Social Media Analytics in the entertainment industry

Granis Charilaos
SID: 3301140005

SCHOOL OF SCIENCE & TECHNOLOGY
A thesis submitted for the degree of
*Master of Science (MSc) in Information and Communication Systems*

DECEMBER 2016
THESSALONIKI – GREECE
Social Media Analytics in the entertainment industry

Granis Charilaos

SID: 3301140005

Supervisor: Prof. Pramatari
Member: Mrs Fraidaki
Member: Prof. Theotokis

SCHOOL OF SCIENCE & TECHNOLOGY
A thesis submitted for the degree of
Master of Science (MSc) in Information and Communication Systems

DECEMBER 2016
THESSALONIKI – GREECE
Abstract

This dissertation was written as a part of the MSc in ICT Systems at the International Hellenic University. The goal of this dissertation was to conduct a research in the entertainment sector based on actual data, obtained by different types of social media with the use of open source tools such as gephi and the use of analytics in network theory.

Initially, the potential and the challenges of social media analytics are stated and then the way the Media & Entertainment industry uses insights from social data today. A general methodology of harnessing social data is described based on a generic workflow that the industry follows today. Afterwards, a research is being made for the available tools for data analytics, both open source and commercial. An overview of Gephi tool follows and finally a suggested methodology based on two case studies is proposed. The thesis ends with the possibilities for future development and the conclusions.

At the outset, I would like to thank my instructor and guide Dr.K.Fraidaki for guiding and mentoring me throughout my project.

Grinis Charilaos
23/12/2016
## Contents

ABSTRACT .....................................................................................................................3

CONTENTS .....................................................................................................................4

1 INTRODUCTION .......................................................................................................6

2 LITERATURE REVIEW ............................................................................................7

3 PROBLEM DEFINITION ..........................................................................................10
   3.1 THE PROBLEM .................................................................................................10
   3.2 WHY BIG DATA ARE IMPORTANT ......................................................................10
   3.3 THE SOLUTION .................................................................................................11

4 POTENTIAL OF SOCIAL MEDIA ANALYTICS ..........................................................13

5 ANALYTICS CHALLENGES ....................................................................................16

6 HOW THE MEDIA & ENTERTAINMENT INDUSTRY USES INSIGHTS FROM SOCIAL DATA .................................................................18
   6.1 HOW THEY EMPLOY BIG DATA ANALYTICS ..................................................18
   6.2 DETECTING EMERGING TOPICS ....................................................................20

7 HARNESSING SOCIAL DATA TO BUILD ACTIONABLE INSIGHT .................23
   7.1 METHODOLOGY ..............................................................................................23
     7.1.1 Listening .....................................................................................................23
     7.1.2 Refining .....................................................................................................24
     7.1.3 Workflow and integration ...........................................................................24
     7.1.4 Strategy ....................................................................................................24

8 SMA PROCEDURE AND TECHNICAL WORKFLOW .............................................26
   8.1 METHODOLOGY ..............................................................................................26
   8.2 FINDING THE RELEVANT BLOGS AND POSTS ............................................27
   8.3 SENTIMENT DETECTION .................................................................................28
   8.4 COLLABORATIVE FILTERING ........................................................................28
   8.5 MEASURING INFLUENCE AND AUTHORITY ..................................................28
   8.6 NETWORK VISUALIZATION .............................................................................29
8.7 **IMPORTING GRAPH EXCHANGE FORMATS**.........................................................30
8.8 **EXPLORING NETWORKS** ........................................................................30

9 **ANALYSIS TOOLS AND LIBRARIES - TECHNIQUES** ..............................34

9.1 **TOOLS AND PLATFORMS** ........................................................................34
  9.1.1 *Scientific programming tools* .................................................................36
  9.1.2 *Business Toolkits* ..................................................................................36
  9.1.3 *Social media monitoring tools* ...............................................................36
  9.1.4 *Text analysis tools* ................................................................................37
  9.1.5 *Data visualization tools* ........................................................................37
  9.1.6 *Social media analytics platforms* ............................................................37

9.2 **PLATFORM COMPONENTS** .....................................................................37

9.3 **TERMINOLOGY** .......................................................................................38

9.4 **RESEARCH CHALLENGES** .................................................................38

9.5 **TECHNIQUES** .......................................................................................39
  9.5.1 *Social media data* .................................................................................39
  9.5.2 *Social media providers* .........................................................................40
  9.5.3 *Text cleaning and storing* .....................................................................43

10 **GEPHI GRAPH ANALYSIS SOFTWARE** .................................................46
  10.1 **OVERVIEW - CAPABILITIES** ...............................................................46

11 **CASE STUDIES – ACTUAL DATA** ............................................................55
  11.1 **CASE STUDIES** ....................................................................................55
  11.2 **DATA FORMATS** ..................................................................................57
  11.3 **WORKFLOW WITH GEPHI (CASE STUDY 1)** ....................................60
  11.4 **WORKFLOW WITH GEPHI (CASE STUDY 2)** ....................................71

12 **FUTURE DEVELOPMENT** ........................................................................81

13 **CONCLUSIONS** .......................................................................................83

BIBLIOGRAPHY - REFERENCES ..................................................................85

APPENDIX .....................................................................................................87
# Introduction

Social Media play a vital role in the entertainment industry and the communication of the varying stakeholders takes different forms and instances. Most of entertainment organizations in various entertainment sectors like cinema, music industry, sports and other, utilize not only traditional mediums like television, radio, publications to reach their stakeholders, but also online methods to reach possible customers. During the last years social media has grown with great rhythm in many industries and has become extremely popular especially in the entertainment industry. Huge percentage of online users are using social networking sites. Platforms like Facebook, Twitter, Instagram, Youtube, Google+, Linkedin are being extensively used and they are efficient in reaching many audiences with great speed and lot of content offering the capability of interaction. These facts give the opportunity to the users to connect with the each other, build a brand, create customer relationship management and manage their reputation.

The purpose of this study is to conduct a research in the entertainment sector based on the existing literature and on actual data obtained by different types of social media in order to identify how the industry uses insights from social data and how it employs big data analytics. The analytics potential and the challenges are going to be mentioned and described as well as the available procedures and workflows and the available commercial and open source software tools that are needed in order to implement such analysis. The tool that is going to be used in the specific research is going to be Gephi Graph Analysis Software, an open source package which will help to explore and analyze social networks. The research will be based on general theoretical facts and on specific case study which will be presented in the following chapters.

The study will begin with a review of relevant literature in entertainment communication integrating social media. Then the methodology of the study will be discussed followed by the theoretical findings and the practical research. Finally, the future development and the conclusions that can be extracted from the findings will be discussed and a general guide of best practices on utilization of social media in communication will be proposed.
2 Literature review

Prem Melville, Vikas Sindhwani and Richard D. Lawrence are mentioning the power of the blogosphere for marketing purposes [5]. The rise of the created blogs has given the power to the users to influence the public opinion about products, services etc. Eventually, this influences also the profitability of the businesses. This means that organizations need to take this in mind and extract insights from these blogs. This can be done through Social Media Analytics. Seeking relevant topics, measuring the influence and authority of the bloggers and detecting their sentiment about specific topics are just the basic requirements for having successful marketing insight.

The Next Generation Of Analytics-Based Marketing Seeks Insights From Blogs [10] paper also demonstrates the power of analyzing blogs in marketing. Following almost the same methodology it provides an approach for the automated analysis of blogs and related social media. Again, measuring relevance, influence, authority and sentiment helps to find trends that exist in the various discussions. This is also the target for social media analytics, to automate the process of detecting patterns that emerge in the various blogs.

Suzanne Clayton in the The New Digital Supply Chain journal [7] states that the successful analytics are implemented through the gathering not only of social media data, but also of data from other resources like emails, the web interactions, POS data and other digital sources. Only in this case the appliance of analytics will gain insights about the business. With the help of analytics businesses can create more personalized marketing communications and can understand how connected consumer diversity affects consumption behavior.

J.P. Benedict focused on power of social media for Broadcast business insight. The programming content of media and entertainment companies can be maximized by integrating a broadcasting-focused intelligence program which continuously uses social media metrics and analysis tools across function silos. Broadcast industry and especially TV is very active in its use of social applications such as Twitter. There are many insights into the TV business that can be gained from social data like the response to promotional campaign events, the greater confidence in predicting
award winners, better targeting and result of promotional campaigns, better results in the procedure of gaining sponsors and partners and many more. The key to achieve these is to map social metrics and align the analysis tool with the organization’s specific goals.

Mylynn Felt [12] compares three free-use Twitter application programming interfaces through which users can capture tweets and enable analysis. These tools are Storify, Netlytic and DMI-TCAT. Moreover, he states the limitations of refining data through the social media and the difficulties that researchers might face in accessing valuable data.

Soumitra Dutta [14] focuses on personal social media strategy. Today’s businesses must use social media for many reasons. They provide a low cost platform on which you can build your brand, they allow you to engage rapidly with the public, customers, employees etc. and they give you an opportunity to take unvarnished feedback from instant information. The social must-dos are stated and some critical questions are answered in order to formulate a personal social media strategy in the right way. On the other hand, the risks of being online in social media are also demonstrated leaving to reader the option to find the right balance.

Nazneen Fatema Rajani [15] proposes a recommendation system based on opinion mining. The used sources are Facebook and Twitter and it explores the way the sentiment estimates can be filtered in a way that the noise can be extracted. The main conclusion it derives is that although it is possible to predict an individual’s mood based on the tweets, it is very difficult or even not possible to do it with big accuracy. In this thesis, the procedure is described and the performance results are stated.

Falko Schulz [17] explores networks with SAS Visual Analytics and states that it is very important to understand the relations between the different parts of the data and not just the statistical details. Network analysis and network visualization are the most helpful ways to understand this. The paper describes the procedure of exploring networks with SAS Visual Analytics.

Bogdan Batrinca and Philip C. Treleaven [18] present a review of software tools for social networking media, wikis, blogs, newsgroups etc. It also provides a sample methodology and gives a personal opinion on these tools. It describes how to use these tools in order to capture, scrape, cleanse and analyze the available data of the various social media. The variety of web-based application programming interfaces
(APIs) provided by the social media like Facebook, Twitter and other platforms has lead to the creation of many tools and services for scraping and analysis.

*Centrality Measure Of Social Networks Using Gephi As A Visualization Tool* [21] examines social network communities on Facebook using Gephi for visualization tool. Measures for the persons involved are taken with the help of centrality measure. The data extraction is done using Netvizz tool. As a case study the authors use the university’s Facebook profile and they are lead to some conclusions like that page likes and sharing are strongly supported to make profiles more social instead of writing comments on the profile page.
3 Problem definition

3.1 The problem

The evolution of technology and the automation of many everyday tasks have lead to the creation of huge amount of data in communications. The digitization of every simple procedure we follow in everyday life has brought us to a new era with the help of Internet of course. Internet connects all people and with the help of sensors embedded in smart devices people are connected with “things” from complex and heterogeneous resources and networks (Internet of Things). This means that people turn to use more and more the Internet and tend to be online almost during the whole day. Eventually, the exchange of data in communications is huge and the information in these exchanged data can be really useful. These raw data are known as Big Data, a term used for datasets that are large and complex which are very difficult to deal with.

In parallel, the use of social media networks has grown rapidly and users connect their daily activity to these networks. Huge amount of people use and live through social media. The main problem is that people do not know how to handle big data. They don’t know how to turn these raw data into useful information and information into knowledge. Regarding to the Social Media networks, the exchange of information is also huge while most users have various accounts in many social networking sites. Each site is a different network which can offer valuable information. So how can users extract knowledge from social data and what will they gain from this?

3.2 Why Big Data are important

Taking into consideration the exchange of big data in the communication systems and the observations that have been done during the last decade about the use of social media, we can easily conclude that the importance of analyzing big data can be vital. So, let’s see the following facts:

Twitter facts:

175M Twitter accounts
100M tweets per day
20% of tweets contain brand or product reference
Source: RJMetrics and Twitter.com, February, 2010

Facebook facts:

<table>
<thead>
<tr>
<th></th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Members</td>
<td>850,000,000</td>
<td>1,100,000,000</td>
</tr>
<tr>
<td>Checking</td>
<td>31% every day</td>
<td></td>
</tr>
<tr>
<td>Spend time per day</td>
<td>Avg. 20 minutes per day</td>
<td></td>
</tr>
<tr>
<td>Like</td>
<td>2,700,000,000 per day</td>
<td></td>
</tr>
<tr>
<td>Connection</td>
<td>100,000,000,000 per day</td>
<td></td>
</tr>
<tr>
<td>Installed applications</td>
<td>20,000,000 per day</td>
<td></td>
</tr>
<tr>
<td>Photos</td>
<td>250,000,000 per day</td>
<td></td>
</tr>
<tr>
<td>Access via mobile</td>
<td>450,000,000 per day</td>
<td></td>
</tr>
</tbody>
</table>

Source: RJMetrics

LinkedIn facts:

<table>
<thead>
<tr>
<th></th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Members</td>
<td>161,000,000</td>
<td>200,000,000</td>
</tr>
<tr>
<td>Checking</td>
<td>60,000,000</td>
<td>every month</td>
</tr>
<tr>
<td>company pages</td>
<td>2,000,000</td>
<td></td>
</tr>
<tr>
<td>Searches</td>
<td>4,200,000,000 per year</td>
<td></td>
</tr>
<tr>
<td>Groups</td>
<td>more than 1,000,000</td>
<td></td>
</tr>
<tr>
<td>Students</td>
<td>35% students search for jobs</td>
<td></td>
</tr>
<tr>
<td>Traffic from mobile</td>
<td>22%</td>
<td></td>
</tr>
</tbody>
</table>

Source: RJMetrics

Blog facts:

200 million blogs world-wide
32% post opinions about products and brands
71% of internet users read blogs

Source: Universal McCann 2009

Taking in mind these statistics, business users can have access to a huge market through social networks. They can use social media to attract new customers, to collaborate with other organizations, to recruit job candidates, to promote their services, as well as for PR activities.

### 3.3 The solution

The solution in gaining knowledge from Big Data is Analytics. The vast majority of the content that the users produce, such as that in blogs, tweets and many online communities demands an automated solution. Regarding to the social media net-
works the solution must have the ability to manage social data in all its forms effectively, even the data is structured, unstructured or semi-structured, including both video and audio content. This should be able to overcome linguistic and language issues. The problem becomes bigger especially in the social media communities where the language is like making conversations, full with similar expressions, slang and emotions (e.g. excitement, disappointment etc.). These conversations are stated so briefly that is difficult to distinguish the context. An automated, comprehensive social media analytics solution has to be flexible in such way that it can adapt to any changing topics and trends. It must also include functionalities like parsing through online blogs, articles or other websites for relevant generated content, identify trends or identify a group of people who may have similar interests. Social media analysis tools should provide results that are understandable and in such form that the user can take action by using them. Successful social media solutions are using meaning-based technology, semantic understanding and they always have built-in analytical engine to provide the best outcomes from social media traffic. A social media analytic solution extracts the meaning of the content and gives the opportunity to the users to proactively respond to some situations by making use of various social media applications. Through a set of connected APIs to social media networks it “listens” and understands what is being told and responds accordingly to emerging topics. The right solutions are supposed to be enhanced by several complementary technologies. These technologies can be advanced mathematical modeling so that a conceptual and contextual understanding of information can be formed in any format and language. This important capability enables the automation of key functions that are very important to handle knowledge and social media content.
4 Potential of Social Media Analytics

Every day, users and customers engage in online conversations about products on networking sites like Facebook, Twitter, Instagram and offer valuable feedback to businesses. Organizations which want to be competitive in the market can use social media monitoring and analytics tools to find and analyze that data. Social media analytics offer also to businesses the ability to identify patterns in customer sentiment and measure their marketing effectiveness. It is essential that today’s businesses must embrace social media for many reasons. Firstly, they allow them to interact quickly and simultaneously with customers, employees or the broader public, especially with younger generations. They represent a platform that enables interaction among various participants. Secondly, they represent a low-cost way on which businesses can build their brand, communicate with various stakeholders within and outside the company. Thirdly, they give a chance to learn from instant information and feedback.

The information that the social media provide is the most important characteristic of such networks. Every day users generate thousands of messages and they offer valuable insights in public forums without any cost. Marketers are provided with all the information and details they need so that they can have a full image of consumption habits, opinions and trends. From the moment that these data are combined with specific strategy and performance metrics, rich knowledge can be obtained which can help in taking critical business decisions. For example, in case of TV industry there are many insights that can be gained from social network data:

- Day by day view of customer behavior and sentiment.
- Sentiment based on demographics or on specific demographic market area.
- View into what people like or don’t like about specific content.
- Response to promotion events or campaigns.
- Faster awareness of trends and emergency situations.
- Better accuracy and confidence in predicting award winners.
- Better understanding of people who have a key role in a network.
• Better accuracy and better results in targeting of messages that the brand wants to promote.
• Better targeting of sponsors.

All of this kind of information can be collected through social networks. The basic problem is how do organizations may start to find the valuable information through raw data? A basic key to the discovery of this for businesses is to implement an integrated framework of social strategy which consists of tools and processes, along with specialized analysts who understand the business and the metrics that need to be defined. Of course, in order the framework to be successful, the businesses have to insert the logic of mapping social metrics to specific goals and implement this way of thinking in their overall social analytics strategy.

One other case at which the social media analytics can be used is the human resources management as a new source of modern human resources recruitment for the needs of the business. Organizations intend to attract creative professionals, proven leaders, young, talented and educated staff. Social media are used as an external source of human resources recruitment.

Moreover, they can be used as an innovative solution for achieving internal efficiency. The benefits of their use can be utilized in various ways like: cooperation (social media provide consistency, transparency establishing effective cooperation with employees, job applicants, customers, suppliers), talent management (SM can help attract highly competent people). Not only Linkedin is a good example of such tasks, but also other platforms like Facebook which many do not consider it as a serious site, but if you penetrate deeper in the analysis of the information offered in the personal accounts of the members, we will conclude that it is a contemporary job search tool.

Also, big data analytics can be used for the predictability of social events. Twitter data have been analyzed to predict winners during contests. The fraction of tweets that contain geolocation information gives us to map the fan base of each contestant showing the strong regional polarizations that occur. Online systems offer a huge amount of data which permits the real time gathering of various indicators that may be able to predict future events and opinions. The collection of massive data on social behavior is a unique opportunity to observe and study social phenomena. Search engine queries or social media posts have been used to forecast elections.
trends, market behavior etc. However there are challenges that have to be faced with this kind of data like intrinsic biases, uneven sampling across location of interest etc. Therefore, open source data is very important for the forecasting of various events and this can be achieved through the analysis of blogging activity for example. Especially in the entertainment industry, for example in TV shows, the data from the social platforms are becoming more and more a serious part of the show. This fact amplifies the importance of the indicators that can be extracted from these media in monitoring various things. Moreover the increased use of smart phones and mobile devices create geolocalized information about the social media activity that we explore. The geographical information is a key ingredient in achieving predictive power. Today’s business environment demands the existence of the ability to predict the outcome of future events accurately. For example, sales predictions and prediction of future consumer behavior can really help businesses reconsider their research and production planning. The first thing they have to do is to identify social media information sources, which include opinions relevant to the event set and extract the appropriate information. In a second step, a method is applied for processing content, for example sentiment analysis for textual input. Marketers evaluate the content that is generated every day and update accordingly their actions. Then they submit buy or sell offers for products which are related to a specific indicator or offers for the products they consider most successful for example.
5 Analytics challenges

Marketing and entertainment organizations need answers to tough questions like: “How do I identify the relevant blogs with a specific topic?”, “Who are the key influencers?”, “Which are the emerging topics in this field?”, “What is the sentiment about these relevant topics?” and many more. These kind of questions drive various sets of analytics like relevance classification, sentiment classification, casual influence in the blogosphere, topic evolution, measuring influence (viral potential) in facebook or twitter networks, concept mapping for learning and information filtering. This growth has affected how people process and interpret new information. Having in mind that most data originates and resides in the Internet, a serious challenge is to determine how computing technology should evolve to let us access, analyze and take actions on big data.

Businesses have to take into account various obstacles when implementing social media analytics plans in order to ensure that they provide value to their business. Without a plan that combines social media analytics and a set of technologies that can effectively support that procedure, businesses take the risk of missing the chance in trying to implement the isolated insights gleaned from social networking data into strategic business intelligence. But treating social media data as an independent network is a big mistake. Social media is just a piece of the big part of the game.

Eventually, some key challenges are the following:

- The available data has its limits. All of the data being generated in social forums are not readily accessible for analysis. For example most companies don't have access to the complete Twitter data and are only able to capture publicly available information on Facebook.
- Text analytics software isn't perfect due to the limited ability of text analytics tools to interpret the nuances of written content, including sarcasm, slang and irony. Assessing the meaning and sentiment of social media posts is a complicated procedure. Even widely used abbreviations can confuse the tools.
• Organizational obstacles are many. To rate a program as successful, we need more than just choosing the right social media monitoring tools and analytics software. First, companies need to consider what are the resources and skills they have inside the business and decide if they need hiring new employees or bringing in outside contractors. They also need to deal with issues such as who "owns" a social media analytics program and who is responsible for its success. The creation of a social media committee with representatives from various departments like customer service, marketing, product development etc can really help to integrate social intelligence strategy.

• Reaching too many data may have opposite results. Too much information can be negative for social media data analysis initiatives. Trying to find useful information in every corner of the social media universe might not be a good idea. Companies must have the ability to focus on what matters to the business' bottom line and not look at every detail of the product.

The data you can dig into on social networking sites represent only a small part of what's being said about a company or a product.
6 How the Media & Entertainment industry uses insights from social data

6.1 How they employ Big Data Analytics

In order to gain insights not previously possible, companies have to understand firstly who is consuming what, when, where and how. If they achieve this, then they can apply predictive analytics. They will understand this by collecting and combining consumer data from the Web, email, social media and other digital sources. Media and entertainment companies are asking themselves: What resources do we need for predictive analytics? How much data is “big data”? Should Hadoop be part of our information strategy? Where should we focus first? How can we be more data-driven? And can data make our supply chain more efficient? If we understand patterns about content consumption it is sure that we will uncover previously unknown insights about the connected consumer. Some companies are moving to a Hadoop-based platform such as Cloudera to take advantage of big data. Analytics can increase your understanding of how connected consumer diversity affects consumption behavior. This knowledge is useful to create more personalized marketing communications. This knowledge can be built by:

- Building a general view of the consumer by gathering data from various sources like social media, Web interactions and POS data.
- Optimizing campaign management by measuring campaign performance.
- Producing clever customer segmentations, so when interacting with these customers, they can engage in relevant, real-time content.

The second bullet will tell you what is working to attract the consumer and what marketing activities and channels have the most impact. The other two will tell you how to attract and keep the consumer. By analyzing social media data you can realize what consumers actually have in mind about specific content. In case of TV industry where you have to decide about content production such information can be important before ratings are aggregated, or to evaluate how audiences are receiving your ad campaign. The accuracy of the forecasts on new releases is basic factor to
have a more efficient content supply chain. For example, by using a data mining tool we can determine which upcoming movie is similar to which previously released in order to better forecast box office performance.

Social media analytics help us to discover topics with important meaning, demographics and other relevant information. This knowledge discovery requires automated solutions sophisticated enough to detect trends, sentiment, influencers and real time issues during the time they arise. The power of search and discovery analytics engine is critical. For example, analytics help government agencies to interact effectively with citizens. It helps people be more informed of what the government is doing, makes government agencies and officials more accessible. The best engines are that which can analyze online conversations and interactions through specific techniques. They tend to use probabilistic modeling and they treat words as symbols of general meaning and sentiment rather than words with specific definition. These engines can understand slang and context-specific words and adapt to the dynamic nature of language automatically. Of course, these kind of engines must operate across the variety of blogs, forums, social networks and online communities. When the analysis is automatic and proactively responsive, it is sure that a company does not miss an emerging trend or topic. These Engines can be used to monitor what is being said continuously or over a specific period of time. This possibility influences the strategy and the future approaches of the companies and eventually reduces the costs. Moreover it is a good tool to build a good marketing strategy in order to build brand and good reputation. A sophisticated Social Media Engine has also the possibility to quickly react to dissatisfaction when it monitors negative comments or conversations with bad sentiment which means that a company can avoid damage on sales.

The continuous creation of new blogs and forums has given the average consumer the power to influence the opinion of the public and of course the profitability of the companies. Marketing organizations need to be mindful of what people say in blogs and forums and how the expressed opinions could impact their targets. Social Network Analysis, Machine Learning, Data Mining, Information Retrieval (IR), and Natural Language Processing (NLP) are some concepts that have emerged in order to achieve this. But how can we identify a subset of blogs that make discussions for products or other concepts that are relevant to these products? And how can we
identify the most influential bloggers in these forums? How do we detect an emotion for a product in a blog or forum?

The first thing we have to do is to filter down from millions to hundreds the most relevant blogs to the product we monitor. The objective is to filter not only the blogs that directly talk about the specific product, but also competing products in order to understand how we can gain potential new customers. There are two kinds of techniques to achieve this, the one is the IR where keyword searches are used to retrieve relevant blogs, and the other is the Machine learning, where the filtering is implemented with standard classification. These techniques have differences. In the first case, a list of keywords is produced related to the product being monitored, while in the second case, specific blogs are pointed as positive examples in order to train a classifier. The less human supervision we have in discovering blogs with relevant content, the better the engine is. The real need is to have scalable methodology which combines various techniques and integrates various sources of information in a specific manner.

6.2 Detecting emerging topics

The main goal of social media analytics is to automate the human process of detecting and summarizing patterns that are emerging in the various blogs and forums. The posts that are useful for us to read are filtered and extracted measuring criteria of relevance, authority and sentiment. If we consider the fact that there is a huge variety of posts in the internet, it is very challenging and also valuable to find patterns of what people are talking about, and determine topics that emerge in the discussions.

The technology today create platforms where the users generate Big Data and they are transformed in a way into patterns. Of course those patterns are traceable. Activities, connections and products are collected, saved, and then analyzed. Big part of these data is collected through social media networks like Twitter. But what kind of analytic tools are researchers employing now that Twitter has become more restrictive? Are researchers focusing primarily on Twitter? To what extent do researchers compare data across platforms? Recently Twitter altered its API structure, making more difficult for analysts to use the tweets. Analysts have now to purchase the data they need through a Twitter reseller such as Gnip. Twitter has the advantage that the posts are not necessarily limited to a specific group of users like Face-
book “groups” or “friends”. The posts are public by default and may also be found by visitors searching the site or tracking the Twitter stream. This fact has played a significant role in the researchers’ choice. They analyze Twitter feeds, because they are more accessible to them. They focus on Twitter, because it answers many social questions and expresses the public sentiment about things and persons. It is a barometer for revealing everything, from the occurrence of natural disasters to the public perception of political candidates. However, analysts have to realize the limitations of information provided through data analytics as well as the algorithms and processes involved in capturing such information. On the other hand, the users who are the persons who produce big data have the least control over their data. There are lot of difficulties which prevent platform users from understanding of how their information is used. The common users produce the data, they don’t analyze it. The researchers who have not aim for profit seek for free portals to data analytics. This can give the possibility for improvement of the platforms. Moreover, the implications of social media practices can be studied better. There are free Twitter API tools like the storify.com, netlytic.org or the DMI TCAT. These free tools offer data for dynamic visualizations of networks and Twitter content. A social network analysis reveals who the influential social media users are in a specific network. The visualization of users reveals differences between them and presents those who are connected with lots of other users. These are the highly connected users who have a strong voice in the network. These tools give the possibility to researchers to see various topics of discussion and discover relationships between the participating actors. Patterns among the actors and patterns of discussion are also emerged.

The choice of the right tools to use in social media analysis is very important. The design plays a significant role. For example, API tools are constrained by Twitter’s policies. Social media offer many answers to many social science research questions. Twitter is of particular interest, public nature, and flexibility. However, changes in the policies over the last years have made access to tweets more costly and restrictive. There are API tools for researchers that need a payment or subscription in order to have access to more “deeper” data. Storify and Netlytic are capable of capturing data also from other social media except Twitter. Both TCAT and Netlytic combine capture with analysis. Storify keeps the original visual presentation of posts and active links but it doesn’t generate usable metadata. Netlytic is a user-friendly tool and generates quality visualizations without significant prior training.
However, it has a price associated with it. The metadata available for export is not as rich as what TCAT generates.
7 Harnessing social data to build actionable insight

The entertainment companies need to make their broadcasts more popular. One way to achieve this is by programming their content in order to be more focused on the public interest. This can be implemented by having an intelligence program with instant use of social measurement and analysis tools across various functional silos. Many marketers in Media and Entertainment are not aware of the true benefit of the social Web. Many companies build profiles in social media with a narrow way of thinking, just focusing on getting likes and followers. The approach they make is usually simple, they use dissimilar tools and have difficulties in understanding which social metric to analyze in order to gain business insight. Marketers still use traditional forms of measurement. Marketing departments continue to use social tools for promotion, but this is not enough. They must create business critical KPIs and then apply meaningful intelligence based on the measured data.

7.1 Methodology

Social technologies can potentially unblock big amount of money in the market. A large amount of this money can be unlocked from improved market intelligence in the transactions with customers. Another amount can be unlocked from improved communication and collaboration opportunities. This will impact the market through the ability to listen, understand and act appropriately on social data. Broadcasters can use this data for content optimization and better engagement with the customers. An effective social media analytics program should consist of listening, refining, workflow and strategic elements. The objective is to align the analysis tool with the organization’s social data goals and ensure that the reporting data can be integrated rightly into the workflow of the business.

7.1.1 Listening

The first important step is listening. This is the part where access to all of the “discussion” available in various platforms (e.g., the Twitter, Facebook etc.) is achieved with the help of data feeds data feeds. The extracted data from social me-
dia applications are either unstructured or less structured than the usual data pulled from common marketing data sources. A big part of the social “chatter” needs to be refined in order to be used in effective way, because the formats are not usable. However, there are solutions to this problem. For example Crimson Hexagon’s social listening tool offers an algorithm that users can train over time to structure social data into a more useful form. There are also more solutions which help in this direction.

7.1.2 Refining

The refining of data is where all of the pulled data from a social listening program (Tweets, posts, etc.) is organized into a structured, usable format. Refining data is a process which requires knowledge to manage the complexity of transforming the data into meaningful forms. Of course, it must be automated so that no posts or comments are missed.

7.1.3 Workflow and integration

Companies usually have problems with the integration of social media intelligence between the various departments. Each single department use different APIs, schemas and formats which leads to a siloed structure of operation. This means that the valuable data are trapped within the single departments, when this data could provide value to other areas of the organization. The value that can be created through the integrating social intelligence into the business process is huge. An effective workflow can create value in almost every department of an organization.

7.1.4 Strategy

The main problem with social media programs is their lack of a comprehensive strategy. Departments in organizations are having independent social strategies to improve particular business units like research, ad sales, production etc. For example, a marketing department might create its own Facebook page and Twitter account to promote an event or a TV show. On the other hand, the production department creates its own hashtag to interact with the event or TV show in real time. These activities could be done with coordination and planning so that the benefits of social media could be maximized. A standalone social strategy is not the solution. Of course, this kind of strategies should be always supported by an executive team.
which is engaged to this part of work. This team also has the responsibility of giving solutions to the existing silos in order to unlock the value of social data across the organization. The goals should also be connected to specific metrics and the outcomes should be reported continuously. In this way the normal procedures of the company could be optimized and bring better results.

The practices that entertainment and media companies are a good idea to established are many. Giving quality content is always a good idea, the gamification of social media efforts and the personalization of fans. The companies can reach stakeholders through many ways including television, radio, publications, and online efforts. Specifically in the case of social media they represent mobile and web based technologies that provide interactivity to the public to share, post, discuss or modify content. Companies use social media in various ways that most times are bringing the desired results. For example, they use blogs and other tools to engage customers, they use both traditional and Internet-based promotional tools to reach the audience and they support problems that are important to consumers. A clever tactic is not to appear as a marketer to the consumer but to establish a right relationship with him/her. This helps in building trust. Most marketers point out that the most important aspect of a social media presence is the content to have good quality.
8 SMA procedure and technical workflow

Social Media may influence the thought and consuming behavior of a large number of potential buyers. Big percent of consumers trust opinions posted online by other consumers. The changes that the social media in traditional marketing causes offer also a good opportunity to companies to adapt their strategy by leveraging social media to their advantage. This requires not only new way of thinking, but also automated analytic capabilities. Social media analytics contain a number of disciplines like social network analysis, machine learning, data mining, information retrieval and natural language processing as it is mentioned in previous chapters. However, several questions are created for the automated analysis of social media, blogs etc:

- From the moment that we know that the blogosphere is huge, how can we identify the set of blogs that are discussing not only a specific product, but also more sophisticated concepts that are relevant to this product?
- What sentiment is expressed about a product or concept in a blog?
- Who are the most influential bloggers in this relevant set and have the big authority?
- What are the emerging topics in these discussions?

8.1 Methodology

The big amount of social data is provided by various resources which mainly try to sell these data. It is very important for researchers to have access to open-source big data in order to make experiments. If this does not happen social media research could become the exclusive domain of major companies.

Research requirements can be grouped into: data, analytics and facilities.

- Data

Researchers need online access to real-time social data to conduct a research. Some types of this kind of data are the following:

- Social network media data
- News data: historic data and real-time news data sets.
• Public data: scraped and archived data which are usually available through RSS feeds.
• Programmable interfaces: application programming interfaces (APIs) to scrape and store data sources that may not be automatically collected.
  ➢ Analytics
For the analytics researchers need access to:
• Analytics dashboards: non-programming interfaces which give access to raw data.
• Data analysis: tools for analysis across multiple social media and other data sets.
• Data visualization: visualization tools where information can be visualized in graphic forms which is more clear and understandable.
  ➢ Facilities
• Data storage.
• Computations.

8.2 Finding the relevant blogs and posts

As previously said the first target is to filter the quantity of relevant blogs to the minimum possible. The “topic” can be a specific product and the objective is therefore to identify all blogs, groups or posts discussing this product and perhaps competing products as well. Tracking a specific product and tracking a wider concept impacts the methodology we have to use in order to find the relevant set of blogs. If the interest is only in a specific product, it is easier to identify blogs by using a search engine for example. Such an approach has no good results we the interest is for broad topics, because discussions on a topic may not contain reference keywords. As we know relevant blogs are likely to link to other relevant blogs and an approach to text classification could be to research the structure of the blog cross references. This simple approach seems to be more efficient than keyword search. When product search terms have multiple meanings, e.g.,“Agora” is an ancient market, a documentary film and a Greek delicatessen shop and brand, it is unlikely that blogs talking about Agora will reference blogs discussing one of these meanings. An important consideration is to avoid dragging the parts of the relevant blog sub-communities that are irrelevant from a marketing perspective.
8.3 Sentiment detection

Sentiment detection or analysis or opinion mining refers to the application of natural language processing, computational linguistics and text analytics to identify and extract information in various sources. Sentiment analysis is used to determine the attitude of a writer, speaker regarding to a discussion, topic or the overall polarity. The attitude may be the speaker’s evaluation or emotion. A basic task in sentiment analysis is classifying the polarity of a given text at the document or sentence, if the expressed opinion is positive, negative, or neutral.

This classification limits the quantity of blogs. But which of the relevant blogs should you read? A basic concern is to detect strong sentiment especially when it is negative. Since it is impossible to read all the relevant blogs to the topic of interest, there is a big need to develop automated features which extract the tone and sentiment in these posts. One big challenge in sentiment classification is that the expression of sentiment is specific to the domain and the set of domains change often. Thus we require classifiers that can rapidly adapt to new domains without requiring a big number of training examples of positive and negative sentiment.

8.4 Collaborative filtering

Collaborative filtering (CF) is the process of filtering for identifying information or patterns using techniques involving collaboration among multiple agents, data sources, etc. It is usually implemented in large data sets. Collaborative filtering is a procedure of making automatic predictions about the interests of a consumer by collecting preferences and information from other users. The concept is that those who agreed in the past tend to agree again in the future.

8.5 Measuring influence and authority

One basic concern is how much does the opinion of one blogger matter in the community? Are the rest people listening or reading what he/she says and will these opinions influence others? Bloggers that have big impact on others may or may not be experts, but it is sure they influence others’ opinions while discussing about a topic. There are bloggers who have lots of connections and spread what they say in the community with very good results. The common way to determine a blogger’s authority is to use the links between blogs. In this case we have a network of di-
rected edges indicating the links between blogs or posts. From social network analysis we take more indications. For example, the authority of a blog can be extracted based on the number and authority of other blogs that link to it (using the PageRank algorithm), while the influence of a blog can be captured by the degree to which the blog contributes to the flow of information between other bloggers (Flow Betweenness). But how can the different measures of authority and influence be calculated. In social media high authority have those who influence the thinking and subsequently the content blogged by others. If a blogger is influential his ideas propagate to other blogs.

8.6 Network visualization

A network is defined as a collection of nodes, which some of them are connected by links. A link can represent any type of relationship. Networks are represented by graphs. The interconnected nodes are represented by vertices and the links that connect some pairs of vertices are called edges. Vertex attributes can be mapped to visual node characteristics like color, size and shape. Edge attributes can be mapped to link width and color. The direction of edge attribute is important. There are relations that are undirected or directed (asymmetric). For example, the “friends” relationship in Facebook is undirected. But the “follows” relationship in Twitter are directed, or asymmetric which means that a link from A to B doesn’t imply a corresponding link from B to A. In this case, there will be arrows on the edges of the visualization to indicate the direction of the relation. The biggest challenge in network visualization is finding an optimal spatial placement for the nodes that makes the main characteristics of the network structure clearly visible. In a few specific cases, like transportation networks, node position is a simple mapping of node attributes that represent spatial coordinates. This is how Visual Analytics overlays network diagrams over geographical maps. But in general, node positions are generated dynamically based on the network structure. This is accomplished using a special type of algorithm called a network layout, which attempts to minimize edge length and crossings for maximum readability. Exploratory data analysis is done through an intuitive graphical user interface that enables manipulation of graphs in real time. Visualization techniques enable users to detect patterns and extract information of interest. Visualizing networks in gives the user additional insights about the under-
lying structures of associations between entities or objects. Analyzing social networks shows connections between people and reveals entire communities or group of friends. The user can assign colors to nodes and links as well as assign node size or link width based on variables in the source data.

8.7 Importing Graph Exchange formats

Network data sets are available in many different formats. You might want to analyze a pre-existing network that was stored in one of these standard network data formats such as GEFX and GraphML. GraphML is an XML-based file format for graphs supported by open-source graph visualization applications like Gephi [Gephi]. It supports attributes for nodes and edges, hierarchical graphs, and benefits from a flexible architecture.

- Graph Exchange XML Format (GEXF)

GEXF is a XML-based language for describing complex networks structures.

Based on the context of your network, you will need to add network metrics in order to understand relationships in your data. For example, in social networks you might want to add metrics such as Betweenness or Eigenvector centrality so that you can perform social network analysis by finding key actors within a social group.

For further analysis we are also going to group our nodes based on their social role in the network. Community detection is to group nodes into communities such that there is a higher density of links within communities than between them.

8.8 Exploring networks

- Random Networks

How are networks formed? If networks are the common structure behind many different systems in nature, could we also find common laws that govern their evolution? Nature grows networks by connecting its nodes randomly. The highly complex structures found in networks were basically a consequence of this randomness. This idea, the Theory of Random Networks, dominated our view of networks for many decades. And most natural phenomena (imagine the height of your coworkers or your blood pressure measurements) fit this model. Their measures are characterized by a bell curve, with a peaked distribution that decays rapidly on both
sides. Most values are close to the average, and extreme values are rare. For random networks, that means that most nodes will have about the same number of links, with a few exceptions to either side, and no extreme outliers. For example, let’s think at the network formed by Greek cities and the highways that connect them. Every city is connected to at least another city by a highway, most cities have about the same number of roads connecting them, and there is no city connected directly to hundreds of others.

- Scale-Free Networks

But as time went by, scientists and researchers started to find networks where this rule simply didn’t apply. The most obvious example is the World Wide Web. The first “crawlers” came back reporting that some sites had a disproportionate number of links pointing to them, and they were far from being as rare as predicted by the distribution associated with random networks. Soon other examples were found, from social networks to citations in scientific journals and many other.

- Social Media

Social media are complex systems. They have complicated structure and the various components influence each other. Who you know affects what you do, and vice versa. The network consists of players or actors, but sometimes other even of automated agents. What the actors do is a function of with whom they relate. Actors in critical places in the topology can affect the entire group. So, what is the underlying structure of the network? Do we have single groups, or strong and linked communities? Who are the influencers? Who are the key actors in our network? Key actors build the structure of the network while communities define the scope of their influence.

- Community Detection

Community detection, or clustering, is the process by which a network is partitioned into communities with the criteria of density of the link connections within the community. The nodes in these communities have common properties or similar preferences. Therefore, identifying the communities in a network can further the understanding of how network function and topology affect each other. This has direct applications in marketing, sociology, and many other areas.

- Key Actor Analysis
There are different ways to measure who or which is important or central to a network like:

- **Degree** which reflects how many players are connected to a specific player. This metric counts direct relationships like number of friends, followers etc.
- **Betweenness** which measures the number of shortest paths an actor is on, which indicates how often actors can reach each other through it.
- **Closeness** shows the relative distance to all other players which is given by the shortest path between a pair of players. When a player has a high score it means he is close to everyone.
- **Eigenvector centrality** is proportional to the centrality of an actor’s neighbors. Google’s PageRank algorithm is an example of this metric. A high score indicates the actor is popular among popular players.

A visual representation of these measures using color can be seen in the figure below. It can be seen how they emphasize how different aspects of the network structure are reflected on the centrality, or importance, assigned to different nodes.
Key Actor Analysis identifies significant nodes in a network by calculating Eigenvector centrality and Betweenness. Players with high score on both measures are important nodes in the network. Outliers might sometimes be key actors playing specific roles. An actor with high Betweenness and low Eigenvector centrality might provide the only path to a central actor. These are “gate-keepers”, connecting actors to a session of the network that would otherwise be isolated from the core. On the other hand, an actor with low Betweenness but high Eigenvector centrality might have unique access to central actors. These are “pulse-takers” which are well-connected actors at the core of the network. As illustrated in the following figure the bubble plot makes it easy to visually identify the key actors and the links between the two visualizations indicates these players in the network structure.
9 Analysis Tools and Libraries - Techniques

9.1 Tools and Platforms

In order to make decisions in the industry we have firstly to understand the used data sets and the structure of complex systems. For these data sets we use a visualization tool called macroscope. With macrosopes we can synthesize related elements and detect patterns, trends, and outliers. At the same time we gain access to many details. Macrosopes let us observe things that the human eye and mind is difficult to notice and understand. What is missing so far is a common standard for the design of modulars, compatible algorithms and tool plug-ins (also called “modules” or “components”) that can be easily combined into scientific workflows. There is a big variety of mining tools for sentiment analysis, from simple open-source tools to libraries, toolkits and platforms.

Due to the availability of web-based application programming interfaces (APIs) provided by social media platforms, feeds have become a major aspect of research and business. This has led to the spreading of tools for scraping and analysis. However, this possibility of easy access on social media data may change significantly due to commercial pressures. In addition, the tools available to researchers are far from ideal.

One of the biggest benefits of the use of social media for the company’s communication is that they can measure their campaign’s efficiency. The term Social Media Analytics has to do with the procedure of measurement, analysis and interpretation of interaction and correlation between people. A business can export valuable conclusions for the demographics of its customers, their preferences, their reaction regarding products and services and many more. After a detailed analysis, a business can also recognize the trends of the market and based on that to adapt better its strategy. Most companies are using the available tools mainly for measuring the efficiency of their campaigns, while other reasons are the brand analysis, finding competitive advantages, customer support etc. Of course for the use of such tools some expenses have to be made. There are free specialized tools, but there are also
commercial tools, especially for big companies. There can be also cost for the education of the employees who are going to be associated with these tools or for the digital marketing company which will do this task. Some of the best tools of social media analytics are offered in a very good price. The best choice of such tool is that which will help us in a specific sector to export the best conclusions, that which fits more to our demands. Some popular tools are the following:

**Google Analytics:** It is a free tool which Google provides to everyone who wants to analyze various data for the traffic on their website. Google has also added the possibility of monitoring data on the site which are obtained from social media like the percentage of visitors who come from social media or how many visitors of the social media are discussing something relevant with your business and many other features.

**Hootsuite:** Hootsuite is a tool mostly popular for the management of multiple social media accounts, but the analytics possibilities that offers are also valuable.

**Cyfe:** one more online platform with analytics tools which is offered in free version and in monthly subscription, where you can collect a variety of data for the social media and sales.

**SumAll:** This platform offers furthermore the benefit that it supports electronic payment services, so it’s very useful for those who work in the e-commerce. In this way it combines many useful analytics tools in one platform.

**Social Report:** One more solution which cooperates with e-commerce platforms and variety of other sites and services. It can be configured to to create daily reports, while it doesn’t cost a lot.

Most of the available tools have trial versions which the company can use initially for a short period in order to choose which tool fits better in the company’s demands. Additionally, there are even more tools some of which are complex and others more simple. Some of them are the following: **SAS Visual Analytics, Tweetdeck, Radian6, Crimson Hexagon, Fisheye analytics, Fizziology, Attensity, Lithium, Jive Software, SCi2, Gnip, Datasift, NodeXL, Gephi, DMI Issue Crawler.**

In general, we divide social media tools into:

- Social media data: social media data types (e.g., RSS feeds, blogs, blogs etc.) and formats (e.g., XML).
• Social media programming access: data services and tools for scraping textual data.
• Text cleaning and storage tools: tools for cleaning and storing textual data like Google Refine and DataWrangler.
• Text analysis tools: individual or libraries of tools for analyzing social media data once it has been scraped and cleaned.
• Social media platforms.
• Social network media platforms: platforms that provide data mining and analytics on social media networks.
• News platforms: platforms such as Thomson Reuters providing commercial news feeds and associated analytics.

9.1.1 Scientific programming tools

There are scientific analytics tools that are used for sourcing, searching and analyzing text like MATLAB which is used for numeric scientific programming and Mathematica used for symbolic scientific programming (computer algebra). Apache UIMA (Unstructured Information Management Applications) is one more scientific programming tool which is an open-source project.

9.1.2 Business Toolkits

Business Toolkits are tools that allow users to search and analyze text for a range of commercial purposes. SAS Sentiment Analysis Manager, part of the SAS Text Analytics program is an example that can be used for scraping content sources. It creates reports that describe the expressed feelings of consumers in real time. RapidMiner is another toolkit which provides data mining and machine learning procedures.

9.1.3 Social media monitoring tools

Social media monitoring tools are sentiment analysis tools which track and measure what users say about a product, a person, company or any other topic in the web. Some of such tools are Social Mention which provides social media alerts similarly to Google Alerts, Amplified Analytics which focuses on product reviews and marketing information and many others. Google Trends also which shows how often a specific term is searched in the web.
9.1.4 **Text analysis tools**

Some text analysis tools are Stanford NLP, LingPipe, Python NLTK (Natural Language Toolkit), GATE and Lexalytics Sentiment Toolkit.

9.1.5 **Data visualization tools**

Two notable visualization tools are SAS Visual Analytics and Tableau.

9.1.6 **Social media analytics platforms**

Social media platforms are different from tools and toolkits since platforms are more comprehensive and provide both tools and data. We can group them in:

- News platforms like Thomson Reuters providing news feeds and associated analytics.
- Social network media platforms which are platforms that provide data mining and analytics on social media platforms like Facebook and Twitter. Some examples of such platforms are Brandwatch, Salesforce Marketing Cloud or Radian6 and Syosmos MAP (Media Analysis Platform) which measure sentiments, demographics and they include text analytics on online consumer conversations. They also have user-friendly interfaces for customizing the search queries, dashboards and reports.

9.2 **Platform components**

- **Back-end services**: this is the core of the platform. The back-end services are a set of services through which connections to data providers are allowed, processing of data feeds and execution and maintenance of models.
- **Front-end client APIs**: this is a set of graphic and programmable interfaces that can be used to interact with a platform in order to apply the analytical models. These give the opportunity to the user to visualize data in different ways and in dynamic manner.
- **Connectivity engine**: this is the element which communicates with the outside world. Of course each connected element of the outside world has a dedicated connector element which is responsible for communication.
- **Aggregation database**: this is a DBMS system used for the entrance of data, which is then filtered, cleaned and stored in big data facilities.
- **Client SDK**: this is a set of APIs (Application Programming Interfaces) from which we can develop and test new analytical models with use of the Integrated
Development Environment. The SDK allows connection from the IDE to the server side of the platform so that all the functionalities are available to the user.

- **Shared memory**: the shared memory maximizes the speed of data delivery and the speed of analytics elements of the platform while it reduces the memory usage requirements. The concept is to have a central memory that will manage and provide data from the central point for recent or old data. Since the memory is shared, the model will not have to store and process the history by themselves.

### 9.3 Terminology

In order to understand the techniques related to analyzing unstructured textual data we have to mention the following definitions:

- **Natural language processing (NLP)**: is a field of computer science which has to do with the process of computers which extract meaningful information from natural language input and produce natural language output.
- **Opinion mining**: it has to do with the attempt to make automatic systems to determine human opinion from text written in natural language.
- **Scraping**: it is the collection of online data from various Web sites and social media networks in the form of unstructured text.
- **Sentiment analysis**: it is the procedure of identifying and extracting subjective information in data sources.
- **Text analytics**: it refers to information retrieval (IR) and text analysis in order to discover word frequency, patterns, or any data mining techniques.

### 9.4 Research challenges

If we go deeper into the analytics procedure, we’ll easily realize that we have to face some hard challenges:

- **Scraping**: although social media data is accessible through APIs, it is really difficult to obtain access to the ‘raw’ data of social networks. Many social networks have charges for accessing their data.
- **Data cleansing**: it is really hard to clean in perfect way unstructured textual data, especially high-frequency streamed real-time data.
• Data protection: the data we obtain from various resources have to be secured and with right ownerships and access levels for the various users, because the retrieval of some data has to be restricted.

• Analytics dashboards: there are platforms that require from users to write APIs to access feeds or program analytics models. This is not easy for most researchers.

• Data analytics: the analysis of social data for sentiment analysis brings up lot of challenges due to variety of languages and the existence of slang.

Social media networks are the largest, richest and most dynamic pool for discovering patterns of human behavior and understanding individuals or social groups. Marketing professionals are continuously finding ways to automatically collect and analyze data. The initial social media analysts were companies from the field of finance and retail. Retail companies use social media for obtaining brand awareness and for advertising reasons. In finance, social media is used for measuring market sentiment.

9.5 Techniques

9.5.1 Social media data

The services providing access to social networking media are evolving rapidly. New sources of data are still being discovered. Eventually when we talk about textual data analysis, we should into account multiple sources (e.g., social networks, blogs, RSS feeds) enriched with geospatial, financial data or even audio and video data. The future of analytics depends on the usage and compatibility of various input data sources.

➢ Types of data

• Historic data sets: previously retrieved and stored data.

• Real-time feeds: live data feeds from streamed social media.

• Raw data: unstructured data taken from source that may contain errors or be unprocessed.

• Cleaned data: corrected data. All the dirty data, mistakes, missing bits, outliers have been removed or normalized and corrected.

• Value-added data: data that has been cleaned and analyzed and have offered some value regarding the knowledge.

➢ Text data formats
There are some standard formats which are used to markup text. These are: XML, CSV, HTML and JSON.

- **XML**: Extensible Markup Language (XML) is the markup language for structuring textual data to define elements.
- **CSV**: files which contain comma-separated values (CSV).
- **HTML**: HyperText Markup Language (HTML) is the markup language for web pages that can be viewed in a web browser. HTML consists of HTML elements, which include tags enclosed in angle brackets within the content of the web page.
- **JSON**: JavaScript Object Notation (JSON) is a text based open standard designed for human-readable data interchange and is derived from JavaScript.

HTML and XML are markup languages that define syntax rules for encoding documents in an appropriate format in order to be readable both from humans and the machines. Such languages consist of start-tags (e.g., `<tag>`), content text and endtags (e.g., `</tag>`). JavaScript Object Notation (JSON) is based on a subset of the JavaScript Programming Language. JSON does not depend on specific language and it uses conventions that are familiar to the C family languages, including C, C++, C#, Java, JavaScript, Python, and others. JSON’s basic types are: Number, String, Boolean and Array.

### 9.5.2 Social media providers

Social media providers are data input sources which can be grouped in:

- **Open source databases**: databases that can be freely downloaded, e.g., Wikipedia.
- **Data access via tools**: sources that offer controlled access to their data via specific tools with the goal of providing easy access to users, but also to prevent users from 'sucking' all the data from the repository.
- **Commercial sources**: data providers who charge the access to their data. Gnip and DataSift are an example of such providers.
- **Data access via APIs**: social media data databases which offer access to the data via APIs (e.g., Twitter, Facebook etc.).
  - Open-source databases

Wikipedia is an example of open source database which provides free copies of all available content to interested users (Wikimedia Foundation 2014). These databases can be used for database queries and social media analytics. Another example of freely available data for research is the World Bank Databank which offers many
databases, such as Gender Statistics, Health Nutrition and Population Statistics, Global Economic Prospects, World Development Indicators and Global Development Finance etc. Tools to customize and display reports in table and charts are also provided.

- **Data access via tools**

Many providers offer commercial access to social data via online tools for various reasons, for example to control the access to the raw data and to get value from the usage of the data. Google Trends and InSights is an example of such tools. The following figure displays the interface of Google Trends and how we do a search, in this case ‘Greek wine’. With this tool you can make comparisons for several topics and also check how often those topics have been searched and mentioned in the web and in which geographic areas.

![Google Trends Interface](image)

- **Commercial sources**

There are companies like Twitter which restrict free access to their databases, but are licensing commercial resellers to offer data in charge. An example of this kind of resellers are Gnip and Datasift.

- **Data feed access via APIs**

APIs are interfaces programmed using HTTP-based protocols which make it easy for marketers to access the data. Wikipedia provides large open-source repositories of user-generated (crowdsourced) content in HTTP-based APIs that allow programmable access and searching that returns data in a variety of formats like XML. The used API for Wikipedia is part of MediaWiki’s open-source toolkit.
Most of the social networks offer a proportion of their data and not the whole volume through the APIs. Some other networks do not provide at all APIs for scraping data like LinkedIn and Skype. It is observed that the last years many big social networks are restricting more and more free access to their data.

Tweets from public accounts are available in JSON format through Twitter’s Search API for requests of past data and Streaming API for real-time data. The Search API is used by querying Twitter for recent Tweets containing keywords and requires authorization before extracting any results. The Streaming API is a real-time stream of Tweets, filtered by user ID, keyword or geographic location. In social media, streaming APIs are often called Firehose which is a feed that publishes all public activities as they happen in one big stream.

Facebook has various APIs like the Graph and Public Feed APIs. It stores all data as objects and in order to access the properties of an object, its unique ID must be known to make the API call. The Facebook Graph API search queries require an access token during the request. Searching pages requires an ‘app access token’ while searching for other types requires a user access token.

Most web sites offer access to their content through RSS feeds. This is a connection standard for the publication of content updates. The format of the feeds is based on a type of XML file that resides on an Internet server. RSS feeds can be constructed manually or even automatically with the help of dedicated software. An RSS Feed Reader reads the RSS feed file, discovers which content is new, converts it to HTML and displays it. Feeds are available in various formats, but mostly as XML documents, JSON or CSV files. They contain numerical values, tags and other properties.

During the last decade much of the geospatial social media data are produced from mobile devices that generate location and time sensitive data. We can categorize four types of mobile social media feeds:

• Location and time sensitive: data relevant to specific location at one specific point in time.
• Location sensitive only: data relevant to specific location, which are tagged to a certain place and read later by others.
• Time sensitive only: mainly mobile applications with data exchange relevant to time. For example a Facebook status update.
• Neither location nor time sensitive: applications with data exchange irrelevant to time. For example watching a video on Youtube.

The evolution of smart devices like smart phones, pads, smart watches and other have added to the communications geographical identification called ‘geotagged’. These geospatial metadata are usually latitude and longitude coordinates, but they can also include altitude, bearing, distance and other properties. GeoRSS is an emerging standard to encode the geographic location into a web feed. There are two encodings the GeoRSS Geography Markup Language (GML) and the GeoRSS Simple.

9.5.3 Text cleaning and storing

The quality of the data we use in analytics play important role in the model we will build. The type and the structure of the data will also influence the database. Unstructured textual data can be noisy and will not bring the desired results. Therefore, data cleaning is really significant task. The process of this cleaning may need to remove typographical errors or outliers or correct values. The entries may contain misspelled words, quotations, unreasonable spaces and line breaks, special characters and so on. If we want to achieve a good quality text mining, we have to clean the data at a first stage. We have to remove duplicated values, make spell checking, remove spaces and strange characters, fix numbers and outliers, correct date and time formats, rearrange columns, rows and many other things. In general there are some usual problems with the data which are the following:

• Missing data: when a part of text existed but was not included in the data. Problems may occur and a sentence can have different meaning when numeric data is blank or there is a missing value which is replaced by ‘zero’. In textual data when there is a missing word (like ‘not’) problems can also be created.
• Incorrect data: when a part of information is specified or interpreted in a wrong way.
• Inconsistent data: when part of the data is inconsistently specified. For example, with numeric data, a differentiation in date and time formats like 2016/12/18, 18/12/2016 or 12/18/2016 can create problems.

➢ Cleansing data
Most practices and applications who do this job pull data into a spreadsheet or similar table and then reformat the text. For example, Open Refine is a standalone application for data cleaning and conversion to various formats.

- Tagging unstructured data

The data in social media are created by humans. This means that these data are unstructured and they don’t have specific structure or data model which will transform the input entries into structured data. These data need to be preprocessed, tagged and then parsed in order to be analyzed. Tagging the unstructured data involves adding extra information to the data. This can be done manually or via engines, which seek patterns or interpret the data using various techniques. Usually during the procedure metadata are added.

- Storing data

Social media networks have specific rules about how their data can be stored. These are usually described in the Terms of Service of each platform. Some types of databases that store data can be:

- Flat file which is a two-dimensional database like a spreadsheet containing unstructured records that have to be searched sequentially.
- Relational databases which are organized in tables and there are relations between stored items of information, fact that allows more complex relationships between the records.
- NoSQL databases which are a type of database management system (DBMS) which are different with the classic relational database management systems (RDBMS). NoSQL databases are non-relational, distributed, opensource and horizontally scalable. NoSQL databases improve the speed processing time that relational databases have.

There is an increasing community of researchers that use Apache software for social media analytics. The Apache Software Foundation offers three kind of relevant software which are:

- Cassandra databases: Cassandra is an open source (noSQL) distributed DBMS providing a structured store.
- Hadoop platform: is a Java-based programming framework that supports the processing of large data sets in a distributed computing environment. A given task
is broken into many small parts (fragments or blocks) that can be run on various systems and nodes.

• Mahout: service which provides distributed analytics algorithms running on the Hadoop platform. Mahout supports four classes of algorithms clustering, classification, recommendation mining and frequent item set mining.

  ➢ Social Media Analytics Techniques

  Sentiment analysis is a method to gain knowledge from the vast amounts of text and news content which users produce on the web. Such content is characterized by textual disorder and high diversity. Natural language processing, computational linguistics and text analytics are used to identify and pull out subjective information from such text. The target is to define the attitude and the feeling of a writer or the general polarity of the text based on specific topic. Automated sentiment analysis of digital texts uses elements from machine learning. The techniques are involved with the following fields:

  • Computational statistics which are statistical methods including resampling methods, Markov chain methods etc.

  • Machine learning which is a method of obtaining autonomously knowledge learnt from experience and analytical observation. Machine Learning aims to solve the problem of having huge amounts of data with many variables.

  • Complexity science which uses complex simulation models of systems that are difficult to predict.
10 Gephi Graph analysis software

10.1 Overview - Capabilities

Gephi is an open source software package which helps to explore, analyze, filter, manipulate and export any network. It includes algorithms for layout graphs and modifying visualization properties. Moreover, it gives the opportunity to calculate graph metrics like degree, pagerank shortest path etc. It offers new opportunities to work with complex data sets and produce valuable visual results. Basic requirements for such an exploration tool are the high quality layout algorithms, clustering, data filtering and statistics. All these features must be embedded in a flexible and user-friendly application. With the help of developed modules Gephi can import, visualize, spatialize, filter, manipulate and export all types of networks. The visualization module uses a special 3D render engine to render graphs in real-time. This technique uses the computer graphic card, like in video gaming, and leaves the CPU free for other kind of computing. It can handle with large networks and it uses multi-core processors. Node design can be customized and it can be a texture, a panel or a photo. The layout algorithms are configurable and can be run in real-time on the monitor. Moreover many algorithms can be run in the same time without having a crash in the software. Any data attribute associated with the nodes can be shown on label and Label Adjust algorithm can be run to avoid label overlapping. The user interface is developed into Workspaces, where you can work separately and independently. Very good work has been done to the extensibility of the software. An algorithm, filter or tool can be easily added to the program, without having programming knowledge. Filters can filter the number of shown nodes or edges with thresholds, range and other properties. Filters are sequentially chained which means that each box takes as input the output of the previous filter. Graphical modules like Size Gradient or Color Gradient can change the network design. Graphical modules take a set of nodes in input and modify the display parameters, like colors or size, to enhance the understanding of the network structure or content. Networks can be explored in visual way and they can be exported as a SVG or PDF file. Font and label
features are very well designed. For example, small labels can be drawn on edges to immediately display the neighbors of a node.

Gephi supports graphs whose values or content varies over time. The system continuously queries all nodes and edges and updates the visualization module. Therefore dynamic networks can be played as movie sequences. They can also be imported from external file. During the run, the source sends data to the controller and the results are displayed at that time on the visualization module. The architecture is flexible and can be customized in order to communicate with third party databases or web-services.

Version 0.9 is fairly new. It still has bugs that you may or may not encounter depending on your machine and operating system. Gephi has a number of community-contributed plugins available from the Tools → Plugins menu, as well as online at marketplace.gephi.org/plugins. At present, many of the plugins do not yet work with Gephi 0.9.1. That’s why Gephi 0.8.2 was chosen with Java version module jdk 1.7.79. The etc subfolder of the Gephi folder contains a file named gephi.conf. This file is used to change the startup settings of Gephi, for example the maximum available memory that the software can use. To increase that, we have to modify the value that follows the string -Xmx in the configuration file to the desired number of MB. If we are working with very large graphs and want to increase Gephi’s memory to several GB, we need to have a 64-bit OS and a 64-bit version of Java. The software interface has the following basic tabs:
Overview Tab: Calculate network and node measures, change visual properties and generally set up the network visualization. This is where we do most of the work: resizing, coloring, and laying out the nodes and edges.

Data Laboratory Tab: Examine and edit the network data (node & edge tables). This is also where we can import CSV and Excel files.

Preview Tab: Configure the rendering settings (colors, sizes, line width, opacity, etc.) and preview the visualization.

There are a number of options to manually interact with a graph

The Appearance box of the Overview tab allows you to change the visual characteristics of nodes, edges, and labels. The node and edge tabs have two options: Unique and Attribute (1). Unique changes all visible elements uniformly to the selected color or size. Attribute changes visual characteristics based on a selected attribute from the dataset.
To change parameters based on attribute, select the appropriate element (node or edge), select the “Attribute” tab (1) and choose an attribute from the drop-down menu. The visual parameters that can be changed for nodes from this box include: color (2), size (3), label color (4), and label size (5). The parameters for edges are the same except for size (edge width can be set based on the edge Weight attribute). This also means you can visualize different types of relationships in by coloring their respective edges in a different color. When you set nodes or edge color based on a numeric attribute, the attribute can be treated as categorical (separate colors for each level) or as continuous (a color gradient, usually with darker or more intense colors representing higher attribute values). To switch between categorical and continuous, press the cubes/palette button at the bottom left corner (6).

- Filter options

You can apply a variety of filters to select specific nodes and/or edges from your network. The filters are applied by dragging and dropping them onto the Queries window (1). Filters based on attributes include Equal, Partition, Range, Inter-edges, Intra-edges and so on.
Filters based on edges allow you to select graph edges with different properties e.g. a particular range of weights. Topology filters allow for selection based on network structure components, K-cores, degree range, etc. Operators permit you to combine other filters in different ways (find their intersection, union, complement, etc.).

- **Statistics**

To extract a statistic, we simply click the "Run" button next to it. In the result window, many measures will be available and can be used in the visualization. The newly generated attributes are listed in the drop-down Attribute menu of the Appearance box. Note that to get betweenness and closeness centrality, you need to calculate the average path length.
Some of the statistics algorithms are the following:

Pagerank: An iterative algorithm that measures the importance of each node within the network. The metric assigns each node a probability that is the probability of being at that page after many clicks.

HITS: It computes hubs and authority. The authority number indicates the value of the page (node) itself and hubs indicates the value of the links outgoing from the page (node).

Average Shortest Path: The graph distance between two nodes $K$ and $N$ is defined as the minimum number of hops required to cross the network starting from node $K$ and end at node $N$.

Node Betweenness Centrality: This metric estimates how often a node is found on a shortest path between two nodes in the network.

Node Closeness Centrality: This metric estimates how long it will take for information from a node $K$ will take to reach other nodes in the network.

Modularity: this metric estimates the strength of division of a network into clusters. Networks with high modularity have dense connections between the nodes within clusters, but sparse connections between nodes in different clusters.

Gephi offers a variety of layouts, including several relatively fast home-grown force-directed options (Force Atlas, and the faster Force Atlas 2). The classic force-directed algorithm of Fruchterman & Reingold nicely distributes the nodes in the available area, but is too slow for large graphs. It is also provided a multilevel force-directed layout that performs well for larger graphs. For undirected graphs, OpenOrd also provides a nice layout that aims at separating clusters. The Geo Layout allows you to place nodes according to geographic coordinates in the projection of your choice. In order to use it, you need to have node attributes for geographic latitude and longitude in your data. If you do not see the Geo Layout in the layout menu, go to Tools → Plugins and install
GeoLayout from the Available Plugins tab. Several available algorithms are meant to help further improve your layout like preventing node overlap (Noverlap) and label overlap (Label Adjust); rotating the network (Rotate), and scaling the network (Expand and Contract) to make it more spread out or condensed. Each specific layout parameter is nicely explained in Gephi (1) once you click on it or hover over it with the mouse cursor.

- Node options

Right-clicking on a single node or group of nodes allows you to perform a number of actions including deleting, moving or copying to a new workspace, selecting the corresponding row in the data, etc.

One interesting option is “Settle” (1) which locks the node to its position, so that applying layout algorithms will not change its position. To release a settled node select the option “Free” (2).

THE DATA LABORATORY TAB

The data laboratory tab contains the data tables for node and edge attributes. This is where you can import CSV files, export data tables from open network projects, modify or delete graph elements. You can change individual values directly by clicking on them...
and typing their new values. The Data Laboratory also provides a variety of ways to manipulate data columns. For instance, you can duplicate columns or copy values between them, as well as create new columns by combining existing ones based on certain rules.

THE PREVIEW TAB
This is where you configure the rendering settings to polish and finalize the look of your visualization. Use the right-side Preview Settings menu to change the rendering options: size, color, transparency, and other properties of nodes, edges, and labels.

If you want to preview a smaller sub-graph rather than the full network you can do it by the Preview Ratio setting (3). After any change the network preview must be refreshed. Note that the background color in Overview won’t match that of Preview you need to change it separately (4). Once you have tweaked the preview settings to your liking, you can save them for future use (2). Later you can restore them by selecting them from the drop-down menu (1).

SAVE AND EXPORT
When the network visualization process is finished, the project can be saved in one of multiple supported formats. We go to File → Export → Graph File and select your preferred format (e.g. JSON, GDF, GEXF, GML, GraphML, Pajek .net, Ucinet DL, CSV, etc.). To save as an image or PDF file, go to File → Export → SVG/PDF/PNG File. Op-
tional Gephi plugins give also the opportunity to export to an interactive web project that can easily be embedded in a website like Sigma.js. If you do not see it in the File → Export menu, you will need to install the SigmaExporter plugin from the Available Plugins tab of the Tools → Plugins menu.

DYNAMIC GRAPHS

Gephi allows you to explore longitudinal networks, as well as conduct some basic analyses on dynamic data. Gephi offers two alternative formats for dynamic data, Timestamp and Interval. After the dynamic data has been loaded or generated, you can click on Enable Timeline to begin using the dynamic features. Once you return to the Overview tab, you will be able to play an animated sequence of your network as it progresses over time. You can also calculate dynamic network statistics such as node degree.
11 Case studies – Actual data

11.1 Case studies

The use of Social Network is increasing day by day because of its simplicity and way of communication to other users, hence it is important to know the users involvement in the community. The actual involvement will provide the exact interest of the person. The information extracted from social networks requires some criteria to understand the important factors of the community. Social network communities on Facebook are examined using Gephi as visualization tool. The involvements of the persons in the social networks are measured with the help of centrality measure. Many social networking sites such as MySpace, Facebook, Twitter provides hundreds of technologies and attractive features for users that give main interest and practices. Some of these sites help to define views, interests or activities. Social networking analysis provides new information and different tools to connect, sharing of online videos and photos. These sites are considered to be a web based service in which the individual make their profiles and making a list of friends and other users who want to connect and share information in such a way that these connection vary from site to site. The backbone of social networking site provides a wide variety of feature that includes user profiles that display list of friends. Social networks are connected through friendship, common interests, financial exchange, knowledge or relationship of belief. The most popular terms of relationship are “contact”, “Friends “and “Fans”. Data in the social networking site can be explored through displaying nodes and ties, its attribute, color, sizes and other node properties. It provides a map that is made up of ties such as friendship. It also determines how much organization shows interaction and provides informal connection between users and also shows the connection and association between different users of an organization. Many social networking sites include visual representation of network that help to understand the network data and provides the accurate results of the analysis. Users’ profiles are considered to be an integral part of social networking site in such a way that profiles give the positive outcomes for their users and create a great sense of presence in which the users are kept updated. The common and important goal of online social communities is forming connection between users. These connections can be visualized using network tools such as Gephi. Gephi is software tool used to analyze the social
network graph. There are many different software tools that are available today for the utilization of network with their advantages, but Gephi facilitates visualization of network. It is open source software and free to use and analyze the network graph.

In this research, the first case study is examining Facebook as an online social network using Gephi as a visualization tool to monitor, evaluating and examining the Facebook friend’s network. The second case study is examining Twitter’s network. The examined datasets are obtained from Stanford Network Analysis Project (https://snap.stanford.edu/index.html).

- Facebook dataset
Facebook dataset consists of 'circles' (or 'friends lists') from Facebook. Facebook data was collected from survey participants using Social Circles app. The dataset includes node features (profiles), circles, and ego networks. Facebook data has been anonymized by replacing the Facebook-internal ids for each user with a new value. Thus, using the anonymized data it is possible to determine whether two users have the same political views, but not what their individual political views represent.

These statistics were compiled by combining the ego-networks, including the ego nodes themselves (along with an edge to each of their friends).


- Twitter dataset
The Twitter dataset consists of 'circles' (or 'lists') from Twitter. Twitter data was crawled from public sources. The dataset includes node features (profiles), circles, and ego networks.
11.2 Data formats

Gephi projects have .gephi format extension. The software can read many other file formats, including GEXF, GDF, GML, GraphML, Pajek NET files, and UCINET DL files. When GraphML is imported, GDF or GEXF files, Gephi should be able to import the position, color and size of nodes when these are present in the data. In order to open a matrix data file stored as CSV (File → Open), we need to be sure that the file is semicolon-separated. If there is a comma-separated CSV file, we simply open it in a text editor and replace all commas with a semicolon.
Gephi can also import Excel and CSV files with data in an edge list format. To import data that way, we need to create two files, one containing nodes and their attributes and another containing an edge list and edge attributes.

The CSV file containing nodes needs to include a column named "ID" containing unique node identifiers as well as any other node attributes you would like to include. The edge list CSV should include columns titled "Source" and "Target", containing the node IDs of the start and end node for each edge, as well as any other edge attributes we
would like to include. Gephi can recognize two more columns if we include them in our data: "Type" indicating the type of each edge (Undirected or Directed), and “Weight” containing the edge weight. To import the files, first we create a new project from File → New Project. Then we go to the Data Laboratory tab → Import Spreadsheet button. We select our file, and the appropriate table type from the As Table drop-down menu (i.e. Nodes Table or Edges Table). We import both files.

![Import spreadsheet dialog](image)

We click "Next" and check that each column of our data is displayed in an appropriate format.
11.3 Workflow with Gephi (case study 1)

As mentioned above, the first case study is implemented with SNAP Facebook dataset. The contained files are the following:

nodeId.edges: The edges in the ego network for the node 'nodeId'. Edges are undirected for facebook, and directed for google plus and twitter. The 'ego' node does not appear, but it is assumed that they follow every node id that appears in this file.

nodeId.circles: The set of circles for the ego node. Each line contains one circle, consisting of a series of node ids. The first entry in each line is the name of the circle.

nodeId.feat: The features for each of the nodes that appears in the edge file.

nodeId.egofeat: The features for the ego user.

nodeId.featnames: The names of each of the feature dimensions. Features are '1' if the user has this property in their profile, and '0' if the user doesn't have the property. This
file has been anonymized for Facebook users, since the names of the features would reveal private data.

So, the dataset contains various data and depending on what kind of knowledge we want to gain, we open and process the respective files.

Let’s assume we want to reveal some information for the ego network of the node 0. We are going to inspect its edges, the most important nodes which will probably be the influencers. From this raw data, we will create a project that incorporates a wide range of Gephi techniques and methods, resulting in a finished network graph that tells a compelling story. We launch Gephi and open the file 0.edges as the following figure shows. The graph type is undirected, because we have to do with Facebook data as mentioned before.
Once the data has been loaded in Gephi, we'll see the following network in a random layout:

![Network Diagram]

We need to get the data into some sort of layout that will help us to understand the network structure. The context window tells us that the network has 333 nodes and 5038 edges. The number of edges is almost 15 times bigger than the number of nodes. Given this information, I am going to begin with the Force Atlas algorithm, which is useful for small to medium-sized networks. Graphs are usually layouted with “Force-based” algorithms. The principle is that linked nodes attract each other and non-linked nodes are pushed apart. We'll see whether Force Atlas effectively displays the network, if not, then another option will be considered. We'll need to make that determination after seeing the results.
We locate the Layout module on the left panel, we choose Force Atlas and we click on Run to launch the algorithm. Through Layout Properties you can control the algorithm in order to optimize the graphic representation. We set the “Repulsion strength” at 200 to expand the graph and we type “Enter” to validate the changed value. The algorithm is run for almost 15 seconds and we can see that the new layout seems to be expanded.

So, it’s better to use a circular layout like ARF or Fruchterman Reingold. After running for 3 min the Reingold algorithm, we take the following layout.
The Fruchterman-Reingold layout which is another force-based approach has slightly different settings available. While we are still working with the defaults, it is possible to adjust settings for this algorithm, although not to the same degree as with the Force Atlas model. The primary adjustments we can make here involve the graph size area and the gravity. Thus, a dense network can be forced to spread out by manipulating the graph area rather than adjusting repulsion or attraction settings. Here, the graph quite clearly displays a number of clusters. However, the nodes of the main big cluster seem to be not very clear. Something else is very clear that this is not a single connected network. Instead, there are multiple cases where smaller sub-networks exist. This is going to influence some of our statistical measures, as we'll see momentarily. For instance, there is no way to calculate a single diameter measure, as it is impossible to traverse the entire graph.

Next, let's begin using some filters to better understand the network. Here are a few questions we can attempt to answer:

- Which nodes are the most influential, as measured by degrees?
- Where are the heaviest edges an indication of frequent collaborations?
- How large is our largest connected component, and who belongs to it?

We have several easy ways to measure degrees.

In the Ranking (color) module the node’s color and size can be configured. So, we locate the ranking module on the top left, we choose Degree as a rank parameter and we click Apply to see the result.
We can see rank values by enabling the result table. As we can see node 56 has 77 links and is the most connected node in the network.

We could also take the data outside of Gephi and calculate a degree value for each node, we can use the Gephi ranking function to size all nodes based on their individual degrees, or we could simply use the filters within the Topology folder to look at Degree Range.

So even though we don't have an explicit field for degrees, Gephi recognizes the network structure and lets us filter using the degree attribute. We can note that the degree range runs between 0 and 77, so let's examine all the values of 15 or greater.
We finalize the setting at 37 which means that now we only view nodes with 27 edges or greater. The graph now is clearly concentrated in one distinct area of the network. Now let's look at edge weights to see where the most frequent collaborations occur. We have multiple ways we can go about this, but in this we can easily note who the collaborators are by navigating to the edges table in Data Laboratory or even from the overview visualization.

If we apply the Giant Component filter to the network to understand what proportion of the network is connected in the largest area of the graph and we will not get many differences in the result. This means that the biggest proportion of the network is connected in the previous distinct area. As we can see, there is a main large connected component of 35 giant nodes with 438 edges which represents almost the 10% of the entire network. The isolated activity is confined. The main traffic is concentrated in this component.

Let's apply some statistical measures to the graph to understand patterns even better. Usually there is a difficulty in calculating diameter across the network, due to the many component groups. However, we still have the ability to run many statistical functions, but must recognize their limited meaning in certain contexts (such as a component with just five members). So our primary goal should be to examine the giant component and its member nodes, as this is where the most significant interactions are taking place. So, we will estimate the average path length for the network. It computes the path length for all possible pairs of nodes and gives information about how nodes are close from each other. We locate the statistics module on the right panel and we click Run near the Average Path Length.
The settings panel appears. We select the “Undirected” and we click OK to compute the metric. This metric computes the average graph-distance between all pairs of nodes.

From the report we exclude that the network diameter is 2 and the average path length is 1.26. General reports and results for each node are produced. The three new values that have been created by the “Average Path Length” algorithm are Betweenness Centrality, Closeness Centrality, Eccentricity.
After that we go to ranking module and select Betweenness Centrality in the list. This metric indicates influential nodes for highest value. The node’s size will be set now. We select the diamond icon in the toolbar for size and we set a min size at 10 and a max size at 45 as in figure below. We click apply to see the results.

We take the following graph.

Now we’ll experiment more with the network now that colors and size indicates central nodes.

- Display node labels
- Set label size proportional to node size
• Set label size with the scale slider

After making these settings we end up in the visualization below.

From this final visualization we can make some conclusions based on the questions mentioned before. The three blue nodes (56, 271, 67) seem to be the most influential nodes as measured by degrees. This happens because they are the most interconnected nodes in network, as we saw earlier at Ranking Degree table. These nodes have the most edges. Moreover this component of network, this cluster is the largest connected part and all the above nodes (circles) with the respective ids (numbers) belong to this component. All these nodes have also the most frequent collaborations in the whole network. The big part of traffic is done in this cluster. For privacy reasons the nodes are anonymous. In case we have the real node names, we can be lead to important conclusions for the key players of the examined network. This fact can lead our decisions. De-
pending on the KPI measures we make in our data, we take the respective decisions about various topics.

Before exporting the graph in a desired format (SVG or PDF), we go to Preview to see exactly how the graph will look like and make corrections if needed. We select the preview tab in the top banner and click refresh to see the preview.

We take the following graph.

In the node properties, we find the “Show Labels” option, we enable it and we click refresh to take the final graph.
11.4 Workflow with Gephi (case study 2)

As mentioned paragraph 11.1, the second case study is implemented with SNAP Twitter dataset. The Twitter dataset consists of 'circles' from Twitter and the data was taken from public sources. The dataset includes node features (profiles), circles, and ego networks. The goal here is to follow the same methodology with dataset from different social media network in order to see the results, differences etc.

So, as in the previous case the dataset contains various data. Depending on what kind of knowledge we want to gain, we open and process the respective file. The dataset contains all the files in the following figure. Again, we want to play with the edges in the ego network for the chosen node. As we said, edges are undirected for facebook, and directed for twitter and google plus. The 'ego' node does not appear, but it is assumed that they follow every node id that appears in this file.
Again, let’s assume we want to reveal some information for the ego network of the node 12831. We are going to inspect its edges, the most important nodes which will probably be the influencers. We launch Gephi and open the file 12831.edges as the following figure shows. The graph type is directed now, because we have to do with Twitter data and the relations between nodes are directed. In the following dialog we choose directed graph type.

Once the data has been loaded in Gephi, we’ll see the following network in a random layout:
We need to get the data again into some sort of layout that will help us to understand the network structure. The context window tells us that the network has 236 nodes and 2478 edges. The number of edges is almost 10 times bigger than the number of nodes. Given this information, once again we are going to begin with the Force Atlas algorithm, which is useful for small to medium-sized networks. As we have already said, graphs are usually layouted with “Force-based” algorithms. The principle is the same, linked nodes attract each other and non-linked nodes are pushed apart. We'll see whether Force Atlas effectively displays the network, if not, then another option will be considered. We'll need to make that determination after seeing the results.
We locate the Layout module on the left panel, we choose Force Atlas and we click on Run to launch the algorithm. We set the “Repulsion strengh” at 200 to expand the graph and we type “Enter” to validate the changed value. The algorithm is run for almost 15 seconds and we can see that the new layout.

We see that the graph is condensed, so we change the repulsion strength from 200 to 500 and we get:

If we run a circular layout like Fruchterman Reingold for 3 min with the default settings, we take the following layout:
Here, the graph doesn’t display any concentrated clusters. The connections seem to be almost equally distributed. We could say that one main big cluster exists although it isn’t very clear. Smaller sub-networks do not exist.

In the Ranking (color) module we configure node’s color and size. So, we locate the ranking module on the top left, we choose Degree as a rank parameter and we click Apply to see the result.

We can see rank values by enabling the result table. As we can see node 1186 has 80 links and is the most connected node in the network. The second most connected node is 528 with 70 links.
If we run the Degree Range filter we can note that the degree range runs between 0 and 80, so let's examine all the values of 20 or greater.

We finalize the setting at 25 which means that now we only view nodes with 25 edges or greater. Now let's look at edge weights to see where the most frequent collaborations occur.

If we apply the Giant Component filter to the network to understand what proportion of the network is connected in the largest area of the graph and we will not get many differences in the result. This means that the biggest proportion of the network is connected in the previous distinct area. We can also understand this from the visualization, because it’s clear. As we can see, there is a main large connected component of giant nodes with many edges which represents almost the big part of the entire network.

Let's apply some statistical measures to the graph to understand patterns even better. So our primary goal should be to examine the giant component and its member nodes,
as this is where the most significant interactions are taking place. The average path length for the network computes the path length for all possible pairs of nodes and gives information about how nodes are close from each other. We locate the statistics module on the right panel and we click Run near the Average Path Length. The settings panel appears. We select the “Directed” and we click OK to compute the metric. This metric computes the average graph-distance between all pairs of nodes.

From the report we exclude that the network diameter is 5 and the average path length is 2.05. The three new values that have been created again by the “Average Path Length” algorithm are Betweenness Centrality, Closeness Centrality, Eccentricity.

We go back to ranking module and select Betweenness Centrality in the list. The node’s size will be set now. We select the diamond icon in the toolbar for size and we set a min size at 10 and a max size at 50 as in figure below. We click apply to see the results. We take the following graph.
Now that colors and size indicate central nodes, we can experiment further with the network...

- Display node labels

- Set label size proportional to node size

- Set label size with the scale slider

After making these settings we end up in the visualization below.
From this final visualization we can make some conclusions based on the questions mentioned before. The two blue nodes (528, 1186) seem to be the most influential nodes as measured by degrees. This happens because they are the most interconnected nodes in network, as we saw earlier at Ranking Degree table. These nodes have the most edges. Moreover this component of network, this cluster is the largest connected part and all the above nodes (circles) in the figure with the respective ids (numbers) belong to this component. All these nodes have also the most frequent collaborations in the whole network. The big part of traffic is done in this cluster. For privacy reasons again the nodes are anonymous. In case we have the real node names, we can be lead to important conclusions for the key players of the examined network.

Before exporting the graph as a SVG or PDF, we go to Preview and follow the same procedure as in the first case study to see exactly how the graph will look like and make corrections if needed.

We take the following graph:
As we can see from the two case studies, the methodology is almost the same, but it varies on what we want to measure. For different measures we have to use different algorithms in order to visualize the network in the right way or to obtain the desired result. In these two cases the questions were the same:

- Which nodes were the most influential, as measured by degrees?
- Where are the heaviest edges an indication of frequent collaborations?
- How large was the largest connected component, and who belonged to it?

Since the questions were the same, the goal of the two cases was the same. Eventually, the methodology was also similar with few differences. In a pool of various data the same “kind” of data were used, the nodeId.edge files, for the same reason, to measure the edges in the ego network for a specific node. Moreover, the same layout algorithm was used although this could differ depending on the allocation of the nodes in the network, the same statistics, filters and ranking algorithms. So, this procedure could be a common methodology for manipulating such networks for this kind of purposes.
12 Future Development

With the help of new technologies the entertainment and broadcast industry is becoming increasingly competitive and the limitations are not distinct. The business model in the entertainment industry shifts toward Web delivered content, televisions are getting “smarter,” delivery channels are growing, and the applications continue to become more “social”. The rise of “over-the-top” television (OTT) and “TV anywhere” delivery will point to the direction of even further integration with social networks and provide more detailed data about consumption through the means of social media. The need to implement business intelligence and extract full value from strategies is bigger than ever, mainly because of the increased competition. Social measurement and analysis will become a basic part of intelligence applications and business strategies. Social media analytics can unlock great value content, when it is applied in the right way, allowing the organization to better engage the fan base, optimize marketing and make profits. The businesses that incorporate analytics will be rewarded in contrary to the companies that continue with traditional business models.

Till now social networks connect groups of people who have similar interests or habits. In the future, it is expected that social networks will also connect other entities, such as software components, Web-based services, data resources, and workflows. The interactions among people and nonhuman entities will be enhanced more. Understanding social networks will solve a big data problem of businesses that will need to enhance their marketing, sales, and online commerce. Taking out data from various social networks enables data analytics to correlate the different input data and finally have better preview of what is happening. However, the social networking vocabulary varies from one network to another and the need for cross-domain vocabulary mapping as a data preprocessing step is really important. For example, the Twitter glossary defines terms such as “followers” and “tweet”. Facebook defines terms such as “friends” and “status”. Google Plus uses “circles” and “hangout”. To perform cross-domain data analytics, a common practice shall be implemented that will capture the differences and similarities in terminologies and define relations between terms within the network.

Although a big part of the job has been done regarding the software, further work is needed for the development of new features and tools. Improvements will be integrated to the data structure to support grouping and navigation within a network hierarchy.
Gephi is trying to make a more user friendly interface in order to adapt it to the users’ need. Gephi has also tried to speed up the analysis process, from data import to map export and has given a lot of attention to the development of dynamic features.
13 Conclusions

Every day, users offer feedback and engage in online conversations about businesses on social media sites. Businesses can use social media monitoring and analytics tools to find, sort and analyze that data. They offer the ability to identify patterns in customer sentiment and behavior. Today’s businesses must embrace social media, because they provide a low-cost platform on which they can build their brand and they communicate effectively with stakeholders within and outside your company. They allow them to interact rapidly and simultaneously with peers, employees, customers, and the broader public, especially with younger generations. Social media represent a modern platform that enables interaction among various participants.

Because most of the data both originates and resides in the Internet, one basic challenge is determining how Internet computing technology should evolve to let users access, assemble, analyze, and act on big data. Especially in social media case, the access to datasets that might be useful for analysis seems to be difficult task, due to restrictions that the social networks place, mainly for copyright or commercial reasons.

Moreover, social media programs lack of specific and comprehensive strategy. Most departments within an company are acting independently and follow their own social strategies to benefit particular business units. For instance, a marketing department can have its own Facebook page and to promote a show. This is not the right procedure. A well defined plan that puts social media analytics into a broader enterprise context and a set of technologies that can effectively support that process would be the correct procedure. Otherwise companies can be trapped on isolated insights on social networks. Treating social media data in this way is a big mistake. Social media is just a part of the broader procedure.

It is very important for the businesses to discover the most effective ways they can utilize social media tools, and why those tools work for them. Advanced mathematical modeling will be needed to form a conceptual and contextual understanding of information in any format and language. This will enable the automation of many key functions that are vital to handle knowledge and social media content. The appropriate methodology requires tracking, measuring, analyzing the data and adjusting the model.
Regarding to the network analysis in Gephi the more difficult task is to realize and choose what we want to visualize and how to move from a simple idea to a complete graph.

Further work is required for the development of cross-domain data analytics. A common practice that will capture the differences and similarities in terminologies and define relations between terms within the network must be developed.
Bibliography - References

There are a lot of papers available on the Web dealing with various aspects of social media analytics, as well as multiple books available in PDF versions. The following list provides some of the resources I used in order to write this thesis.

[1] Rick Lawrence, Social Media Analytics: Strategic Use of Advanced Analytics in the Public Sector, CMU Heinz / IBM Joint Workshop, April 13, 2011.


[25] *Collecting the data to build your Facebook network*, June 2015, https://linkurio.us/

## Appendix

<table>
<thead>
<tr>
<th>Chapters</th>
<th>Content</th>
<th>Length (pages)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cover</td>
<td>Social Media Analytics in the entertainment industry</td>
<td>1</td>
</tr>
<tr>
<td>Title Page</td>
<td>Social Media Analytics in the entertainment industry</td>
<td>1</td>
</tr>
<tr>
<td>Abstract</td>
<td>Abstract</td>
<td>1</td>
</tr>
<tr>
<td>Contents</td>
<td>Contents</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Literature Review</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Problem definition</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Potential of Social Media Analytics</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>Analytics challenges</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>How the Media &amp; Entertainment industry uses insights from social data</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>Harnessing social data to build actionable insight</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>SMA procedure and technical workflow</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>Analysis tools and libraries – Techniques</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>Gephi Graph analysis software</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>Case studies – actual data</td>
<td>26</td>
</tr>
<tr>
<td>12</td>
<td>Future Development</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>Conclusions</td>
<td>2</td>
</tr>
<tr>
<td>Bibliography</td>
<td>Bibliography – References</td>
<td>2</td>
</tr>
<tr>
<td>Appendix</td>
<td>Appendix</td>
<td>1</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>87</strong></td>
</tr>
</tbody>
</table>