispensable to track which modifications are done on the specific data. The combination of data coming from different sources can permit the discovery of hidden patterns and insights. Big Data can be integrated through a federated model, virtual integration, or through a data warehouse, physical integration. There are traditional technologies that support integration and combination of sources like relational databases, suitable for most kinds of tabular data, or emerging technologies like semantic web that instead is suitable with nontabular, nonnumeric and defined by rich networked relationships data. Using both the technology types gives to the analysts the capacity to explore and discover hidden knowledge in a more effective and efficient way.

Data exploitation

The last macro phase is data exploitation; after being gathered, prepared and integrated the data are finally ready to be analysed. There are a lot of very different techniques in relation to the data utilized and the algorithms implemented on them, but the scope of all these methodologies is to extract knowledge and insight to simplify decision-making process. The first activity of the exploitation phase is just the **analyse**; it consists in the elaboration of the Big Data, different techniques and methodologies are applied to find something that is not notable looking the raw data. It is part of the analysis maintaining the provenance between the input and the result and keeping track of everything in the metadata, so that other analysts can start from those result and increase their validity. Then **visualize** phase starts; visualization consist in the presentation of the results to the decision makers, good

results without a good presentation are useless because they cannot be exploit. So, it is very relevant to choose the right visualization tool depending on the situation, a lot of tools exist like static report or interactive audio and video support. The critic point is to select the right format to the right audience, to permit them to make decisions. **Make** decisions is the last activity of the chain and is the final scope of the Big Data utilization. Also, in this phase metadata are essential to track from what source the data comes and what analysis have been performed on it.

As already said the phases and activities vary a little bit according to the kind of techniques used and the area of interest. In the next chapter there is a deeper description of the Big Data analytics utilized in the social media environment.

1.7 Big Data applications

After this introduction to the world of the Big Data, it is significant to treat which are the area of applications of this important resource. Due to the high volume and the high variety of the data the spectrum of possible utilizations is very wide and various. The Big Data utility can be exploited in a lot of areas; not only in business or literature areas, but also in daily situations of people life. Indeed, the adoption of data perspectives does not bring only economic value, directly linked to a profit, but also a social value, Big Data, if used in the right way, can improve significantly the wealth of citizens. Below (Chen H. et al., 2012) there is an introduction to the major applications of Big Data in the business and in the social area.

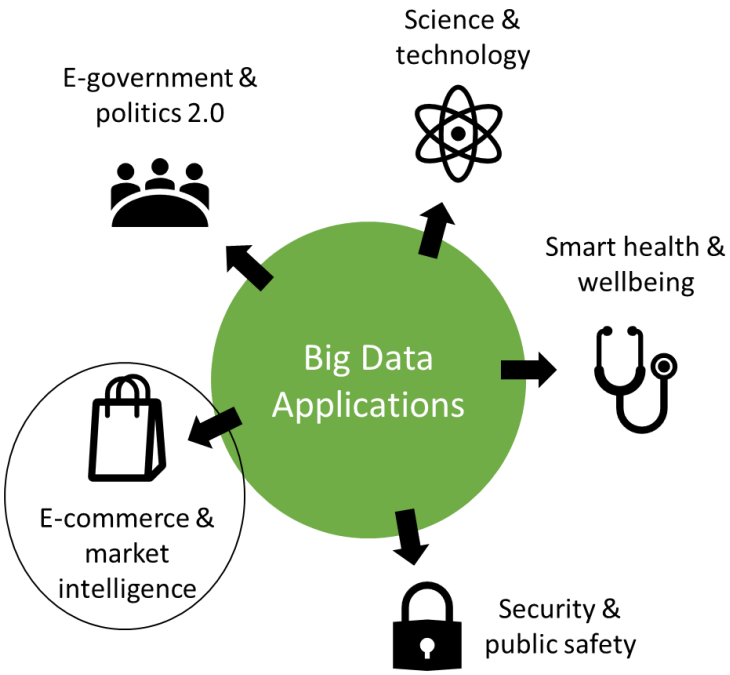

Figure 3-Big Data Applications

E-Government and Politics 2.0

In the recent years the paradigm of politics 2.0 takes place; it refers to the increasing participation and engagement of people in the politic world; politicians are open to a direct dialogue with citizens, people can create a relationship of asking and answering questions with them. This can happen only thanks to the advent of internet and Big Data. Indeed, the elaboration and analysis of data, through techniques like information integration, sentiment and affect analysis and content and text analytics, permit politicians to know what people think about political and social issues; in this way for example they can plan in a better way their campaign and target the right citizens for the right message. Through the utilization of this new channel of communication it is also possible the establishment of e-governments, where citizens have equal access to public ubiquitous government services. The goal of making the public administration transparent and of increasing the direct participation of the citizens to the decisional process is also the one of open government that pushes the utilization of open data, data that everybody utilize, reutilize and distribute without any sort of limitation, only with the obligations to mention the source and to maintain the data set open. Open data, for their nature, are an incredible source of value for, not only companies, but for every person that through internet has the possibility to access to them.

Science and Technology

The high volume of data and the increasing possibility for a larger number of organizations and people to access them benefit also the scientific field. Researchers and experts in different area, like medicine, mathematics and natural sciences, thanks to Big Data can discover new theories or develop existing concepts in a more efficient and faster way. For example, through text mining techniques they can check an infinite number of documents knowing which topic is treated in them or thanks to transfer learning techniques can used information from different disciplines in different languages in the same work. Big Data has also push improvements in the technology's world, to manage so high volume, high velocity and high variety data there is the need to better and better infrastructures and algorithms.

Smart Health and Wellbeing

One of the scientific fields where the utilization of Big Data has a huge impact is the healthcare. Even if traditionally the adoption of Big Data in healthcare was slower than in other industries, because of the resistance of doctors to be helped in making treatment decisions by a software, now data have an important role in this environment. For example, to control the satisfaction of the clients or to disease detection and diffusion, techniques like sentiment analysis, community detection and link prediction are used. The creation of social value thanks to Big Data is evident also in the improvement of the wellbeing of the citizens; in the area of transportation and education for example data are used to optimize and reduce costs of the services.

Security and Public Safety

Big data is largely utilized also in critical environments like security and public safety. In banks and in other businesses where transactions take place, Big data is essential for fraud detection, data platforms can analyse transactions and claims to discover in real time patterns and anomalous behaviours of the users to anticipate and stop possible fraud. Through community detection techniques is also possible to detect suspected organization, for example terrorist cells, to prevent attacks and massacres protecting the safety of the citizens.

E-Commerce and Market Intelligence

From business perspective one of the main advantages brought by Big Data is the fact that this huge amount of data permits to firms to know in a better way their clients and their potential buyers. The advent of IOT, smart devices, smartphones and tablets creates a continue stream of data that could be very valuable for firms if correctly analysed. Indeed, through data it is possible to cluster customers targeting the more profitable clients that fit with the offer, to create personalized advertising messages and to personalize the experience and value proposition for each cluster of costumers. The focus of the rest of the paper will be on this area of application.

1.8 Social Media Environment

After the description of the Big Data world, what it is, its characteristics and main applications; now there is an introduction to the social media environment, that will be the area of interest of the analysis of the next chapter. Social media is a generic term that refers to technologies and online practices that people used to communicate and to relate with others through the sharing of contents; (Zeng, D., 2010) "it is a conversational, distributed mode of content generation, dissemination, and communication among communities". In this definition is evident the centric position of the users, they are not only passive audience of the media, like could happen with more traditional media channels, television or radio, but they are also content creators. The rise of social media is strictly related to internet and mobiles, the smartphones' diffusion gives the ubiquitous possibility to be connected and to take part to the social life. In the figure below there are some numbers that in an immediate way make understandable how huge the rise of the social media world has been.

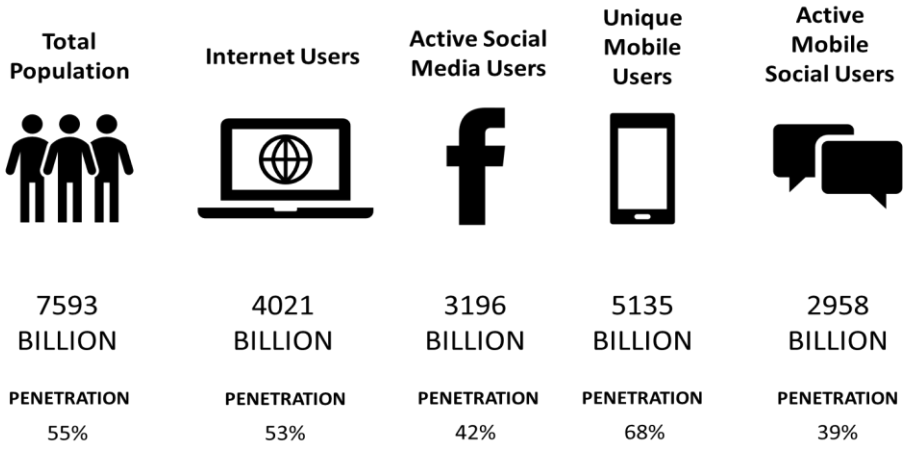

Figure 4- Social Media Environment (web site 3)

The social media world is very various; (Zarrella, D., 2009) different types of social media exist and each one has its peculiarities. For this reason, a brief description of the types of social media with their characteristics is now introduced.

- **Blogging.**
  A blog is a type of content management system, it is a personal web site where there is the possibility to publish short articles, from 100 words to many pages, called posts, that usually are more effective if develop a single topic. Blogs have social features; readers have the option to comment or to subscribe, or to use other social tools like blog rolls, list of links to other recommended blogs, and trackbacks, notifications to have been mentioned.
- **Twitter and Microblogging.**
  Microblogging is a form of blogging with limitation on the length of each post, in Twitter, the most famous site, the limit was 140 words up to September 2017, when the firm has stretched it for some users to 280. Twitter is a social site on which there is the opportunity to subscribe and to create an own account. Like for the social network it is possible to follow, reply/comment and direct messaging. Another feature of twitter is the retweet tool, that permits to share content created by other users. To manage the feed of news of each user, algorithms as trending topics and hashtags are utilized.
- **Social Networking.**
  Social networks are website connecting friends, that could be also offline friends or online-only ones. The most famous and used social networks today are Facebook, LinkedIn and Instagram. They all have common features; the building block are user pages, called profiles; where it is possible to find sensible information about the user, to have the possibility to link with other users using connecting or grouping options. It is also possible to do a wide range of actions like sharing photos, creating and organizing events and using various applications.
- **Media Sharing.**
  Media sharing sites goal is to permit to everyone to create and upload multimedia content, in the era of the smartphones, being always online, the number of user-generated content is huge; on YouTube, the most famous media sharing site, every minute 400 hours of video are uploaded (web site 2). Although media sharing sites provide social features to members, most of them are not registered and use this platform only for content viewing.
- **Social News and Bookmarking.**
  Social news are websites that permit users to submit and vote content from around the web, in this way the most appreciated links are isolated and made more visible. Bookmarking sites have a similar scope, they allow users to collect and store links that they found interesting and may wish to revisit. Counting how many times the links are stored, they highlight the most valuable ones.

- **Rating and Reviews.**
  Sites like TripAdvisor were born with the goal of give advices and suggestions. In the web the possibilities to make reviews and rating the products or services are infinite, this has changed the behaviours of the clients that, before to buy something, check online how other customers perceived the offer. So nowadays for a company is essential to monitor online rating and reviews.
- **Forums.**
  The main goal of forum sites is to create constructive discussion and sharing of ideas; the members can publish posts and respond to posts of other users. More a member stands out with good information and answers more he is respected by the community. Unlike for the others social media, do not exist few major forums, on the web the number of available forums is huge; each one speaking about a single topic or community.
- **Virtual Worlds.**
  Virtual worlds are computer-based simulated environments where users can create and personalize a specific character to explore independently and simultaneously the virtual world overcoming difficulties and interrelating with other players.

2. Social Media Analytics

As it is possible to note in the introduction the social media environment is really various, so, satisfying the social needs of different categories of people, it manages to interest a lot of users that are not only passive entities but, with their content generation, permit the continuous development of the networks. Due to this dual valency of users, content consumers and content generators, and due to the huge number of users, social media world is an incredible source of value that can be exploited only through the analysis of the data available inside it. In the first two sections of this chapter two different types of analytics that through the process of data create value from social media are introduced; structure-based analytics, that exploits the information that emerges from the analysis of the interactions and relationships between entities within social network, and content-based analytics, that extract knowledge and information from the user generated contents. In the last section of this chapter instead descriptive social analytics for the calculation of metrics are introduced.

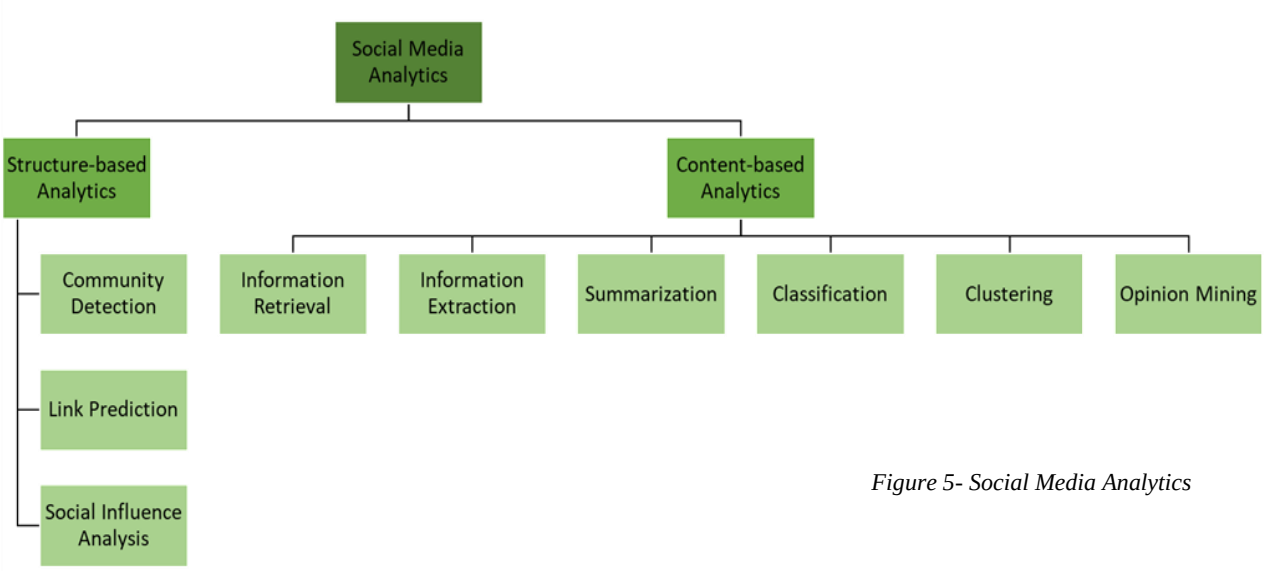

Figure 5- Social Media Analytics

2.1 Structure-based Analytics

Structure-based techniques aim at exploring the structure of social network, identifying the main structural features and extracting intelligence on the evolution and development of the net and of interactions between entities; in this perspective three fundamental issues that need to be explored emerge; community detection, link prediction and social influence analysis.

2.1.1 Community Detection

The expansion of the web and the constantly increasing numbers of user of social media, above all social networks, have permitted users to easily interact and to create relations. In network science (Bedi, P., Sharma, C., 2016), which through mathematical theory analyse and describe the behaviours and characteristics of networks, social networks are often represented as graph in which nodes correspond to individuals or entities and edges to the interactions among them. Within networks nodes whit similarities; they could show similar preferences, ideas or interests, have the tendency to group creating communities. The research and discovery of them into networks is called community detection and it can bring benefits and applications in different areas, from marketing to disease diffusion. The figure (Muhammad, A., J., et al., 2017) below tries to model the several existing community detection algorithms. Then there is a brief and general description of these techniques.

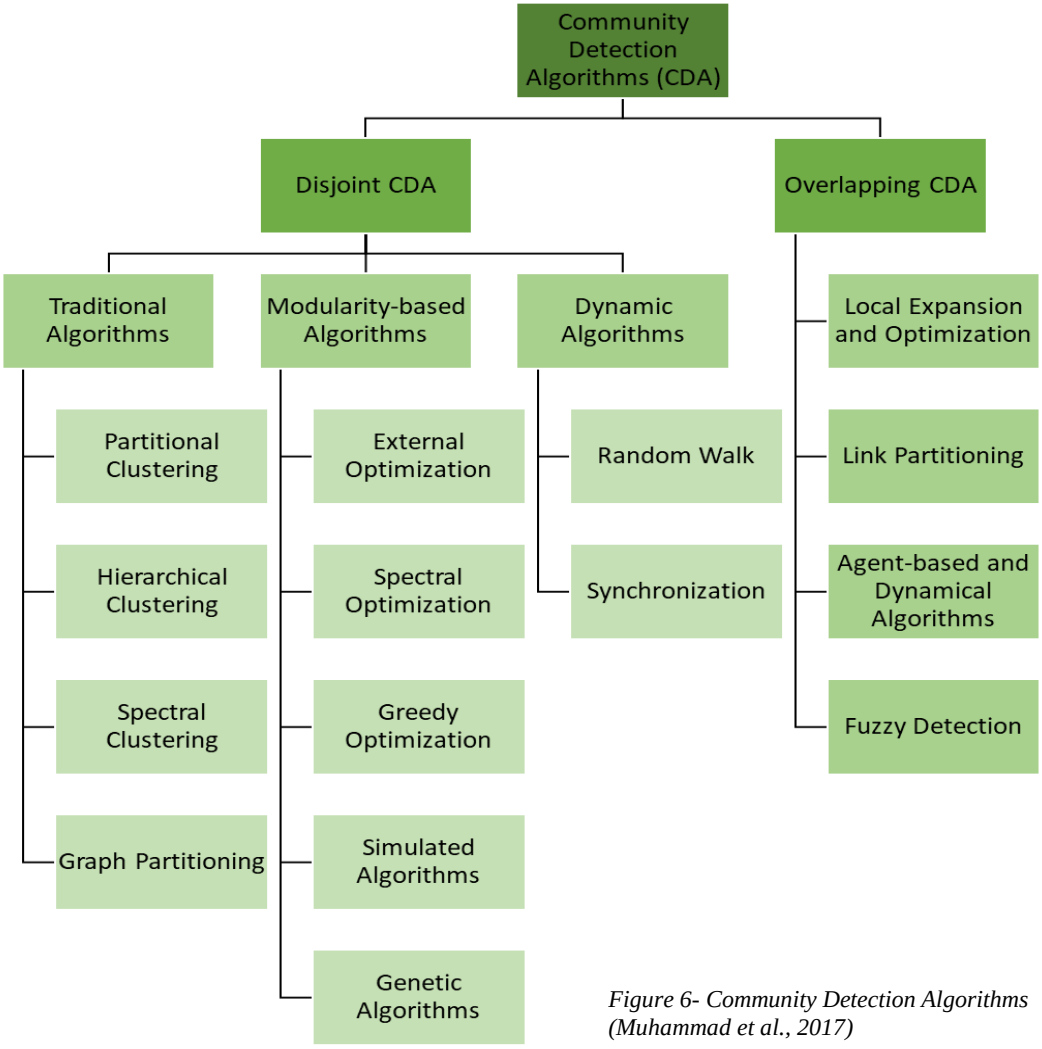

Figure 6- Community Detection Algorithms (Muhammad et al., 2017)

2.1.1.1 Disjoint Community Detection Algorithms

The first distinction is between disjoint or overlapping communities. In disjoint communities, nodes belong to a single community, instead in the overlapping ones some nodes can belong to more than one community. A lot of algorithms have been proposed to detect disjoint communities, some of them are briefly described below.

Traditional algorithms

Traditional algorithms are based on clustering techniques, in statistics cluster analysis is a branch of data mining and business intelligence whose objective is to find and select homogeneous groups within a determined set of data. There are four main clustering algorithms applicable to the community detection in network.

Partitional clustering. It separates the nodes into k clusters, the number of clusters is predefined, by maximizing or minimizing a loss function based on distance between them. For example, in minimum k-clustering the algorithm tries to minimize the maximum distance between two nodes of the same cluster, in k-center instead, after the definition of a centroid, the algorithm minimizes the largest distance of the nodes from their centroid. The most famous and used partitional clustering method is k-means clustering, it identifies k clusters minimizing the sum of the squared distance between the nodes belonging to the same group. The limitation of these algorithms is the definition at the beginning of the number of clusters that the algorithm must create.

Hierarchical clustering. The idea of hierarchical clustering is to create a dendrogram, a tree where there are different levels of division, more the tree is developed more groups are created and more the nodes in the same group are similar; the number of clusters is no more an input. The first type of hierarchical approach is the agglomerative one; at the starting point each node represent a separated cluster, every step clusters with the maximum similarity index are merged. Instead the divisive algorithms are top-down hierarchical clustering approach, they start with a unique cluster that includes all the nodes, then it is iteratively divided in smaller clusters removing edges that connect low similarity vertices. The limitation of these algorithms is the difficulty to scale them up.

Spectral clustering. Spectral clustering refers to all the techniques that use the eigenvectors of the similarity matrix of the data, an input information, to divide a graph into clusters.

Graph partitioning. The starting point of this class of techniques are the number and the size of the clusters and the objective is to minimize the number of links between the identified groups. The Kernighan-Lin algorithm is a heuristic approach belonging to this category, it tries to minimize an evaluation function that represents the difference of the inter-community and intra-community links.

Modularity-based Algorithms

Modularity is a measure of the structure of a network, in particular it permits to evaluate the quality of the division in clusters; network with high modularity have dense intra-connections but sparse edges between different clusters. Modularity-based algorithms approach is to maximize modularity $Q = \sum_i (e_{ij} - {a_i}^2)$, where e_{ij} is the amount of edges that connect group i with group j. Most of modularity-based algorithms are heuristic methods.

Extremal optimization. Extremal optimization is a heuristic and iterative method. It starts dividing randomly the set of nodes in two clusters but having an equal number of nodes in the two groups, each iteration the entity having the lowest fitness value, for a vertex the fitness value is its modularity divided by its degree, is moved in the other group. The algorithm ends when does not exist node shift that increase the value of the function Q.

Spectral optimization. Differently from traditional spectral clustering algorithms, these techniques use spectral information of given similarity data matrix to optimize the modularity function through eigenvalues and eigenvectors. Moreover, as in extremal optimization the result can be further improved if some nodes are moved in other clusters to improve modularity value.

Greedy optimization. Greedy optimization can be seen as an agglomerative algorithm, step by step nodes, which their union increases the value of the modularity function, are joined. The result of the algorithm is the creation of large communities with a poor value of modularity maxima.

Simulated annealing. It is based on probabilistic theory for global optimization, it is utilized also in relation to k-means algorithms, in this case it permits not only to detect clusters maximizing the modularity function but also it permits to identify central node of each community.

Genetic algorithms. These optimization algorithms take inspiration from biological evolution process; to detect clusters, genetic algorithms try to maximize the modularity function Q of the network. Moreover, an advantage of this typology of techniques is that the number of groups to be created is not a starting requirement.

Dynamic algorithms

The type and functioning of community detection algorithms is strictly related to the topology of the network, a network could be static, its features and characteristics do not change over time, or dynamic, it evolves and changes. Most of the real-world networks are dynamic and for this reason, dynamic algorithms are fitted to analyse the development of them over time.

Random walk. Random walk has a bottom-up approach and consists in passing randomly in a graph over the nodes to join different groups forming clusters. This type of techniques can be also used in weighted graphs, in which to each edge is associated a numerical weight. In accordance with random walk close edges are likely to be in the same groups.

Synchronization. Synchronization model can be used in environment in which nodes interact each other changing the state of their neighbours. In this perspective to detect clusters an oscillator with random phase is placed at nodes, to establish to which community the single node belongs is sufficient to detect with which group it synchronizes its phase. In social media the synchronization is related to influence and opinion changing.

2.1.1.2 Overlapping Community Detection Algorithms

The case of overlapping community is now introduced by analysing some of the existing techniques for overlapping community detection. Differently from the above algorithms, in the below ones can occurs that two communities intersect and so that a node can belong simultaneously to two different groups.

Local expansion and optimization. This technique could be divided into two steps, in the first one a rank removal (RaRe) algorithm is implemented to eliminate the nodes with the high rank until small disjoint core groups are created. In the second phase a process called iterative scan (IS) starts from these cores, called seeds, and with a greedy optimization enlarges the seeds creating dense communities. The process ends when the density function:

$$f(c) = \frac{w_{in}^{c}}{w_{in}^{c} + w_{out}^{c}}$$

where w_{in} and w_{out} are the internal and external weights of the community c, cannot be further increased.

Link partitioning. In this type of algorithms, the focus is no more on nodes, but it is on edges. The clusters are formed no more with a partition of the nodes but partitioning the links between them, the partition of the network in groups can be performed with a hierarchical approach based on edge similarity. In this perspective a node is considered as overlapping if the links connected to it belong to more than one cluster.

Agent based and dynamical algorithms. The basic assumption of label propagation algorithms (LPA) is that if nodes belong to the same labels they are more likely to belong to the same cluster. In this direction the algorithm step by step defines for an entity its label of belonging on the basis of the votes of its neighbours. The main advantages of LPA are its simplicity and time efficiency that is linear with the number of links in the network. An extension of LPA for communities overlapping is COPRA (community overlap propagation algorithm), in this case to update step by step the belonging coefficients the averages of the coefficients, that each node received from neighbours, are utilized.

Fuzzy detection. Fuzzy detection are label propagation methods that permits, through the computation of belonging factor and soft membership vector, to detect overlapping clusters. It can be seen as a nonlinear algorithm of optimization that needs to satisfy some constraints. The function to be minimized is:

$$f = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (\tilde{s}_{ij} - s_{ij})^2 \quad \text{with} \quad s_{ij} = \sum_{c} a_{ic} a_{jc}$$

Where w_{ij} are the weights, s_{ij} the similarity and $\tilde{s}_{ij}$ the prior similarity and a_{ic} the fuzzy membership. The main disadvantage of this method is that the dimension of the belonging factor is an input and so must be defined before the beginning of the algorithm.

2.1.1.3 Applications

Community detection within a company have mostly implications on the marketing function and on the customer service, all the above presented techniques can be used to detect groups of people, that could be already clients of the company or even they could not know it, with similar features or behaviours helping the design of marketing or customer service strategies. Community detection applied to social media environment facilitates the process of clustering, people that belong to the same online community have a high probability to belong also to the same marketing cluster. This assumption can be used both for who is already client of the company both for who is not client. A clustering approach have a huge number of implications; the creation of personalized offers or advertising messages, the detection of groups with a high level of satisfaction with whom a company have the possibility to create a relation that increase their customer lifetime value, the chance to get insights from unsatisfied groups that can bring improvements to the customer service area or to whatever company function, finally the discovering of valuable groups of people that could be targeted to become new customers.

2.1.2 Link Prediction

After the analysis of the techniques to detect communities, the focus of the research now moves of how these communities and in general networks evolve over time. In this direction (Martìnez, V., Berzal, F., Cubero, J.C., 2016) different empirical theories have proved that discovering new relations between entities is possible analysing the topology of the network and the features of its elements. Link prediction permits to know the behaviour of the network by predicting future, missed or hidden (existing, but not observed) relations based on the characteristics of the current nodes and on actual links. This type of techniques aims at extracting implicit information of the network, at identifying non visible relationships between nodes and at predicting future behaviours and features of the network. The basis of link prediction is the empirical evidence that two elements are more likely to interact if they are similar. In most domains, the number of relevant paths between nodes could be consider as proxy of similarity between them. In the figure below a taxonomy (Martìnez, V., Berzal, F., Cubero, J.C., 2016) of link prediction techniques is presented, the focus is on methods used in undirect networks, composed only of bidirectional edges for whom is not possible to define a source and a destination node, exploiting derived topological features.

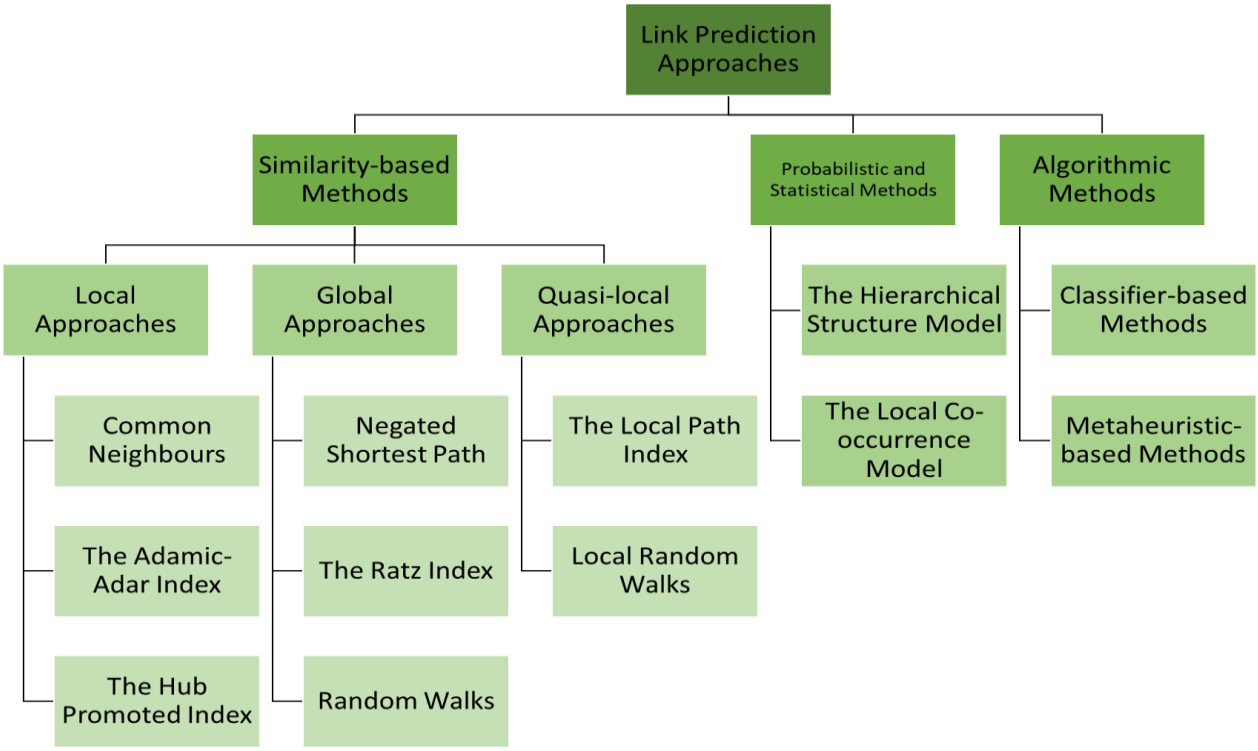

Figure 7- Link Prediction Approaches (Martìnez et al.,2016)

2.1.2.1 Similarity-Based Methods

Similarity methods stem from the principle that is more likely to detect hidden links between similar nodes, two entities are similar if they are linked to other similar entities or near according to a given distance definition. For these algorithms is essential the initial definition of a matrix to assign a similarity measure for each pair of nodes, the calculation of similarity is done for each pair with no observed links. Then pairs of node are ranked with decreasing similarity values, so the top pairs are more likely supposed to be linked with missing relations. The similarity function can vary, this bring to a big number of proposal methods.

Local approaches

Local similarity approaches utilize node neighbourhood, defined as the group of nodes connected with a relation to another node, the advantages of these methods are their rapidity, with respect to nonlocal ones, and their efficiency in very dynamic and changing networks, for example in social networks. The limitation in using local information is that similarity can be measured only for distance-two nodes, so if there is a hidden link between two nodes with a distance greater than two this is not detected. The number of local approach techniques is very high (Martìnez, V., Berzal, F., Cubero, J.C., 2016), so in this report only few examples are presented.

Common neighbours (CN). Common neighbours technique defines as similarity the number of shared neighbours between two elements, because it is more likely that two nodes meet if they have more common contacts. In this perspective the similarity function to be optimized is:

$$s(x,y) = \left|\Gamma_x \cap \Gamma_y\right|$$

Where x and y represented the two nodes and Γ_x and Γ_y are the neighbourhoods, the sets of nodes directly connected, respectively of x and y. This is the simplest local technique, but despite of that its performances in most real-world network are high, beating more complex methods.

The Adamic-Adar index (AA). Proposed by Lada Adamic and Eytan Adar, this algorithm assumes similarity related to the shared characteristics between the two nodes. Each feature is logarithmically divided by its frequency of apparition, so in this way each shared feature is penalized by its degree; if a node shares a characteristic with a lot of node its weight will be low. The similarity function is:

$$s(x,y) = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{\log|\Gamma_z|}$$

The hub promoted index (HPI). This method fits with networks in which a hierarchical structure emerged, small groups whit a huge number of internal connections isolated from the rest of the network. The objective of this technique is to avoid relations between hubs, nodes with a degree, number of links connected to it, that exceeds the average; but it promotes the creation between hubs and low-degree nodes. Below there is the used similarity function:

$$s(x,y) = \frac{\left|\Gamma_x \cap \Gamma_y\right|}{\min(|\Gamma_x|, \left|\Gamma_y\right|)}$$

Global approaches

In global approaches the focus is no more restricted only on neighbourhood and on neighbours' features, but the algorithms to calculate similarity between nodes used as input topological information about the whole network. In large and complex environment could be unfeasible considering the whole picture, this make global approaches inefficient in this context. Even for global approaches, due to the high number of possible algorithms, only a restricted number of examples are presented.

Negated shortest path (NSP). According to its name this technique requires to calculate the shortest path between two nodes, for doing this could be used the Dijkstra's algorithm that utilizes adjacency matrix, a square matrix to represent the presence or absence of edge between the nodes, and a heap data structure to define priority queue. After the calculation of the shortest path similarity function can be computed as:

$$s(x,y) = -\left|shortest\ path_{x,y}\right|$$

Even if it is a basic global approach method, negated shortest path is complex due to the time to compute the shortest path between each element.

The Katz index (KI). Katz index method sum the influence of each possible paths between two different elements, longer is a path less its weight in the sum is.

$$s(x,y) = \sum_{l=1}^{\infty} \beta^l \left| paths_{x,y}^l \right| = \sum_{l=1}^{\infty} \beta^l (A^l)_{x,y}$$

where l is the length of the paths, A is the adjacency matrix that represents the adjacency between the different nodes and the parameter β is a value between 0 and 1, introduced to damp the effect of long paths, if β is closer to 0 the weight of long paths is minimum.

Random walks (RW). Random walk in graph theory works in this way; it starts from a node, then iteratively a neighbour is randomly selected until the arrival to a destination node. The starting point of the algorithm is the definition of $\boldsymbol{p}^x(t)$ as the iterative approximation of the probability of reaching any node starting from x through a random walk:

$$\boldsymbol{p}^x(t) = M^T \boldsymbol{p}^x(t-1)$$

where M is the transition probability matrix computed starting from the adjacency matrix A normalized by rows, with $M_{i,j} = A_{i,j} / \sum_k A_{i,k}$. When, after the iterative application of the equation, the stop condition:

$$\sum_{i \in V} (\boldsymbol{p}_i^x(t) - \boldsymbol{p}_i^x(t-1))^2 < \epsilon$$

where ϵ is a number close to 0, at the end the similarity function could be expressed as $s(x,y) = \boldsymbol{p}_y^x$, the probability of reaching the node y starting from the node x with a random path.

Quasi-local approaches

Quasi-local approaches are recently emerged as a middle way between local and global approaches. Indeed, they are almost efficient as local but do not take into account only neighbours, but exploit the topological information of the whole network, even if quasi-local algorithms do not consider any paths between all the different nodes. As for the two above categories below, now a list of examples of quasi-local approaches techniques is presented.

The local path index (LPI). The idea on which local path index is based on is very similar to the one of Katz index, but being a quasi-local technique, it considers only finite numbers of path lengths. In this perspective the similarity matrix can be defined as:

$$S = \sum_{i=2}^{l} \beta^{i-2} A^i$$

where β is a damping factor, if β is close to 0 the outcomes are similar to Katz index ones, and l the maximal length, if $l = 2$ the algorithm is equal to the common neighbour technique.

Local random walks (LRW). It is very similar in the procedure to global random walks but in this case the iterations are limited by a predefined small value l. The similarity function is computed as:

$$s^{x,y}(t) = \frac{|\Gamma_x|}{2|E|}\boldsymbol{p}_y^x(t) + \frac{|\Gamma_y|}{2|E|}\boldsymbol{p}_x^y(t)$$

where $\boldsymbol{p}_y^x(t)$ is the probability vector of random walk at time t. Due to the limited number of iterations the time complexity of local random walks is less compared to the one of random walks.

2.1.2.2 Probabilistic and Statistical Methods

Probabilistic and statistical notions can be used to study and analyse network formation and evolution, in this perspective many link prediction techniques based on probabilistic and statistical concepts are emerged. Starting from the assumption that networks have a known structure; these algorithms try to model networks structure and the parameters that regulated it, through those parameters is then possible to compute the probability of creation of missed and hidden links in order to rank potential edges on the basis of their realization probability. Below there is the presentation of two main probabilistic and statistical link prediction methods.

The Hierarchical Structure Model. The approach of hierarchical structure model is to identify in the networks a hierarchical structure and to calculate the parameters that regulate it. The basic assumption of this type of algorithm is that entities with higher degree likely have a lower clustering coefficient, hub nodes have a low probability to belong to dense clusters. Given a dendrogram D, that has $|V|$ leaves and $|V| - 1$ internal nodes, if p_n is the probability of an edge between entities of both branches descending from it, the likelihood of the dendrogram D, that can change according to the choice of the internal nodes, can be computed as:

$$\mathcal{L}(D, \{p_n\}) = \prod_{n \in D} p_n^{e_n}(1 - p_n)^{l_n r_n - e_n}$$

where l_n and r_n are the numbers of leaves in the left and right subtrees. Fixed the dendrogram, the algorithm ends with the maximization of its likelihood changing the value of $\overline{p_n} = \frac{e_n}{l_n r_n}$.

The Local Co-Occurrence Model. For large dataset the previous method leads to high computational time and high complexity, to avoid these issues a scalable probabilistic technique was introduced, the local co-occurrence model; it solves the problem of

computational complexity basing is process on the local topological features of the network. The algorithm starts with the selection of a set of relevant entities for each pair of nodes called the central neighbourhood, they are the nodes in the most frequent, without cycles, paths that link the two nodes. These central neighbourhood entities are utilized to learn a Markov random field (Martìnez, V., Berzal, F., Cubero, J.C., 2016) iteratively used respecting the constraints of a set of nonderivable itemsets (NDI), their occurrence statistics cannot be inferred from other itemset patterns. The algorithm ends with the calculation of the probability of existence of each link.

2.1.2.3 Algorithmic Methods

The approaches to link prediction described in the previous sections are based on scoring non-observed links on the basis of calculation of similarity functions between edges or probability functions of the future presence of interactions between elements. Instead the methods presented now benefit also on algorithmic approaches that could include optimization and supervised learning techniques.

Classifier-Based Methods. In this case the link prediction issue is seen as a supervised learning problem, specifically a classification problem; the goal of this type of algorithm is to predict to which label, existence or absence, each possible link belongs. To train the algorithm several features of the entities can be used, from topological properties to any other link prediction measures. The main disadvantages of these algorithms are class imbalance problem, the nodes that belong to the label "absence" are really higher than the "existence" ones, and the lack of the possibility to rank the links, making their comparison harder.

Metaheuristic-Based Methods. All the techniques introduced above could be seen as heuristic because they are based on assumptions about the schema of formation of the links. But in this way the link prediction issue is simplified and in some cases the assumptions could not fit with the real status of the network. In this direction metaheuristic-based methods were introduced, they are based on the idea that multiple formation heuristics can coexist and cooperate; an optimization algorithm that considers the influence of the several predictors based on similarity and probability approaches is developed.

$$s(x,y) = \sum_{i=1}^{|w^{(u)}|} w_i^{(u)}\, s_i(x,y)$$

where w_i are the weights of each predictor and $s_i(x,y)$ are its similarity functions.

2.1.2.4 Applications

The opportunity to predict future links and connections between nodes belonging to a graph have a lot of implications for a company that can exploit Big Data gathered from social media in general, but mostly from social networks like Facebook, Twitter or Instagram. An immediate and direct implication of link prediction is trend analysis, a company, knowing

the evolution of the relations and interactions between entities in a network, can predict the future dimensional changes that a community or cluster of people can shows, starting from that information it will focuses on the ones with a greater probability to become bigger. Another important aspect for the definition of a marketing campaign, as well as the analysis of the actual situation of the influence propagation process, is the detection of future possible relation, more audience a marketing message will have better will be for the company. The possibility to conduct a trend analysis can help also to discover new tendencies spread across society permitting the company to be prepared for any kind of future changes. Link prediction have a central role also in recommendation system, it permits to improve similar user's selection leading to better recommendation results. Finally, if a company has its own community, link prediction helps to understand which are the persons with an high probability to come in contact with the company, in this way the organization can focus its energy on the easy to reach new customers.

2.1.3 Social Influence Analysis

Another issue and related methodologies regarding the analysis of the structure of social networks is social influence analysis. As (Li, K., Zhang, L., Huang, H., 2018) already said the basis of networks is the creation of relations between entities and so the exchange of information between them. Social influence analysis explores information diffusion among nodes into the graph and so the phenomenon of social influence. Social influence occurs when user's opinions, feelings and behaviours change after the interaction with another node. In social networks environment to study how this phenomenon happens has a great importance, which are the nodes that has a bigger influence on the others and if it is possible to monitor and control the influence diffusion process. There are a lot of field of application, for example in marketing in the area of viral advertising and recommendation process. In the sections below a description of the principal models and methods to analyse social influence process are presented.

2.1.3.1 Social influence analysis models

To represent social influence propagation two types of models are widely used; microscopic models and macroscopic models.

Microscopic models

Microscopic models focus on individuals and their interactions; the objective is to discover the role of these relations in the diffusion influence process, when and in which condition a node through a link can influence another neighbour. Independent cascade (IC) and linear threshold (LT), their variations and improvements are the mainly used models to figure the structure of the influence process. IC is an iterative model in which in a network represented as a graph exists a subset of activated nodes, that for example know a particular information; at each iterative step there is the probability P_{ij} that an activated node i influences its neighbour y activating it, in the example, sharing its information. In linear threshold model instead for each link between neighbours is associated a weight, the sum

of this weights for each node must be less than or equal to 1; a node i can be considered active when the sum of the weight of the links overcomes a randomly selected value chosen between 0 and 1. Many variations of these two models have been developed to improve and overcome their limitations, some take in consideration the variable of the time and the fact that relations and communications are not always synchronous, others variations introduce a third state between inactive and active, latent active state.

Macroscopic models

Macroscopic models do not take into account the individual nodes and the topological characteristics of the network; every node has the same power of influence and transmission probability; they start knowing the same information. For these reasons, the accuracy of macroscopic models is lower than microscopic ones. Macroscopic models focus on the big picture, analysing how the influence spreads into the network; in this perspective the nodes are divided into classes, the percentage of nodes in each class is determined by differential equations, and for each class an analysis of the evolution of the nodes is conducted.

2.1.3.2 Social influence analysis methods

When social influence analysis is applied to specific problems of the real world is not sufficient to model and figure the diffusion process of influence, but some interesting points and issues to be solved arise. In this direction several methods and techniques have been developed to determine a solution for each sub-problem.

Influence maximization. The objective is to maximize the diffusion of influence, a typical example of this will is viral marketing; for doing that the algorithms try to discover which is the most influential group of users to exploit them in spreading information, feelings or behaviours. Two different types of algorithm exist for influence maximization; **greedy algorithms**, they are iterative techniques, each step among the existing nodes the one with the maximum marginal gain is activated, due to the repetitions the algorithms are complex, the execution time high and so the efficiency low. **Heuristic algorithms** instead have been introduced to reduce complexity and increase efficiency. In heuristic algorithms the choice of the most influential group of nodes is simpler; each iteration a node is selected with respect to easy to compute metrics, like degree, number of edges incident to the node, or PageRank, consider the quality and number of links to reach a page. In this case the limit of the algorithms is the accuracy of the results.

Influence minimization. In influence minimization algorithms the objective is the opposite of the previous one, these techniques try to minimize the diffusion among entities searching the solution with the lowest number of nodes activated. The function to be minimized could be computed in this way:

$$D = \sigma(S|V\ D)$$

where the influence minimized is expressed by formula, with D the subset of k, given as a constant, nodes blocked, entities that cannot influence other entities. For example, these techniques are used when bad rumours are diffused in social networks.

Flow of influence. Flow of influence through pattern mining and other techniques try to monitor in a concrete way how the process diffusion works analysing the behaviours of entities and how the information spreads.

Individual influence. Individual influence focuses on the microscopic model, the objective of these techniques is, through the analysis of features of the single nodes and topological characteristics of the network, to detect which are the influencer nodes in a graph.

2.1.3.3 Applications

Influence analysis have a great importance in social media marketing in which, differently from traditional marketing channels, the main goal is not to reach the largest audience possible but to reach people with a considerable power to influence other persons. In this perspective it is essential for companies to detect how social influence propagation process develops and which are the most important nodes in the graph, the ones that have the biggest number of links with other entities. In social media environment a message becomes viral when the ones reached by the organization share the contents with their contacts and followers.

2.2 Content based analytics

Differently from structure-based techniques, content based analytics work on the data generated by the users, in social media three different typologies of content can be created; text, audio or video contents. The volume of data that everyday are posted on social network is huge, it is mainly unstructured, noisy and dynamic data. Moreover, the three types of data are really different among them and need totally different techniques to be analysed; in this paper for simplicity only text mining techniques are analysed in detail.

Text Mining

The continuous growing of the availability of text data coming from several areas, but mostly from social media, has created the need of specific advanced techniques. Indeed, text mining refers, not only to techniques used to make more efficient the access to data, the right information to the right users at the right time, but also to methods and algorithms to extract high quality information and patterns from text data. Text data needs specific techniques because of its peculiar characteristics; text data is sparse and high dimensional, the lexicon has a wide range of words but in a document only a little amount of them can be used. For this reason, text data can be represented through a sparse term-document matrix where the normalized frequency of each word in each document is indicated. The documents are also represented through bag-of-words or string of words approach, the first one does not conserve positioning information but is widely used for its simplicity and immediacy. Other more complex approaches instead represent text information semantically to perform more meaningful analysis. Before the analysis of the text data a preprocessing phase needs to be developed in order to prepare the documents to make the successive analysis faster and more effective. As it is notable in the figure below it is possible to distinguish (Irfan, R., et al., 2015) between two different preprocessing activities; feature extraction (FE) and feature selection (SE).

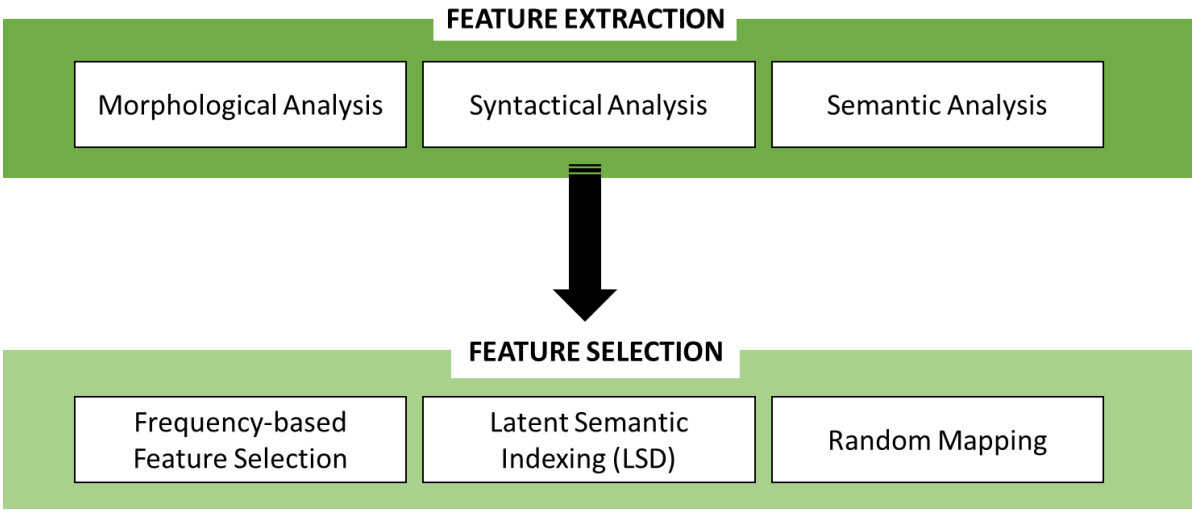

Figure 8- Preprocessing Activities (Elaboration of Irfan et al., 2015)

Feature Extraction. Feature extraction aim at making the words and sentences ready to be utilized, eliminating noisy and redundance in the data. It is composed by three different

tasks, the first one is morphological analysis (MA) that through toketization, the sequences of characters are divided in tokens (words and phrases) eliminating punctuation marks, through remove-stop-word, words such “a”,”or” and “the” are removed and through stemming-word, the words are normalized reducing each word to the root form, makes single words more easily to work on. After morphological analysis sintatical analysis (SA) can start; it is essential to make the senteces grammatically correct for a good logical meaning interpretation. SA is composed by part-of-speech tagging (POS tagging), it adds to the words its lexical class, and parsing, it consists to the creation of a tree, with a top-down or bottom-up approach, to represent and make more understandable the grammatical structure of the sentences. The last task of feature extraction is semantic analysis, it tries to understand the meaning of the text data.

Feature selection. Feature selection aim at eliminating the useless and redundant information removing the words that are no usefull and that make the analysis process slower. In this perspective in feature selection a process of scoring each word with respect to the importance that it has in the sentence and in the whole document is performed. Three different techniques for feature extraction exist; frequency based feature selection, it utilizes the normalized frequency matrix, the more a word is frequent the more important it is, latent semantic indexing (LSI), it improves the lexical matching using a semantic approach, and random mapping (RM), it permits a dimensionality reduction with the creation of a map through the contents of a large set of documents.

After this phase of preprocessing the real analysis of the text data can start; different techniques exist and each one has a specific goal. This paper presents five main areas of interest (Dang, S., Ahmad, P.H., 2014) of text mining in social media environment.

2.2.1 Information Retrieval

Information retrieval refers to methodologies and practices whose aim is identifying and selecting documents in which the needed information is present, it facilitates and makes quicker the access to information. In the context of text data information (Waseem, A., Ali, R., 2016) retrieval must face different challenges; data are unstructured, so in different formats, contents are multilingual, ambiguous, in the context of utilization the meaning of some expression can change. For these reasons two phases are needed to develop an information retrieval analysis.

- Data collection. In this preprocessing phase data are collected from different sources such Facebook, Twitter or other social media using respective application program interfaces (APIs), then the data are converted to the same format and merged, the different documents are classified with respect to their language and subsequently translated into a standard language, usually English.
- Now the information retrieval process starts with the creation of an interface to generate queries, then similarity functions like cosine similarity measure are used

to calculate the similarity between documents and the query and at last documents are visualized in order based on the score of similarity measure with the query.

Applications

Information retrieval permits a fast access to any kind of information, for this reason it has a large spectrum of applicability within a company, for any part of the organization and at any level it is essential to manage information in an efficient and effective way.

2.2.2 Information Extraction

Information extraction (IE) (Allahyari, M., 2017) refers to the methodologies and techniques to extract structured information from unstructured or semi-structured text. The goal is to obtain useful information from data text that, without being processed, have no value. Using these algorithms, it is possible to extract entities, relationships and events, so a general picture of the content of the documents could be described. For example, in the sentence "Luca and Anna are going to organize a summer party on July 14, 2019 in Milan." these techniques can extract organizers, date and location of the party.

OrganizerOf(Luca, Summer party)

OrganizerOf(Anna, Summer party)

LocationOf(Milan, Summer party)

In the following sections two fundamental tasks of information extraction are presented: name entity recognition and relation extraction.

Name Entity Recognition (NER)

Name entity recognition task is to find and extract from unstructured data text a word or a sequence of words that refer to real world entities, after being identified the entities are classified in predefined categories that could be persons, animals, dates, locations or many others by the algorithm. Using string matching with a vocabulary to make NER is not possible because, dictionaries have no always all the forms to refer to a determined entity and the meaning and semantic of a word strongly depend on the context in which it is used. (Aggarwal C.C., Zhai C.X., 2012) The first name entity recognition techniques used are rule based-approach, then with the passing of the years more complex methods based on statistical learning method have been developed. In the next sections there is an introduction to these different approaches.

Rule-based Approach

Rule-based approach is the most immediate way to perform information extraction, the algorithm starts with the manually or automated definition of a set of rules. A rule is defined by a pattern of features and by an action, if a specific word or sequence of words fit a pattern, the related action must be taken on the word or sequence of words. A rule can for example make the algorithm categorize as person entity the capitalized word.

Statistical Learning Approach

Nowadays the task of name entity recognition is analysed using statistical machine learning techniques. In this direction NER is interpreted as a sequence labelling problem in which each word in a sentence is treated as an observation that is represented with a feature vector. The objective is to select for each observation the right entity category, in sequence labelling the decision of the label is related not only with the specific observation but also with other observations and labels in the sequence. Two main types of statistical learning approach are now described.

Hidden Markov model. Hidden Markov model, to calculate the probability that an entity belongs to a particular category, considers the predicted labels of the nearby words. In accordance with a Markov process the generation of labels or observations is depended to one or few previous labels or observations (Allahyari, M., 2017). Given a sequence of labels $Y = (y_1, y_2, \dots, y_n)$ for an observations sequence $X = (x_1, x_2, \dots, x_n)$:

$$y_i \sim p(y_i|y_{i-1}) \qquad\qquad x_i \sim p(x_i|x_{i-1})$$

where y_i and x_i represent respectively a label and an observation.

Conditional random fields. Another model for sequence labeling is conditional random fields (CRFs), it is undirected graphical model whose main difference with hidden Markov model is the fact that the belonging of an observation to a specific label depends not only to the previous labels but also to the future ones. In this direction the probability of the labels' vector $\boldsymbol{y}$ for the observations' vector $\boldsymbol{x}$ to be maximized can be computed as:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(x)} \exp\left(\sum_i \sum_j \lambda_j f_j(y_i, y_{i-1}, \boldsymbol{x}, i)\right)$$

where $Z(\boldsymbol{x})$ is a normalization factor of all possible label sequences represented by the function:

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y'}} \exp\left(\sum_i \sum_j \lambda_j f_j(y'_i, y'_{i-1}, \boldsymbol{x}, i)\right)$$

Relation extraction

Entity recognition could be also considered as a pre-processing step for relation extraction task. It refers to the process of individualization and categorization of the relationships between entities, present in text documents, in predetermined labels. Several techniques to execute relation extraction exist, the most common consider this issue as a classification problem; these algorithms mostly focus on a single sentence and try to classify, if present, the (Aggarwal C.C., Zhai C.X., 2012) relations between words, previously labelled as entities. Two relation extraction techniques are briefly presented in the sections below.

Feature-based Classification

Feature-based classification, as its name suggests, treats relation extraction as a classification problem, the objects of the classification analysis are the pairs of entities that could have a relation and so should be categorized in the right label. In this perspective the algorithm could be seen as a two stages process; the first one in which possible relations between entities are detected and the second one in which the selected relations are classified. This method needs a training phase, a part of the data set is used to set the algorithm; relation mentions, annotated manually, are positive training examples, instead pairs co-occurring in the same sentence but not labelled are utilized as negative training examples.

Kernel Methods

As for feature-based algorithms also Kernel methods are based on the solution of a classification problem. In this case to perform the classification kernels or kernel functions, expressions that define (Aggarwal C.C., Zhai C.X., 2012) the inner product of two observations represented in a vector space, are used as similarity measure between the instances. Below two widely used types of kernels for relation extraction are introduced:

- Sequence-based Kernels. It is a basic kernel that takes into consideration the shortest dependency path between a pair of entities, if the paths have the same length and have in common some nodes they are considered similar, and so the relations are labelled in the same category.
- Tree-based Kernels. To measure similarity tree-based kernels focus on the share of common structures. The fundamental assumption is that relations between entities are similar if their trees share common subtree structures.

Applications

The quantity of text data present in social media is huge, starting from blogs to social networks passing through any types of social media the availability of sources of information is not comparable to the one of only five years ago and it is continuously increasing. In the social media environment a business company can discover information of any kind of argument and topic and can obtain knowledge about every aspect of the life, about entities, individuals, issues, events, topics and their attributes. This incredible source of valuable, if analysed and elaborated, information is the starting point for essential activities for an organization; trend analysis, through which the evolution both of the attitudes of customers both of the internal and external context of the company can be detected, a better knowledge about clients and about the situation around them is a key resource for competitive advantage, and insight mining, whose goal is to transform information into knowledge exploitable for making improvement at any level of the companies, from operational to managerial one. All social media, mostly social networks, permits also to know what clients, or not, think about company's brand, products and whatever, the processes and techniques that deal with the discovering of people opinions,

emotions and attitudes are opinion mining and sentiment analysis; because of they use, not only information extraction tools, but also information retrieval and classification ones, they are described at the end of the section of content based social analytics.

2.2.3 Summarization

Summarization techniques idea is to create starting from a data text a summary of it, in which the main information of the original text is present; it is a way to make immediate and quick the reading of an unstructured text data. Two different (Gandomi, A., Haider, M., 2014) approaches for text summarization exist, the extractive approach and the abstractive one. In the first one the summary is created with sentences taken from the original document, so new sentences are not built, the algorithm has no the necessity to understand the meaning of the text. Instead abstractive approach extracts key information from the document by creating new sentences and text units not present in the starting document. In this work the focus is on extractive summarization approach because fits with Big Data in social media, like blogs; because even if it is less coherent than abstractive approach is simpler and more efficient. In order to explain if an effective way how summarization process works (Aggarwal C.C., Zhai C.X., 2012) a description of three essential tasks that are performed by summarization techniques is presented.

Intermediate representation. The goal of this method is to derive from the original data text an intermediate representation that is useful to figure the general picture of the starting document in an immediate way. A widely used intermediate technique is topic representation, in which the text is represented by the topic that it treats, the topic representation approach permits to quick understand what the document is telling about.

Score sentences. Score sentences is a process through which to each sentence is associated a weight, a score that represents the importance of that sentence. For example, in topic representations the sentences with the highest score, and so the most important, are the ones that clarify in the best way the topic of the document.

Select summary sentences. After the scoring of the sentences there is the selection or the most important ones that will create the summary of the original documents. Different methods of selection exist; best n, respecting the length of the summary the combination of the best n sentences is used to create the paragraph, maximal marginal relevance, this is an iterative greedy procedure in which each step the score of the sentences is updated considering above the original score also the similarity with the already chosen ones, global selection, it is an optimization algorithm that tries to maximize the sum of the scores of the sentences, to minimize redundancy and, in some cases, to maximize coherence.

Applications

Summarization key role is to make quicker the process of learning of new knowledge, it could be considered as a preprocessing technique to prepare text data, like documents, posts or comments, to be faster elaborated and analysed. Working with a shorter text data, that

however maintains the main information of the original one, is easier and quicker both for humans both for data algorithms. In this direction summarization process is useful also in opinion mining and sentiment analysis.

2.2.4 Classification

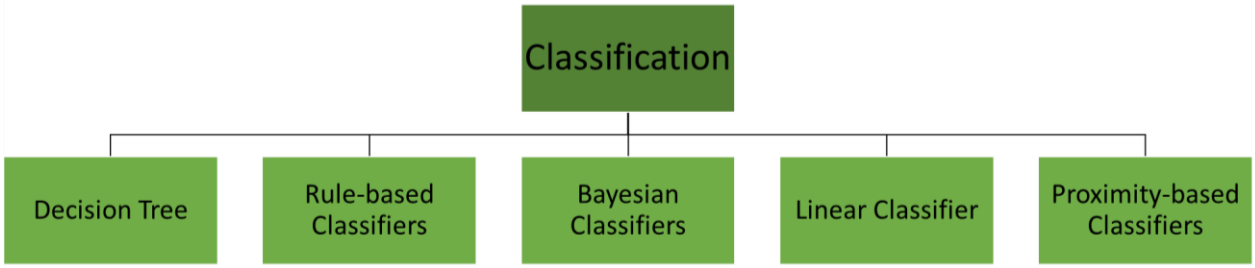

Figure 9- Classification Techniques

Classification techniques (Vercellis, C., 2009) in the field of text mining are used to classify documents in predefined classes. The goal is to create an algorithm that automatically insert new documents in the right label. The starting point for the generation of the algorithm is a set of text data D that is divided into two parts; a training set T used to derive the classification rules, it permits to identify a function that relates each documents of the set $D = \{d_1, d_2, \dots, d_n\}$ to a specific label of the set $\mathcal{L} = \{l_1, l_2, \dots, l_k\}$:

$$f: D \to \mathcal{L} \qquad\qquad f(d) = l$$

The second part V of the set, with $V = D - T$, is then utilized to verify the accuracy of the classification rules found comparing the results of the application of the function f on V with the real known labels that the documents belonging to. Several measures can be derived from the results of the test phase, after the classification algorithm is checked and validated the prediction of the new documents can start. Almost all the classification methods could be used for text data classification, modelling text as quantitative data (Aggarwal C.C., Zhai C.X., 2012) with frequencies on the word attributes permits to use on text also most of methods for quantitative data; below the relevant types of methods and some techniques for classification are presented.

Decision Trees

Decision tree refers to the creation of a dendrogram, called decision tree, starting from the training set; at each step text documents are divided in different groups in accordance with the satisfaction or not of different conditions called predicate. The whole set is recursively separated in smaller groups until the leaf node requirement of class purity are satisfied or the minimum number of entities per group is reached, at the end of this top-down algorithm relevant leaves are determined. After the training phase starts the test one that consists for each document to the development of the sequence of predicates, to check if the real class of belonging corresponds to the one predicted by the decision tree algorithm. To improve

the accuracy and efficiency of the algorithm the nodes not used to construct the dendrogram could be pruned. The main predicates used to split the set are:

- Single attribute split, the division condition is the presence or absence of a word of multiple words (even a whole sentence), at any iteration the word with the maximum level of discrimination between the different groups is selected.
- Similarity-based multi-attribute split, this time to perform the split the similarity between documents is utilized.
- Discriminant-based multi-attribute split, in this split typology discriminants, like the Fischer discriminant, are used to individualize which is the direction along which the classes are best divided.

Rule-based Classifiers

In rule-based approach (Aggarwal C.C., Zhai C.X., 2012) the data are classified in different classes according to the rules that they satisfy, a left-hand condition implies a determined right-hand class. In this perspective the training phase is essential to generate the rules, instead in the test phase the categories of the input text data are predicted through the use of rule paradigm and then compared to the real ones. In many cases the left-hand condition regards the presence or absence in text data of words; most likely the presence because, due to the fact that text data are sparse, the probability that a determined term is not present in a document is too high and so too much documents would satisfy the condition. Different measures to generate and to evaluate the quality of the rules exist, here the two basic ones are presented; they are widely used because their simplicity and intuitive nature:

- **Support** represents the absolute number of documents that are relevant for a determined rule. A rule whose condition is satisfied by many instances have a lot of importance, but this implies nothing of the strength of the rule.
- The strength of the rule can be evaluated through the **confidence**, that represents the conditional probability of belonging to a determined class due to the satisfaction of the related left-hand condition.

The decision tree approach can be assimilated to a rule-based classifier in which a specific path is seen as a rule classification; but the two approaches differ for a main characteristic, rule-based classifiers allow overlaps, instead decision tree classifiers are strictly hierarchical approach. In rule-based approach if overlaps occur, an instance satisfy two or more left-hand condition, a rank-order system based for example of confidence is used to decide which rule must be applied to the overlapping instance.

Bayesian Classifiers

Bayesian classifiers are based on the maximization of the probability that an instance belongs to a class due to the realization of other particular attribute, for this reason they belong to the family of probabilistic classification models. They, (Vercellis, C., 2009) starting from the calculation of the prior probability $P(y)$ and of the conditional probability

$P(\boldsymbol{x}|y)$, derive through the Bayes' theorem the posterior probability $P(y|\boldsymbol{x})$. In this direction the training phase is essential to calculate $P(\boldsymbol{x}|y)$, the probability, given the belonging to a specific class, of the realization of a determined features, and $P(y)$, estimated used the frequencies of the class y. At this point the Bayes' theorem is applied:

$$P(y|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y)P(y)}{\sum_{l=1}^{H} P(\boldsymbol{x}|y)P(y)} = \frac{P(\boldsymbol{x}|y)P(y)}{P(\boldsymbol{x})}$$

The process to set the algorithm ends with the definition of the target class y for the vector x through the maximization of $P(y|\boldsymbol{x})$:

$$y = \max_{y \in H} \frac{P(\boldsymbol{x}|y)P(y)}{P(\boldsymbol{x})}$$

Now the test phase and the prediction one can start.

Linear Classifiers

The objective of linear classifiers is, starting from the features (Aggarwal C.C., Zhai C.X., 2012) of the data in the set, to identify a function $p = \boldsymbol{A}\boldsymbol{x} + b$ that works as a separator between the different classes. In the function $x = (x_1, x_2, \dots, x_n)$ represents the vector of the normalized frequency of the word in the document, $A = (a_1, a_2, \dots, a_n)$, that has the same dimensionality of the feature space, is the vector of linear coefficients and b is a scalar. $p = \boldsymbol{A}\boldsymbol{x} + b$ in a discrete scenario of categorical class labels can be interpreted as a hyperplane whose aim is to divide observations into different categories. A widely used, for its simplicity and interpretability, linear classifier technique is support vectors machines (SVM), it identifies a set of observations, called support vectors, to define the predictor function and so the right position of the separating hyperplane between different labels. As best separator support vector machines technique selects the one with the largest normal distance from any of the observations, so that the margin of separation is maximum. Another support vector machines advantage is that it is robust to high dimensionality data set because it tries to identify the optimum direction of discrimination in the feature space. Moreover, this method fits text data because it is suitable to process sparse and high dimensionality data in which even if few attributes with low importance exist, they are correlated to others and usually organized in linearly separable categories.

Proximity-based Classifiers

In proximity-based classifiers, to classify the different observations, measures of the distances between observations are utilized, these methods are based on the assumption that close documents, in term of similarity measures, are more likely to belong to the same category. It is possible to utilized two different approach to perform the classification:

- The first process consists in the definition of k, predefined number that depends on the amount of observation but usually ranges between 20 and 40, nearest neighbours

of the selected test instance. The chosen class will be the one that recurs mostly among the ones of the k neighbours.

- The second one starts with a pre-processing phase in which similar observations are grouped into clusters belonging to the same class and a meta-document is associated to each group. The algorithm of k-nearest is now utilized on meta-document; to make the process more efficient summarization techniques are applied decreasing the computational complexity.

Applications

Classification process is essential for companies for the management and organization of documents and text data, a better organization permits to simplify the retrieval of information, the data can be classified for example by topics or areas of interest to streamline the process of search and of detection of the interested information. Another totally different application of classification techniques is email classification for spam filtering, spam issue in company management is a useless waste of time. Moreover classification has a key role in opinion mining and sentiment analysis, that will be described after clustering issue.

2.2.5 Clustering

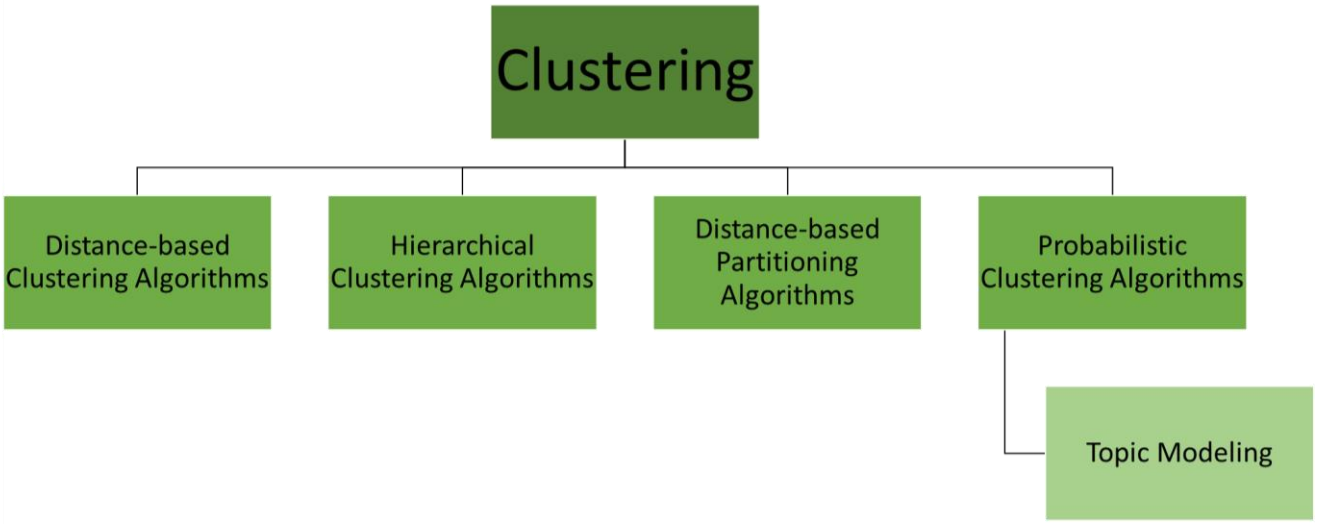

Figure 10- Clustering Techniques

Clustering techniques, traditionally used to process quantitative data, are also utilized in text mining field. The task is, through the calculation (Allahyari, M., 2017) of similarity function, to identify groups of similar observations within the set of data; in text perspective, not only documents, but also paragraphs, sentences or terms can be grouped. One of the main applications of text clustering is the organization of the different data to support and simplify the application of other techniques like information retrieval or document classification. Several text clustering techniques exist, they must take into consideration three different peculiarities and requirements of text data:

- Text data space has a huge dimensionality, but the processing data in most of the case are really sparse. Above all in the analysis of tweets or comments in social network; the words in consideration are few, but the large dimension of the vocabulary implies a huge range of possible words.
- Words belonging to the same documents are usually correlated with each other.
- Given the different number of words that each document contains, it is necessary a normalization of the data before the application of clustering techniques.

In next sections a description of some clustering algorithms types is presented.

Distance-based Clustering Algorithms

Distance-based is a type of clustering algorithm in which data are separated and grouped with respect to their closeness, that is calculated through similarity functions. Several similarity indexes exist, but (Aggarwal C.C., Zhai C.X., 2012) the most used in text mining is cosine similarity function. Given $U = (f(u_1), \dots, f(u_k))$ and $V = (f(v_1), \dots, f(v_k))$, where u and v are the normalized frequencies of the words belonging respectively to the documents U and V and $f(.)$ is a damping function, the cosine similarity between documents U and V can be computed as:

$$cosine(U,V) = \frac{\sum_{i=1}^{k} f(u_i) f(v_i)}{\sqrt{\sum_{i=1}^{k} f(u_i)^2} \sqrt{\sum_{i=1}^{k} f(v_i)^2}}$$

In clustering techniques cosine similarity can be used to compare observations as splitting condition in hierarchical clustering or can be optimize as done in partitioning algorithms.

Hierarchical Clustering Algorithms

As already said the goal of hierarchical approaches is to create a tree (dendrogram), through which is subsequently possible to perform a division in clusters at different levels, the number of clusters and the level of generalization of clusters are not predefined. Two different approaches exist; top-down or divisive one that starts from the whole data set splitting it step by step in different groups until blocking condition are satisfied, and bottom-up or agglomerative in which the starting points are each single node, represented as a leaf, and step by step they are merged to create clusters. The merging rules are based on similarity indexes like cosine function, but they differ depending which pair of the clusters the similarity is calculated on; three different merging rules are:

- Single linkage clustering, in this case as similarity between two groups is considered the greatest similarity between any pairs of nodes from the two groups. Each step the two groups of documents that have the biggest similarity are merged, so the groups that contain the pair of nodes with the largest similarity between them are merged. This technique is very efficient because is sufficient to calculate the similarity between each pair of documents, but the quality of the clustering is low because of the chaining phenomenon; similar pairs of documents can create

dissimilar clusters, in single linkage clustering if document A is similar to B and B similar to C, they are grouped together but does not mean that A is similar to C because the greatest similarity lack of transitivity.

- Group average linkage clustering, the average similarity between each of nodes in the groups is considered as similarity between the two groups. Due to the fact that it is necessary to calculate similarity between each pair of documents in the groups for all the combinations of merging, the computational time is significantly longer. It is possible to reduce it by approximating the average linkage similarity with the similarity between the mean of the two documents, this method works well for text data. The quality of the clustering is greater than the previous method because the chaining phenomenon is avoided.
- Complete linkage clustering, in this case as similarity between two clusters is considered the lowest similarity (worst-case) between any pairs of nodes belonging to the two clusters. As for group average linkage clustering the computational time is high but the chaining effect is avoided.

Distance-based Partitioning Algorithms

In distance-based partitioning approach two are the most used algorithms:

- **K-medoid clustering algorithms.** The objective of this type of algorithms is to find in the set of data nodes around which build clusters. The identification of the optimal anchors (or medoids) that can represent and be the starting point for the creation of groups is realized from the whole set of data with iterative random improvements. Each step the chosen medoids are changed with others through a random process in order to improve the quality of the division in clusters, the replacement of anchors happens if the average similarity, calculated from similarities of each nodes with its representative medoid, is increased. The optimization of this function ends when the convergence is achieved. The main disadvantage of k-medoid algorithms is the high computational time because at each iteration, a relevant number of iterations are necessary to reach the convergence, the objective function must be computed. The second one is correlated to the fact that text data are sparse and so the similarities between documents are low.
- **K-means clustering algorithms.** K-means clustering algorithms also detect representative nodes around which create clusters, however this time the process of selection is not random and the representative centroids are not necessarily chosen from the starting data set. The selection of k nodes as seeds is the simple way to start the process, then every document is added to the group represented by the seed with the closest similarity. In the following iterations the seeds are replaced with the centroids of the groups created in the previous step; this iterative process ends with the achievement of the convergence. In the case of k-means clustering algorithms the number of iterations before the convergence is smaller than k-

medoid algorithms but the disadvantage is the dependency of the results on the initial choice of the representative nodes.

Probabilistic Clustering Algorithms

Those introduced in the previous sections are all deterministic clustering techniques, challenges arise in the use of that algorithms when overlaps between clusters occur. An effective way to solve this problem is the adoption of soft clustering, the belonging of a node to a determinate group is no more deterministic but a probability of belonging or not is taken into account. A famous and widely used method related to the probabilistic document clustering (Aggarwal C.C., Zhai C.X., 2012) is topic modeling.

Topic Modeling

The approach of topic modeling consists in the creation of a generative probabilistic model that permits also the clustering of the new generated documents. In topic modeling perspective the corpus of nodes is seen as regulated by hidden random variables, the goal of the algorithm is to estimate the parameters of the function between the corpus and the random variables. Above the random variables assumptions other two important assumptions exist:

- Every document n present in the set has a probability to belong to one of the topics k, this means that a single node can belong at the same time to different groups and so that a single document can talks about different topics. $P(T_j|D_i)$, where D_i represents the document i and T_j the topic j, indicates the probability of D_i to belong to T_j, it is comparable to a cluster membership probability. One of the main results of topic modeling algorithms is to compute the value of $P(T_j|D_i)$ using as input the predetermined number k of topics.
- Each topic is also related to the terms that recur in it, $P(t_l|T_j)$ is the probability that the term t appears in the topic T; it is another essential parameter to be calculated in topic modeling approach.

Topic modeling process is an iterative one that ends when convergence is reached; the outcome is a soft division of the documents in topics that can be identified as clusters.

Applications

Clustering process has a relevant role in the organization of documents and text data for a better management of the information across the industries, knowing how to search information is a huge advantage in term of time and cost. Another task for which clustering techniques are important is clustering of clients' messages, clusters of similar messages could represent clusters of similar customers, information that could be important for a company.

2.2.6 Opinion mining

The term opinion mining refers to all the activities and practices aim at extracting and studying sentiments, opinions, attitudes and emotions from data text that could be long documents or brief posts or comments. Due to the nature of social media, an environment where users can create and share its ideas and opinions, opinion mining is really useful and valuable in this area of interest. Opinion mining, also called sentiment analysis, review mining or appraisal extraction, is a multi-faceted problem (Ravi K., Ravi V., 2015) because various steps are needed to perform the analysis, for this reason it comprises activities linked to information retrieval, information extraction and classification. There is now a description of how this process of detection, extraction and classification of opinions is articulated.

Subjectivity classification

The first task to be taken into consideration is subjectivity classification, the detection of subjective considerations that must be distinguished from objective entities. In order to accomplish this task, techniques for the recognition of opinion-oriented language from objective language are used. This is a complex issue, it has been performed using machine learning as well as lexicon based approach. At this step is also important to define the opinion holder, it is the entity that gives a determined subjective opinion on an object, the entity on which an appraisal is expressed by a user, and an opinion, it is an attitude, emotion or sentiment toward the object by the opinion holder.

Sentiment classification

After being detected, the sentiment words must be divided in classes with regard to their orientation, the classes could be two, positive or negative opinion, or more. This classification process permits to immediately divide positive emotions and appraisals from the negative ones, making successive analysis easier to be performed. Intermediary classes like neutral one or different degrees of like or dislike toward an object could be created. Moreover this step of opinion mining could be performed through machine learning, with the adoption of the above described techniques like decision trees or Bayesian classifiers, or lexicon based approach. The second group expected the utilization of two approaches; dictionary based approach, that uses an existing dictionary, created in turn with or without ontology, where words meaning opinions and positive or negative emotions are listed, and corpus based approach, that to classify the sentiments exploits the probability of occurrence of an opinion in conjunction with positive or negative set of words. Sentiment classification (Ravi K., Ravi V., 2015) is divided into four sub-tasks; **polarity determination**, necessary to understand the polarity of the sentence, if it expresses positive, negative or a neutral emotion toward the object, **vagueness resolution in opinionated text**, in social media the way, style and language used to interact is loose and many times sarcastic, these could bring to vagueness and ambiguity that need to fast detected and eliminated, **multi-lingual and cross-lingual sentiment analysis**, it is another issue that can emerge in sentiment analysis process, different languages have different ways and possibilities to express opinions and

sentiments, to solve this problem both lexicon-base approach and corpus based one are utilized, and finally **cross-domain sentiment classification**, it is also challenging to work within the same analysis with different domains because of the different levels of subjectivity.

Review usefulness measurement

In social media environment the concept of influence propagation is essential, in some cases, for example in promoting a new product, organizations could create false reviews to push an idea or emotions in the mind of clients, in this direction is very important to review the quality and so the usefulness of the content present in the sentences taken into consideration in the sentiment analysis.

Opinion spam detection

Another aspect of the same problem is spam detection, exist people that in exchange of some returns decide to create and post fake opinion and appraisal. This could create huge problems to sentiment analysis, that must absolutely detect and remove these trivial reviews that can distort the results of the analysis and so the insight coming there.

Lexica and corpora creation

In dictionary based approach vocabulary has a key role, it is the collection of words that represent a sentiment or opinion, each term is associated to sentiment polarity and a strength value to evaluate its expression power. The vocabulary is created starting from a group of words, called seed words, and enlarging this groups with the synonyms and antonyms of the already present words. The process continues until the extension of the list is stopped. This task could follow two different approaches; non-ontology based approach, lexica is created starting from machine learning, lexicon based and hybrid approach, and ontology approach, it provides an explicit specification of the concepts.

Opinion feature extraction and product aspects extraction

An object usually is characterized by more than one aspect, for example to describe a product is not sufficient the analysis of a single feature because it is formed by multiple ones, for this reason to have a complete understanding of what people think about a determined object is necessary to collect appraisals and sentiments of not only the whole object but also of every single aspect that composes it. In this direction it is important that in sentiment analysis the search of information and reviews is not limited to a single domain or area, must it should conduct a across domain analysis.

Applications

For a company knowing what customers and stakeholders think about any possible object related to the firm, from the product features to the quality of customers service, is really valuable. Indeed this key resource has a lot of applications. Sentiment analysis is useful at any level of the marketing function, for example for marketing intelligence and the realization of ad hoc marketing campaign. It permits to evaluate the customers satisfaction

to understand where it is necessary to set in with improvements and mainly which are the groups of clients that have a large probability of churn. In production management opinion mining is utilized to predict the demand of products and so to facilitate production and procurement plans. Among the company's aspects that could be improved with the help of information taken by sentiment analysis are the quality and the recommender system, both elements that can improve the satisfaction of the clients and so increase the sales for the company. Moreover analysing attitudes and behaviours of people through opinion mining permits a better development of trend analysis, understanding how the field in which the company competes is going to evolve, and insight mining, having a clear ideas of what people perceive and feel can bring the company to generate new ideas for improvements and innovations.

2.3 Descriptive analytics and social media metrics

In the previous sections the main techniques used for the analysis of Big Data in the social media environment and its applications are be described, now another basic utilization field of social data and social descriptive analytics is introduced: social media metrics. In the business environment for the companies that need to interact and keep in contact with the clients is essential to have a social media channel and to manage it in the right way, mainly for marketing scope but also in order to take the right decisions with respect to what the customers want. In this perspective to control and monitor through social media metrics, calculated with the help of descriptive analytics, how the firm manages its social channel and how it can be improved is fundamental. First of all it is important to clarify what the term metric means; a metric can be defined as a measurement system used to quantify static or dynamic characteristics (Peters K. et al., 2013), characteristics is a general term that represents all the features, processes, states, trends and evolutions that can be measured in a system that in this case is a social media environment. As just said a metric can calculate variable states, that vary over time, but also stochastic states, that are linked with a probability of occurrence or with a variance degree. Social media metrics, like any others, to be valid requires a theoretical grounding, a metric is considered valid only within a predefined domain and under certain conditions, a diagnostic nature and over time it must be reliable and credible. Moreover to be considered useful a social media metric should be linked to financial performances or to marketing actions, indeed only in this case a metric measurement could bring to an improvement in the management of the social media channel for a firm. Several social media metrics exist, in the table below, introduced by (Hoffman D.L., Fodor M., 2010), the most frequent ones are presented, divided according to the type of social media in which are applied and according to the performance that improve, awareness, engagement or word of mouth.

SOCIAL MEDIA	**AWARENESS**	**ENGAGEMENT**	**WORD OF MOUTH**
Blogs	• number of unique visits • number of return visits • number of times bookmarked • search ranking	• number of members • number of RSS feed subscribers • number of comments • amount of user-generated content • average length of time on site	• number of refences to blog in other media (online/offline) • number of reblogs • number of times badge displayed on other sites • number of “likes”
Microblogging	• number of tweets about the brand • valence of tweets +/− • number of followers	• number of followers • number of @replies	• number of retweets

Social Networks	• number of members/fans • number of installs of applications • number of impressions • number of reviews/ratings and valence +/−	• number of comments • number of active users • number of "likes" on friends' feeds • number of user-generated items • rate of activity	• frequency of appearances in timeline of friends • number of posts on wall • number of reposts/shares • number of responses to friend referral invites
Media Sharing	• number of views of video/photo • valence of video/photo ratings +/−	• number of replies • number of page views • number of comments • number of subscribers	• number of embeddings • number of incoming links • number of "likes"
Social Bookmarking	• number of tags	• number of followers	• number of additional taggers
Reviews	• number of reviews posted • valence of reviews • number and valence of other users' responses to reviews (+/−) • number of wish list adds • number of times product included in users' lists	• length of reviews • relevance of reviews • valence of other users' ratings of reviews • number of wish list adds • overall number of reviewer rating scores entered • average reviewer rating score	• number of reviews posted • valence of reviews • number and valence of other users' responses to reviews (+/−) • number of visits to review site page • number of times product included in users' lists
Forums	• number of page views • number of visits • valence of posted content +/-	• number of relevant topics/threads • number of individual replies • number of sign-ups	• incoming links • citations in other sites • tagging in social bookmarking • number of "likes

Table 2 - Social Media Metrics

All of these metrics taken alone lose of significance, they are not so meaningful if not related to other metrics, in this direction to introduce the concept of dashboard is important. A dashboard is a visualization tool whose scope is to collect a relatively small number of key metrics and related performances drivers that permit to represent with a double view, short and long term views, the actual situation and the evolution trends of the social media environment. The dashboards are essential tools to monitor and improve the overall performances of the social media channel. Its goal is to link marketing inputs to financial outcomes, that correspond to the final object of a company, through metrics. Before a description of the process to manage a social media channel and guidelines to create an

effective and efficient social media dashboard is important to understand through which point of view the social environment should be seen to conduct on it a useful measurement and monitoring process. For this reason (Peters K. et al., 2013) the *stimuli* → *organism* → *response* framework is now introduced, it permits to visualize the perspective and logic that a firm should have to manage social media in the best way possible. S-O-R paradigm starts from *stimuli*, the marketing function gives inputs through information and advertising to the social networks, the social network and all the entities that belong to it represent the *organism*, which if received the right stimuli produces a *response*, the managerial outcomes that can be measured through social metrics. In this perspective in social media channels it is not only important to detect how many people are reached, as it happens for traditional channels, but is also essential to know how these entities react to the stimuli and how they behave towards the firm; in social media environment the focus must be kept on the interaction and relation with the customers.

Process to measure and manage social media channels

The calculation of the metrics starting from the data taken by social media is one of the main tasks of a bigger process that permits to measure the performances and to manage a social media channel, the computation of metrics could result useless without some previous and post activities and practices. Below there is a brief description of the nine steps presented in the paper (Sterne J., 2010).

1. **Identifying goals**
   First of all it is essential to set objectives, what is the reason because of the firm utilizes a social media channel, what goals it wants to achieve and through which direction. The three basic goals of a business organizations are increasing revenues, reducing costs or increasing customers satisfaction, after the selection of the central objective a set of subobjectives, that are directed connected with the central one, are defined in order to create a strategy of action. In this way all the successive analysis and measurements should be aligned whit the main direction.
2. **Reaching audience**
   The first step of an effective management of social media channel is reaching the targeted audience, in this direction Big Data about the traffic, reach, awareness and frequency are necessary to understand if the message of the company is effectively reaching a substantial number of people. In social media environment, differently from traditional communication channels, it is not sufficient to monitor the number of people reached but it is necessary to monitor the behaviours and the reactions of the users induced by the stimuli of the firm.
3. **Identifying influence**
   Reaching a large number of people is not sufficient, it is important to select the right audience, that is interested by the message and willing to share the content. Another point to take in consideration in the influence propagation process within the graph are the nodes/entities that have more power and capacity to influence a large number

of people, in every network exist some nodes that, with respect to others, are considered important and followed mostly, having so a huge possibility to spread what they publish.

4. **Recognizing the sentiment**

   At this point it is significant to detect what type of reactions and behaviours the customers have in response to the stimuli of the firm. Through text mining it is possible to develop sentiment analysis and opinion mining to understand if the clients appreciate the messages driven by the firm through social media channel. Knowing what the clients think about companies, brands or products is extremely valuable to generate insights that permit firms to maintain the right direction satisfying its customers in the best way possible.

5. **Triggering action**

   When the clients have listened, repeated and liked the message it is time to convert their sentiments in real actions, monitoring and measuring if effectively the customers react interacting with the company is another step in which the calculation of metrics through the Big Data gathered in social media environment has a key role. Without the collection and analysis of data it is impossible to track the behaviours of the customers towards the company, for example data permits to know if and when people click to company's website pages or if they engage through the mobile application with the company.

6. **Hearing the conversation**

   After a company succeed in creating relations and in interacting with its clients, it is fundamental that it is concerned about measuring and monitoring its capacity to hear what the customers want to say. Social media environment through its continuous creation of data is the perfect tool to know what people think and say and to monitor if the company effectively exploits this valuable generation of feedbacks.

7. **Driving business outcomes**

   Now it is time to evaluate how the marketing campaign is going, if the efforts done to reach and engage the customers have an impact on the big three goals set at the beginning of the process. If the campaign does not increase the revenues, does not reduce the costs and does not increase the customer satisfaction it is necessary to rethink the strategy and to bring changes to what done till now. To relate directly the social marketing effort to one of the big three objectives is almost impossible, for this reason some key performance indicators and metrics that have an intermediary role exist and need to be measured and monitored.

8. **Convincing colleagues**

   The availability of a tool as dashboard, that permits to measure in real time the performances of marketing actions, make company's managers possible to show and attest to senior managers that their ideas and efforts bring to evident results. Moreover changing the perspective, a so efficient measurement tool permits to

senior managers to control the performances of their subjects and to monitor if they are going in the right direction.

9. **Seeing the future**
   A performance measurement tool can be used also as generator of insights for future changing, knowing the present permits to better understand the future. The constant and intense analysis of social media can make feasible to predict and anticipate possible insights and trends that can be very useful for the company.

Guidelines for an effective dashboard

As easy notable in the previous section, the exploitation of data to manage the social media channel through the calculation of metrics has a lot of importance within an organization. But, due to the huge number of existing metrics and due to the differences between companies in organizational goals, structure and social media selection, a standard method and a list of metrics to be used does not exist, each company should develop its measurement system. Moreover to utilize metrics is necessary the creation of an effective dashboard. For these reasons a list of practices and behaviours (Peters K. et al., 2013) to create a dashboard and so to exploit in an effective way the potentialities of social metrics is developed. The following advices fit with social metrics and highlight the different approach required to manage social media environment compared to the one used in the management of traditional marketing channels.

Transition from control to influence. Due to the egalitarian nature of social media a brand or a company are only a little piece of the whole network, they have no, like happens in the traditional marketing channels, a privileged position that gives them the power and the possibility to impose and control the recipients of their messages. For example managers still can post contents on Facebook, like videos or comments, but it is not sure that the people who notice those contents really read or watch them, so there is no the certainty that the messages are absorbed by someone. The contents to be consumed need to be interesting for readers. Another important aspect, peculiar of social media environment, is that the messages and contents are really distributed only if they reach influential users, users that are followed by a lot of people and so have a high possibility to spread what they publish.

Shift from states and means to processes and distributions. Social media and mostly social networks are dynamic system in continuous evolution. For this reason it is not so important, as happens for traditional channel, to calculate through metrics static features, but the main goal of dashboards is to focus on metrics that have the possibility to monitor the incessant changing over time of the network features. In social media channel is very important not only the snapshot of the actual situation but mostly the trends of this situation in the next future.

Shift from convergence to divergence. In social media channel an organization has a higher possibility to come in contact with negative feedbacks or opinions compared to how it happens in traditional channels. Moreover negative feedbacks can be evaluated in a

deeper way, it is possible to detect and know who writes it, the context and the period of time in which it is posted. These negative comments and evaluations are valuable tools for the organizations, indeed starting from them the companies can notice what does not work well and can fit it.

Shift from quantity to quality. Representing a social network like a graph, it is notable that there are some nodes that, due to the fact that have a lot of edges and links with other nodes, have a huge importance because they can influence a large number of people. In this perspective it is important that the metrics in the dashboards focus not only on the number of people reached but mostly on the quality of the people reached in the sense of possibility to influence other users. A small number of very influential users could be more attractive to be targeted than a big number of low influential users.

Leverage transparency and feedback-loops on metrics. In social media data the number of likes or comments to a post are public, so everyone can analyse and interpret them. The metrics that a company can measure to monitor its progress and trends in social media channel could be measured by everyone. The consequence of this transparency is that users, if for different reasons consider some metrics valuable, can try to modify with some actions the value of those metrics. In this direction a company should try to detect and filter these distorting behaviours.

Balance the metrics. A way to solve the problem of metric gaming, users that modify for their advantage the value of metrics, is the balance of metrics. In social media it is useless and sometimes harmful to focus only on one or few metrics, it is essential to balance the metrics, a key metric often need to be accompanied by other ones as a counterforce that keep it in balance to a good interpretation of the underlying phenomenon. As seen before quantity metrics must be balanced by quality ones; static metrics by dynamic ones, the bases of interest for calculation can change and so new metrics are needed.

Cover general to specific. A good management of social media channel must be at the same time general and specific. A manager should have a comprehensive vision of the social environment but need also to focus on the particular elements and metrics that help in understanding the overall situation. Dashboard helps in apply this approach, it is general because go across all the different social media, but in the same time through determined metrics can deepen the analysis on a specific aspect of a social platform.

Shift from urgency to importance. Social media have a dynamic, as a living organism, nature, unexpected and rapidly changes are the normality, for this reason a good social media dashboard as well as needs to manage and take into consideration metrics to monitor the dynamical evolution of the system, it must detect which are the critical key metrics that, having more importance than the others, requires to be continuously checked and controlled.

Balance theory and pragmatism. All these theoretical advices should be implemented and put into practice, to succeed in social media environment theory and practice must collaborate and converge in the same direction; implementing a metric is not enough, due to the dynamic nature of social media, it must be continuously supported by theory.

3.Conclusion

In the last decades the number of available data sources is increased a lot; every day through internet of things, sensor networks, and devices connected to internet, such as smartphones, personal computers and tablets, a huge amount of data, not even imaginable ten years ago, is generated. This data for its high volume, variety, several types of data are available, and velocity, the rate of generation of data is extremely high, is defined as Big Data. One of the main platforms that permits to every one to create and share their data and contents is represented by social media, there are different types of social media, each one with its peculiarities and that satisfies different needs of the users, but all are characterized by the possibility to interact with other persons sharing ideas, opinions and whatever type of content, text, audio or video contents. In this thesis a wide description of the techniques and practices that a company should adopt to obtain value from this huge amount of data is presented. One of the main concepts that in this work is stressed is that data, without the specific practices and analysis, are useless and has no value, for this reason is essential for a company to be prepared and organized in the management of data to extract from it valuable insights for decision making and for a various number of applications. This thesis divides data analytics, the processes and techniques for analysing data, in three macro groups. The first one presented is structure based analytics, according to these techniques, social media, mainly social networks, are seen like a graph formed by entities connected one with the others through links, called edges; the links can represent for example an interaction between two friends of Facebook. In this perspective the main techniques used are three; community detection, useful to discover and monitor groups that develop within the network, link prediction, to predict the links and so interactions that in the future have a high probability to appear, and social influence analysis, essential to understand how the influence propagation process evolves and which are the more influential entities. The second macro group of techniques taken into consideration are content based analytics, they refer to the analysis of text data, because of, for reason of simplicity, audio and video mining have not been considered. Text analytics are divided into six tasks, the first one is information retrieval, it refers to the techniques necessary for the individualization and selection of documents, posts, comments or whatever text data in which the needed information is present. The second step is information extraction, that permits the detection and separation from the rest of the text of information, for example entities' name or relations between them. Another important task linked to text analysis is summarization, through the elaboration of documents or posts is possible to obtain a summary of them. Moreover machine learning techniques like classification, a supervised learning technique for the prediction of the category of belonging of future observations, and clustering, an unsupervised learning technique for the creation of groups of similar observations, can be used in the text analysis field. The last content based analytic taken into consideration is opinion mining, it refers to the understanding of people's sentiments, appraisals or attitudes through the analysis of text data. The last group of social analytics described in this thesis

are descriptive analytics whose objective is the calculation of metrics useful for example in the management and monitoring of a social marketing campaign. Finally another interesting point that emerges from this script are the several applications that are enabled only with the adoption of the above mentioned techniques.

4.List of References

1. Cavanillas José Maria, Curry Edward, 2016 “New Horizons for a Data-Driven Economy”, “SpringerOpen”

2. Patgiri Ripon, Ahmed Arif, 2016 “Big Data: The V’s of the Game Changer Paradigm”, “978-1-5090-4297-5/16 © 2016 IEEE”

3. Zeng Jing, Glaister Keith W., 2017 “Value creation from big data: Looking inside the black box”, “Strategic Organization 2018, Vol. 16(2) 105 –140”

4. Miller H. Gilbert, Mork Peter, 2013 “From Data to Decisions: A Value Chain for Big Data”, “1520-9202/13 © 2013 IEEE”

5. Chen Hsinchun et al., 2012 “Business Intelligence and Analytics: From Big Data to Big Impact”, “MIS Quarterly Vol. 36 No. 4”

6. Zeng Daniel, 2010, “Social Media Analytics and Intelligence”, “IEEE Computer Society”

7. Zarrella Dan, 2009, “The Social Media Marketing Book”, “O’Reilly”

8. Bedi Punam, Sharma Chhavi, 2016, “Community detection in social networks”, “© 2016 John Wiley & Sons, Ltd”

9. Javed Muhammad Aqib, Younis Muhammad Shahzad, Latif Siddique, Qadir Junaid, Baig Adeel, 2017 “Community detection in networks: A multidisciplinary review”, “1084-8045/© 2018 Elsevier Ltd”

10. Martìnez Vìctor, Berzal Fernando, Cubero Juan-Carlos, 2016, “A Survey of Link Prediction in Complex Networks”, “ACM Computer Surveys, Vol. 49, No. 4, Article 69”

11. Li Kan, Zhang Lin, Huang Heyan, 2018 “Social influence analysis: models, methods and evaluation”, “Elsevier LTD”

12. Irfan Rizwana et al., 2015, “A survey on text mining in social networks”, “The Knowledge Engineering Review, Vol. 30:2, 157–170”

13. Dang Shilpa, Ahmad Peerzada Hamid, 2014, “Text Mining: Techniques and its Application”, ”IJETI International Journal of Engineering & Technology Innovations, Vol. 1 Issue 4”

14. Allahyari Mehdi, 2017 "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques”, “arXiv”

15. Aggarwal Charu C., Zhai ChengXiang, 2012, “Mining Text Data”, “Springer”

16. Vercellis Carlo, 2009, ““Business Intelligence: Data Mining and Optimization for Decision Making”, “© 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-51138-1”

17. Gandomi Amir, Haider Murtaza, 2014 “Beyond the hype: Big Data concepts, methods and analytics”, “Internal Journal of Information Management”

18. Ahmad Waseem, Rashid Ali, 2016, “Information Retrieval from Social Networks: A Survey”, “3rd Int’l Conf. on Recent Advances in Information Technology | RAIT-2016 |”

19. Peters Kay et al., 2013 “Social Media Metrics – A Framework and Guideline for Managing Social Media”, “Journal of Interactive Marketing 27 (2013) 281 –298”

20. Sterne Jim, 2010, “Social Media Metrics – How to measure and optimize your marketing investment”, “Wiley, John Wiley and Sons, Inc.”

21. Ravi Kumar, Ravi Vadlamani, 2015, “A survey on opinion mining and sentiment analysis: Tasks, approaches and applications”, “0950-7051/© 2015 Elsevier B.V.”

22. Hoffman Donna L., Fodor Marek, 2010, “Can You Measure the ROI of Your Social Media Marketing?”, “MITSloan Management Review, FALL 2010, Vol. 52, No.1, SMR363”

23. BDVA, 2016, “Big Data Value Strategic Research and Innovation Agenda, Version 2.0 January 2016”, “available at: http://www.bdva.eu/downloads”

Web Site:

1. http://www.sosprivacy.it/2017/12/big-data-il-90-dei-dati-disponibili-nel-mondo-sono-stati-prodotti-negli-ultimi-due-anni/#.XGKBIzNKjIV
2. https://www.tomshw.it/altro/youtube-400-ore-di-video-caricati-ogni-minuto-vi-pare-possibile-il-monitoraggio-copyright-voluto-dalla-ue/
3. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/