POLITECNICO DI MILANO

Scuola di Ingegneria dell'Informazione

POLO TERRITORIALE DI COMO

Master of Science in Computer Engineering

User Engagement Analysis of Live Events based on Social Media Monitoring

Supervisor: Prof. Marco Brambilla

Master Graduation Thesis by: Roksana Jafari

Student ID Number: 10419544

Academic Year 2015/16

Acknowledgment

My special thanks goes to Prof. Marco Brambilla for his patience, intuition and discipline in leading this research.

I would like to thank Riccardo Volonterio for his valuable suggestions and help in the practical part of my thesis.

My gratitude goes also to all my dearest friends, who were inspirational and supportive during my experience in Politecnico.

A special thank you goes to my family, for their constant encouragement and their loving hearts, without them I wouldn't be who I am today.

I would like to also thank Politecnico di Milano for this amazing experience. I am really blessed to be a part of Polimi's family.

Abstract

Each day more than millions of posts are published on social media networks and their active users are increasing exceedingly. These networks are massive sources of user-generated data, which can be used to extract valuable information in many fields. Some of these networks grant access to their data through their APIs that facilitates many researchers to mine their data for the sake of analysis.

In this thesis we tried to come up with innovative analysis approaches in dealing with cross-related events based on the aggregated data from social media networks. We studied a program that is a series of events, which took place in Milan during Expo 2015. The program is a collection of different events that feature both diversification points and mutual characteristics. The program is clustered into 4 groups of events (Performance, Art, Media, Science) and its dataset was gathered by probing into Instagram and Twitter APIs. The metrics that we used for assessing the data are based on the previous analysis techniques that have been applied to data derived from social media. Our endeavor in this thesis was focused on performing the analysis in different stages of the program to understand the circumstances of each event and their connections with each other. The types of analysis that were applied include text mining, clustering, topic analysis, engagement analysis and correlation analysis. We extracted the temporal stream of the given dataset, top trends of the events, features indicating user commitment, and correlation between events based on their similarities. After that we employed a flexible and scalable visualization approach for modeling the outcomes.

The results of the process of social media analytics gave us great insights on how each element of such a program was carried out, like events, their specifics and connections, and also how did these elements behave with respect to one another.

Sommario

Ogni giorno milioni di messaggi vengono pubblicati sui social network e gli utenti attivi continuano ad aumentare esponenzialmente. Queste reti sono enormi fonti di dati generati dagli utenti, che possono essere utilizzati per estrarre informazioni preziose in molti campi di ricerca. Alcune di queste reti consentono l'accesso al dato attraverso le rispettive API, che facilitano a molti ricercatori il compito di estrarre i dati a scopo di analisi.

In questa tesi si è cercato di individuare approcci di analisi innovativi per il trattamento di eventi cross-basate su dati aggregati provenienti dai social network. Abbiamo studiato un programma, rappresentato da una serie di eventi, che ha avuto luogo a Milano in occasione dell'Expo 2015. Il programma è una raccolta di diversi eventi che caratterizzano entrambi i punti di diversificazione e le caratteristiche comuni di investimento. Il programma è di tipo cluster ed organizzato in 4 gruppi di eventi (Performance, Arte, Media, Scienza) ed i dati sono stati raccolti utilizzando le API di Instagram e Twitter. Le metriche che abbiamo utilizzato per valutare i dati si basano sulle tecniche di analisi precedenti, applicate a dati derivati dai social network. Il nostro sforzo in questa tesi si è concentrato sull'analisi in diverse fasi del programma per comprendere le circostanze di ciascun evento e le loro interconnessioni. I tipi di analisi che sono state applicate comprendono l'estrazione di testo, il clustering, l'analisi dell'argomento, l'analisi dell'impegno e di correlazione. Abbiamo estratto il flusso temporale del set di dati, il top trend degli eventi, le caratteristiche che indicano l'impegno dell'utente e la correlazione tra gli eventi in base alle loro somiglianze. Dopo di che abbiamo adottato un approccio di visualizzazione flessibile e scalabile per la modellazione dei risultati.

I risultati del processo di analisi dei social network ci ha fornito ottime intuizioni su ogni elemento di tale programma, rispettivamente ai singoli eventi, le loro specificità e le rispettive connessioni.

Table of Contents

Chapter 1

Introduction

1.1 Context

The rise of web 2.0 empowered many online social communities to grow day by day and brought many more into existence, which boosts the process of creating, sharing and following information on social media networks. Nowadays online social networks (OSN) are the new phenomenon on the web and hold most of the web traffic as well. The increasing users generate terabytes of data every day, and this data appears in many different forms such as textual status updates, images, videos, links and etc. Understanding the dynamics of such networks, tracing the trends and their user actives, is very important in many fields such as marketing, management, politics and etc.

Many of the social media companies make their Application Programming Interface (API) accessible to third parties, as a part of their business model. In this scenario researchers are able to retrieve and modify digital data using simple software scripts. This way OSNs contribute to the progress of collecting data. Therefore such progress is not barricaded with the ability to collect or store the data, but by the ability to handle, analyze, abridge, envision and detect desired knowledge from that collected dataset, also considering the obstacles of time limitations and scalability. That is where the social media analytics emerges to contribute and accelerate the process.

When our dataset refers to a series of real life events, the problem is to select or define the different metrics in all stages of social media analytics (listening, analysis, visualization). In order to achieve efficiency and avoid the misapplication of time and resources these metrics have to be selected wisely.

1.2 Objectives

The objective of the process is to increase the visibility of a program including live events, through social media and assess the feedback of its audience. We aimed to define appropriate measures to analyze a live event scenario where there is an integrated program of multiple events, possibly diversified for duration, topic, location, and genre, although featuring some kind of cross relations.

The idea is to attain sensible insights from the diverse events that shape up a bigger plot, and study them both individually and as a whole, to understand how they behave and correlate through time and what are their main connections and impressions toward one another.

We applied our approach to an experimental scenario based on 13 events in a program that is held in a specific period and can be divided to sub categories corresponding to the subjects involved. The dataset is derived from social media APIs during the time of the program's occurrence. The events are not analogous but despite their diversifications, there are some cross relations between them. They are clustered in relevant categories. Each event has a timestamp which is particular, yet there are some overlaps throughout the whole program.

Our approach is to make the process of storing, analyzing and visualizing data more efficient. Therefore, we use a NoSQL database because of its flexibility and speed in dealing with unstructured data that was collected from APIs. Then we define some steps to perform the analysis in different levels. The first phase of analysis is to carry out some general and quantitative measurements about the entire program and the stream of data about live events on social media networks, to get the overall impression of the subject. Then we perform different analysis on other levels, for example on the clusters of events to observe how the elements of the clusters behave together and with respect to other clusters. We perform topic analysis, with text mining to understand the highlights of the program, user engagement assessment and investigation of relations between events and their categories. Finally we model the outcomes of all levels of analysis in order to have a proper visualization of the results.

1.3 Thesis Outline

The thesis is organized as follows.

First there are some definitions that were stated in chapter 2 to better elaborate the process based on them. As a background to out study experience, we reviewed the structures of data and the definition of social media analytics and the necessity of it application. Also the stages are briefly explained except for social media monitoring stage, which explains in details what kind of data, we used and how did we collect it from social networks. In addition we added some details about Twitter objects to help in comprehension of the following chapters.
In chapter 3, we reviewed briefly the scientific works that have been done to address the similar issues on social media analysis with respect to their categories, plus our own strategy with respect to them.
Chapter 4 is dedicated to describe our objectives that we want to address with SMA. As there is a different range of audience to this field, we propose different methodologies in analysis to meet their needs. Our proposed metric system includes temporal analysis of data, measuring user engagement in different scales, topic analysis and correlation analysis while dealing with a series of live events.
Our assets and technical methods are described in chapter 5. The sections are mostly concentrated on the methods to extract proper data from MongoDB. Also describing the technique we used to visualize the results of the analysis.
Chapter 6 is devoted to describe our dataset and the outcomes of the analysis that was performed. It is divided into sections that are relevant to each level of analysis, conducted on our dataset from the social media we used.
Finally we review the study with a short summery of what has been done and a discussion of our results. In addition there are some suggestion for the future work, where our approach would be a proper choice as an appliance in other fields.

Chapter 2

Background

This section is devoted to the useful materials that were used to conduct this study. The first section is about how data structures evolved along with rising technologies and the reason why we need to consider new approaches in data storage for new formats of data. The second section discusses the concept of social media analysis (SMA), its process and stages, and it's necessity in today's world. The last three parts are sub sections of SMA, however in the first stage there are more clarifications on what are APIs and how to extract data from them. In particular as this study focuses more on twitter data, there is a section defining elements of this network and their functionality.

2.1 Evolution of Data

The most intriguing question is why do we need evolution in data processing?

The answer lies in the fact that we all use more devices to produce and save all forms of data that are much greater in volume than before, and were non-existent a while ago. Until now files were comprised of just plain text, they were easier to filter and store. However now, as shown in the figure 1, users tend to deal with rich data types, which include pictures, music, movies and all different data formats. The rich data offers a much better user experience but takes up more space and as it increases at a higher rate, it is much more difficult to handle and store.

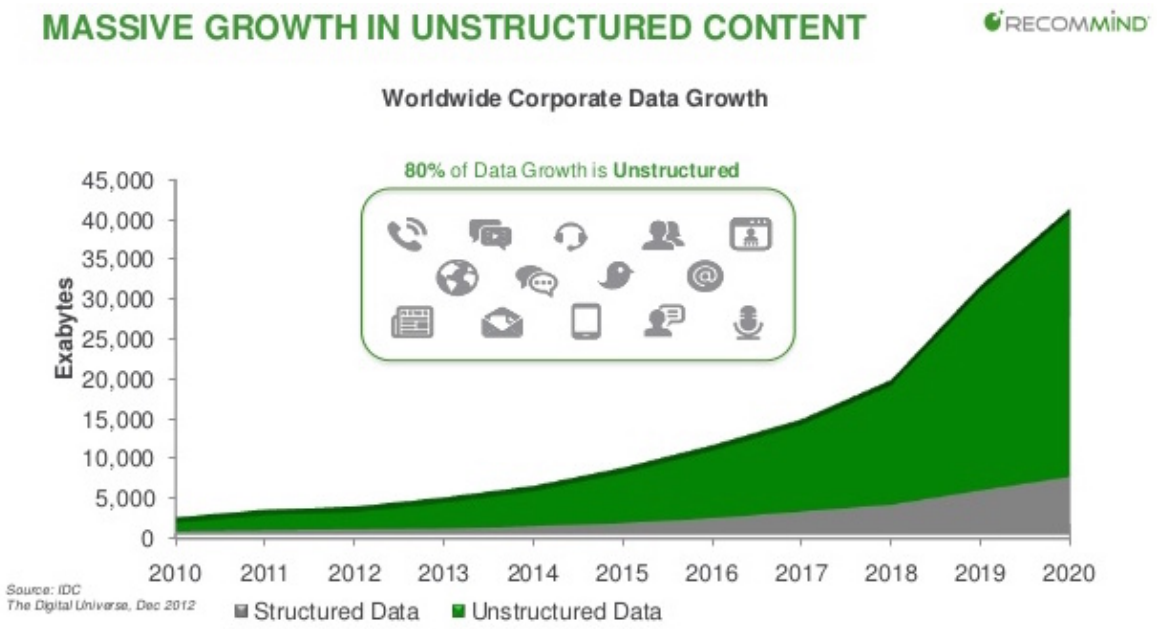

Figure 1: Growth of different formats of data during time

In traditional BI platforms, the flow of data – starting with its attainment from source systems through to transformation, integration, analysis, and reporting – follows a structured sequential process, as in Figure 2 (MongoDB inc., 2015).

Figure 2: Traditional BI Process

The data used was referred to as structured data that was mainly text and could be classified and stored easily in relational databases. In these databases the data is well organized in columns and rows and it is related through key values. This property makes it easy for data mining tools and query languages like SQL to navigate and retrieve desired outcomes.

However with the advancement of web technology and social media, there are many new formats of data available to the customers, who come from different sources like in figure 3, and which is a potential massive source of information to analysis as well.

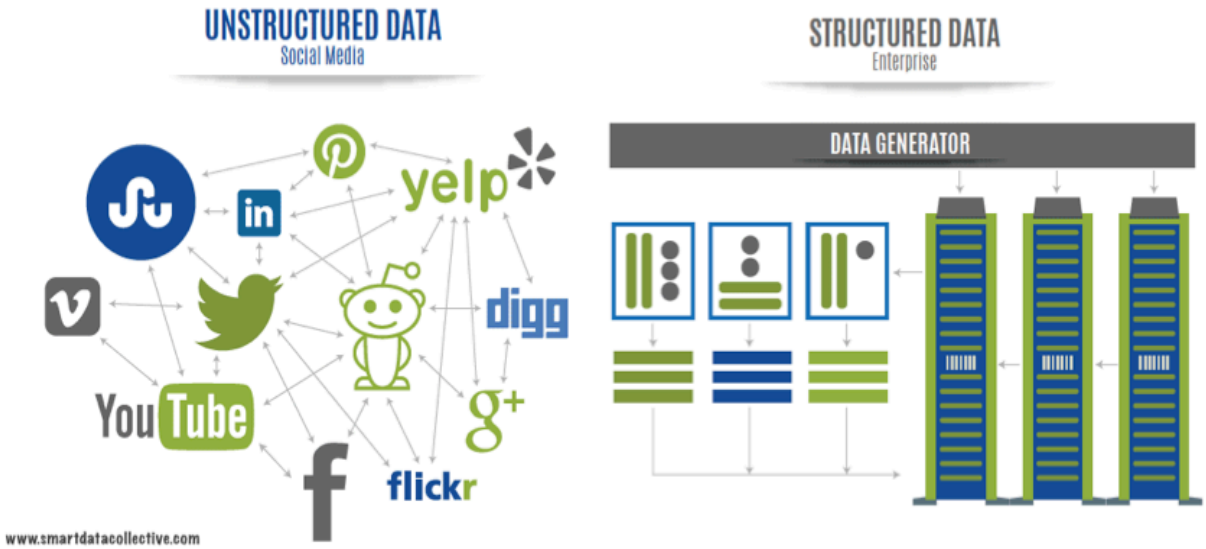

Figure 3: Structured vs. Unstructured Data

"Big Data", also described as unstructured data has characteristics such as the ones in figure 4 that makes it impossible to be stored in rows and columns in a relational database. Storing data in an unstructured form without any defined data schema is a common way of just filing database with information. Examples for unstructured data are the ones from web pages that contain images, texts, videos and etc. altogether. The advantage of unstructured data is that no additional effort for its classification is necessary. A limitation of this kind of data is that controlled navigation within unstructured content is impossible (Sint, Schaffert, Stroka, & Ferst).

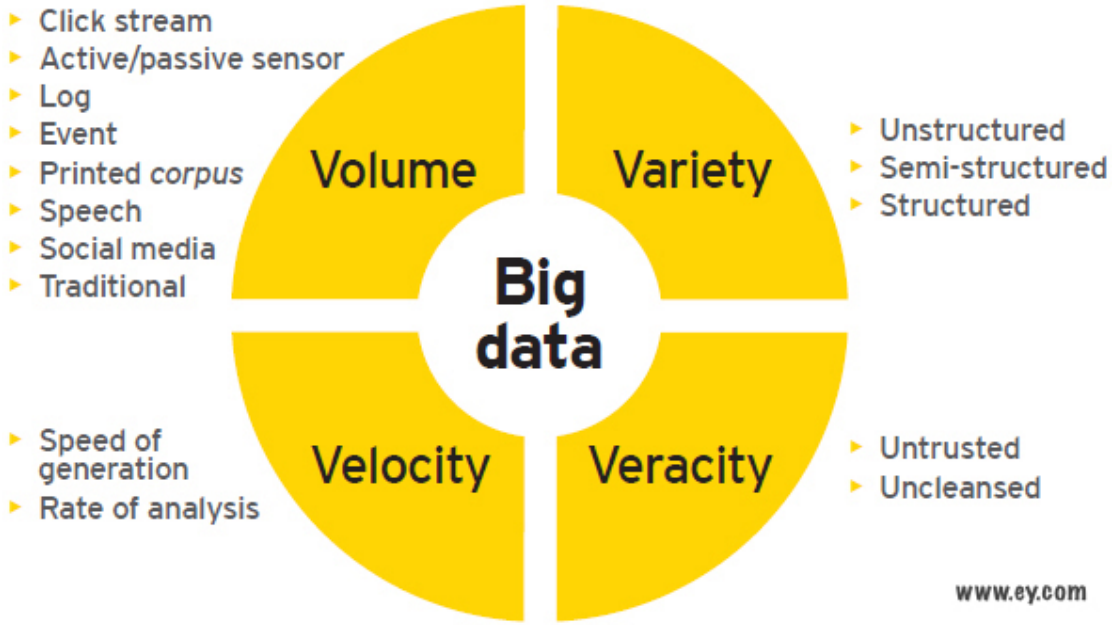

Figure 4: Big Data Characteristics

So in response, new processing methods came into existence like noSQL databases that allow data to be grouped together and diminish the previous restrictions.

MongoDB is noSQL database that models and stores rich data as documents in a binary representation called BSON that extends the popular JSON (JavaScript Object Notation) representation to include additional types. Table 1 shows the relationships between RDBMS terminology with MongoDB.

RDBMS	MongoDB
Database	Database
Table	Collection
Tuple/Row	Document
column	Field
Table Join	Embedded Documents
Primary Key	Primary Key (Default key _id provided by mongodb itself)
Database Server and Client	
Mysqld/Oracle	mongod
mysql/sqlplus	mongo

Table 1: RDBMS vs. MongoDB terminology

2.2 Why Social Media Analytics

In this millennium, social media can be considered one of the greatest sources of unstructured information. It is designed to be disseminated through social interactions (Leskovec, 2011). Users are transformed from passive audiences who browse media, into active prolific authors that consume and create contents. Also, social media platforms enable them to share and contribute to each other's contents and express their opinions or create a link to other users or contents. The interaction is documented as in comments, retweets, shared data, forums and etc. This significant information enables us to mine user behaviors and opinions and gain insights in diverse fields.

- In marketing analytics tracking the pulse of the social media outlets enables companies to gain feedback and insight in how to improve and market products better (Fonseca , Salvador , & Nogueira , 2015). Business brands want to know who is the target client and where, when, how and what are his needs. Plus it is more convenient and economical to trace the feedback of a client with analyzing their online statements.
- For consumers, the availability of information and opinions from many different sources helps them make more informed decisions. They are more likely to trust peer recommendation e.g. one can simply go through reviews of other customers or suggestions of a friend rather than a lifeless catalogue.
- As far As politics and social studies go, social media is a much more enriched source than traditional surveys, and solicits the latest citizens opinions and feedbacks on any desired level, e.g. why does a nation support a trend or what are the issues being raised at specific moments.
- Social media offers a term called "real-time journalism", it provides us with more valuable information than news channels, as every user has the opportunity to record and report online. However it comes with problems like redundancy, which a suitable analysis can form it into noteworthy information.

As you can see in figure 5 social media analytics involves a three-stage process:
The capture stage involves acquiring pertinent social media data by monitoring or "listening" to different SM sources, soliciting and archiving relevant information. This procedure is either done by a company or through a third-party vendor. The understand stage picks out useful data for modeling, filters out the noisy low quality data, and employs various analytic methods to analyze the data retained and attain insights. The present stage deals with displaying findings from the previous stage in a meaningful way.

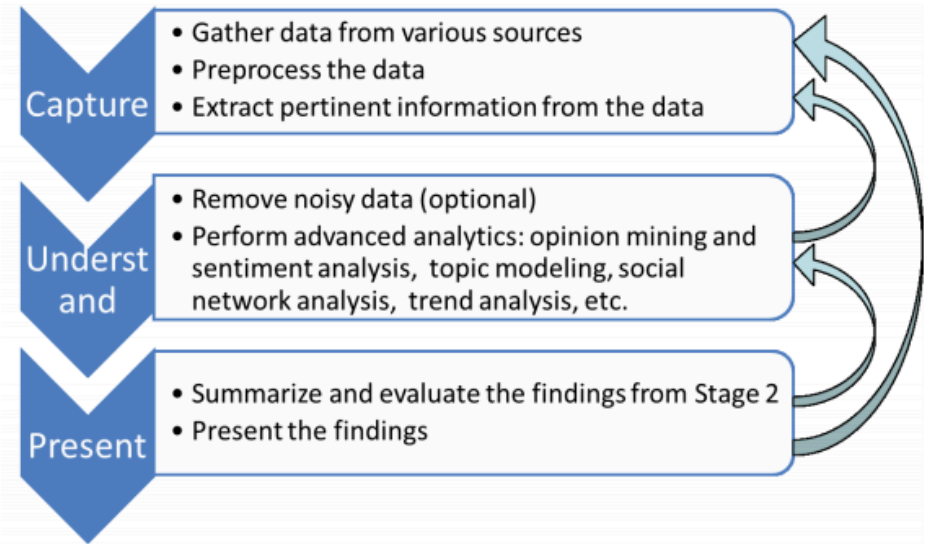

Figure 5: Social Media Analytics Process

However, there is some overlap among the stages. For instance, the understand stage creates models that contribute to the capture stage. Also, analytics support human judgments that complete the understand stage, in addition to help in the present stage. These stages are directed in a continuous, iterative matter rather than strictly linearly. A social media analytics system goes through much iteration before it becomes truly useful (Fan & Gordon, 2014).

2.3 Social Media Listening

Technically the listening part is an endeavor to capture conversations on social media platforms related to activities or interests that are more relevant to our pre-defined goals. The capture stage covers popular platforms such as Facebook, Twitter, LinkedIn, Pinterest, Google +, YouTube, Foursquare, Tumblr, etc. As well as smaller, more specialized sources such as Internet forums, microblogs and blogs, Wikis, picture sharing sites, podcasts and news websites.

The two main flow of information regarding social network originate from newswire and social media, which are mostly, open source. And it is only the matter of using the appropriate platform and tool to extract desired information. Collecting massive amounts of relevant data across hundreds or thousands of social media sources is done using news feeds, application programming interface (APIs), or by crawling the web. (Fan & Gordon, 2014)

Tremendous amounts of data are recorded to meet various needs. Among this data, which is, user generated, not edited and not authorized, there is a great volume of irrelevant data. In order to build a data set for the analysis phase, different pre-processing steps might have to be performed, including data modeling, data/record linking of data from different sources, feature extraction, stemming, part of speech tagging, and other syntactic and semantic operations that support analysis.

It is a fact that that 80 percent of business-relevant information lies in unstructured format, primarily text. (Grimes, 2008) Here the need of a tool to navigate through them and elicit the data, store and sort it in a formidable manner has become more vital especially in BI.

2.3.1 Data extraction from APIs

Application **P**rogramming **I**nterface is an interface with URLs as the controls. Web APIs are a way to shirk all inessential visual interfaces that you don't care about and get at the data. They are considered as a limited shortcut into a web service's database. When it is not possible to access the internal database of a social network, APIs provide us with an even easier and less time consuming way to access it. Some popular free web APIs are Yelp, Twitter, Flickr, Instagram, Tumblr, Foursquare LinkedIn, Vimeo, Facebook, Google+, and YouTube, which most commercial analytical dashboards benefit from.
Usually there is a set of documentation called API specification, which is an instruction manual and explains what all the controls do. What they mainly describe is a list of URLs you can use to retrieve data. API request or API call is a term referred to using these URLs, to call many things, like resources or methods.

Each API call supplies a URL (possibly with extra parameters), and get back a response. Parameters are information transmitted as a part of URL; they define what is needed, in what format, in what quantity or range and etc. Then in case of no failure, some data as response is given back in the format of XML or JSON file.

The limitations are concerning Rate Limit and Authentications. Rate limit checks how many requests are received from one user in order to prevent overloading their servers.
Some APIs are genuinely public, but these days, most APIs require some sort of authentication, in the form of an API key, a long string of letters and numbers that functions like a password. (Veltman, 2013)

As our study mainly focused on data extracted from twitter APIs, the dataset and process of its collection is elaborated in the next part.

2.3.2 Twitter APIs

As in Figure 6, APIs to reach Twitter data can be categorized into two forms based on their design and access strategy:

- *REST APIs* are based on the REST architecture now popularly used for designing web APIs. These APIs use the pull strategy for data retrieval. To collect information a user must explicitly request it.
- *Streaming APIs* provides a continuous stream of public information from Twitter. These APIs use the push strategy for data retrieval. Once a request for information is made, the Streaming APIs provide a continuous stream of updates with no further input from the user.

Twitter APIs can be accessed only via authenticated requests. Twitter uses Open Authentication and each request must be signed with valid Twitter user credentials. Access to Twitter APIs is also limited to a specific number of requests within a time window. These limits are applied both at individual user level as well as at the application level. (Morstatter , Kumar , & Liu , 2013)

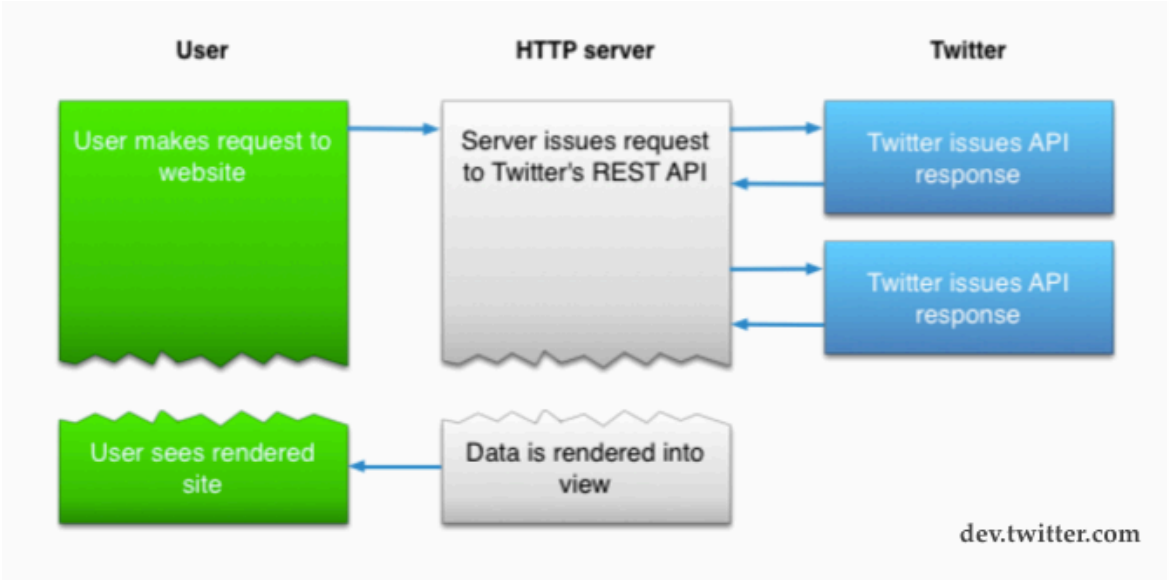

Figure 6: Twitter API functionality

2.3.3 Twitter objects

There are four main "objects" in the Twitter API: Tweets, Users, Entities and Places. By understanding the composed elements of each and the appropriate URL request, it would be a very straightforward task to extract what is needed.

- *Tweets* are the atomic foundations of Twitter network and are known as "status updates". They can be imbedded, liked, unliked, replied to and deleted.

- *Users* can be anyone or anything. They are capable of tweeting, creating lists and following other users. They own a timeline to share and post tweets. Also users can be mentioned and looked up by their peers.

- *Entities* provide metadata and extra contextual information related to the content posted on Twitter, for instance the structured data from tweets including media, hashtags, mentions and resolved URLS, without having to parse the text.

- *Places* are named locations with relevant geographic coordinates and are attached to tweets by a unique ID. Places are not necessarily the exact location of the issued tweet but usually are somehow relevant. It is also possible to search a place.

2.3.4 Twitter Dataset

On Twitter there is a mix of information buried under the concepts of objects previously mentioned. Either by employing programming language to query APIs, or using Twitter search engine, or commercial tools, we are capable to attain data from all those objects as shown in table below:

USER	• Real name • Twitter handle (screen name) • Location • Textual description of the user and his interests • Number of tweets published per user • Network activity information (number od followers and friends) • A verification sign showing if user has been externally verified by Twitter • Profile creation date • URL pointing to a more detailed profile of user
TWEET	• Location • Date • Number of likes • Number of retweets • Replies
ENTITY	• Media • Hashtags • Links
PLACE	• Location name • Address • Geographic coordinates

Table 2: Accessible Elements from Twitter Objects

2.3.5 Twitter Interactions

Most of the data transmission through media is motivated by the behaviors that users have toward one another. These interactions vary from one network to another. In case of twitter there are different ways that users can show their interest or opinion about a tweet.

There is a public profile for each user on Twitter that contains information such as the language, the number of people who follow and are followed by that person and a brief description that indicates the user's occupation and interests. This basic information will give the viewer the first impression of that user and helps him in deciding the further potential interactions. Then a Twitter user can post a tweet or follow the tweets posted by another user. This system builds up a public network of people, that can be followed and / or follow other users.

A tweet is a text message that is limited to 140 characters. There are some features that can be added to such a format when posting it on Twitter. A user can choose to add links to other objects of the network or to external entities in the web, mention other users in his network and use expression called hashtags. Hashtags are combinations of letters with the # symbol in the beginning. These expressions assist in defining the main topics and the highlights in the stream of Twitter data.

After a tweet is posted, it is visible to everyone in the user's network. So other users are able to react to that tweet by marking it as their favorite, writing a comment about it or sharing it. Tweets either can be shared on the same platform as a Retweet which helps spread it through Twitter, or it can be shared on other social medias, simply by sharing the link to that specific post.

In all those cases these reactions to the tweet message can lead us through many steps of analysis to check the engagement of a user or a group of users.

2.3.6 Dataset Storage

As discussed earlier about the data structures, this data explosion requires new data storage paradigms. And NoSQL databases are the pioneers offering easier accessibility. There are several implementations in NoSQL databases, however this study employed and focuses on MongoDB. The notable advantages of MongoDB are:

- *Document-Oriented Storage* – as shown in figure 7 enables database to store its data in JSON-style files. This way raw documents from Twitter's APIs can be stored easily.
- *Index Support* allows for indexes on any field, which helps in creation of optimized indexes for an application.
- Straightforward *Queries*
- *Speed*

Moreover, it works well in a single-instance environment, meaning that it can be set up on a home computer to run examples. (Morstatter , Kumar , & Liu , 2013)

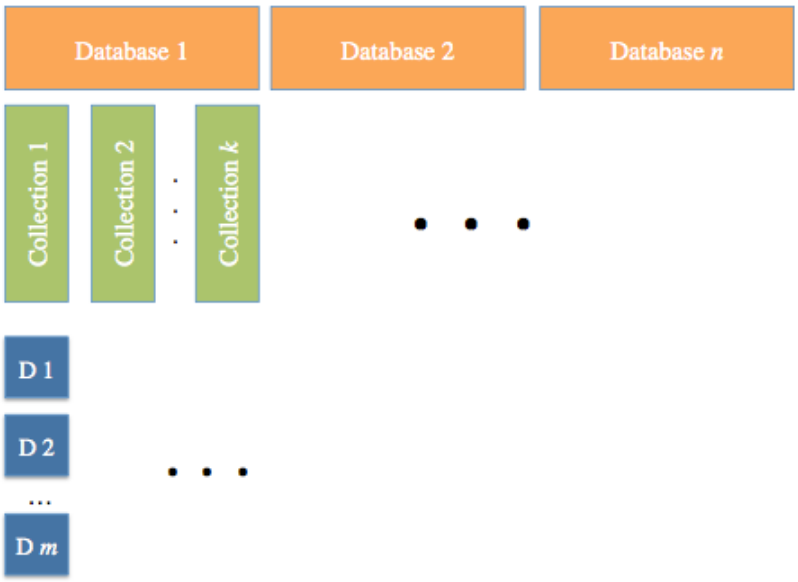

Figure 7: Structure of MongoDB data in documents

2.4 Social Media Analysis

Once the required data is collected, it is time to assess the meaning and generate metrics that helps best in the process of decision-making. This is the understand step as mentioned before. Since in the first step data is gathered from many users and sources, a great portion may be noisy and in need of filtering before analysis.

As we go forward analytics is become a vast area that includes a variety of modeling and analytical methodologies from different fields. As follows, there are some examples, most instrumental in the stages of SMA process.

Sentiment analysis and trend analysis are mainly applied in analysis level. Social network analysis and topic modeling are employed in analysis level, also supporting the listening and presentation stages. Visual analytics concerns the comprehension of data in the present stages. (Morstatter , Kumar , & Liu , 2013)

- *Opinion mining (or sentiment analysis)* is one of the core techniques that leverages natural language processing (NLP), computational linguistics and other methods to delve into users opinions and their sentiments from text sources. Such information conveys major insights in prediction and analysis in domains like top trends, market movements and etc. It develops a lexicon as a sample, which is a collection of words to later examine the data with it – figure 8.

- *Topic modeling* is the concept of filtering large contents of captured text to detect leading topics and trends. The founded themes can be employed to provide labels in exploring text collection. It is a form of text mining that gives an insight on what are the needs of customers or what are the top most popular subject in a field or event and many more similar situations.

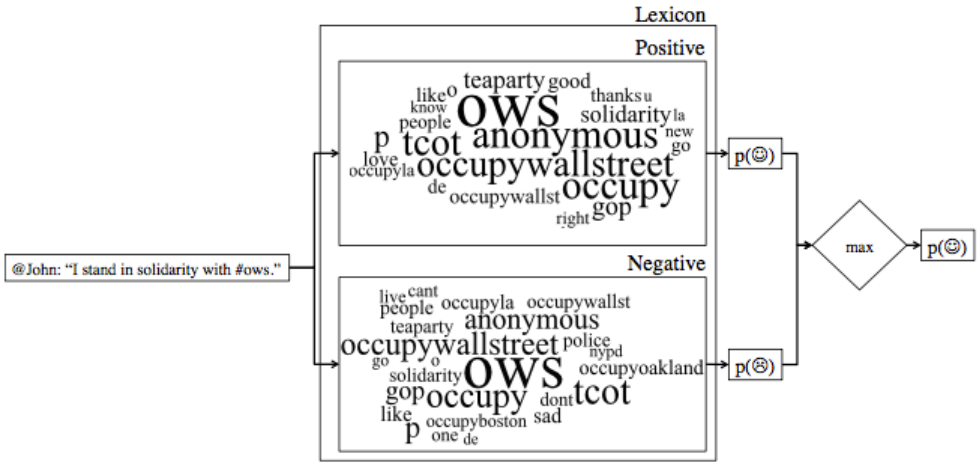

Figure 8: The sentiment analysis workflow

- *Social network analysis* is applied in analysis of a social network graph, which consists of nodes (users) and relationships (edges), to better understand its structure, connections and relative importance of nodes and model social network dynamics and growth.
  It is the main technique in identifying key influencers and sub-communities on social media platforms.

- *Trend analysis* predicts future results or behaviors based on historical data collected over time. It employs statistical methods such as time-series analysis or regression analysis and other more recent modeling techniques such as neural networks and support vector machines.

- *Visual analytics* involves activities, from data collection to data-supported decision-making. Visual analytic systems must be able to process data to reveal their hidden structure as well as their detail. Computational methods for data reduction, displaying correlations among disparate data sources, and allowing the user to physically manipulate data displays, all underlie visual analytics. . (Morstatter , Kumar , & Liu , 2013)

2.5 SMA Presentation

The presentation stage is the ultimate step in the social media analytics process. The outcomes of various analytics will be abridged, assessed, and presented to the audience in an understandable format.

Many visualization techniques can be used to demonstrate information depending on the evaluation and design criteria. One of the most common and popular interfaces that are widely used by both individuals and companies are the visual dashboards, which aggregate and show information from chosen sources. The commercial dashboards seem to be playing an important role as the todays BI components, which brings together all steps of social media analytics. The advantages of such technology is that it does not require previous programming knowledge, meaning that even an ordinary client can use it to visualize his data through the predefined metrics of these tools. They give a suitable grasp in data intelligence to managers or coordinators. Yet there are some disadvantages that make it troublesome to some clients such as exposing valuable data, inability of dynamic modification and being costly when it comes to a higher level of complexity in analysis.

On the other hand, professional type of users prefer to use programming language to deal with raw information and have unconditional access to data with much more flexibility in defining their desired frame. There have been many improvements in this field along with the blast of social media and many programming libraries are being developed to convert such massive data into beneficial diagrams that point out valuable findings.

Chapter 3

Related Work

Twitter has evolved from a popular website into a rich source of notable data, as a result, different researchers dedicate their resources to analyze twitter data for varied purposes. During this process they leverage various analysis mythologies. This chapter focuses on the different types of analysis that have been done before. These metrics gave a basic understanding and in some cases a meaningful insight to the development of our study. They are discussed fairly in each category with respect to their attribution to our analysis. Finally there is a brief section that addresses our approach as an extension to what we have reviewed.

3.1 Clustering and Correlation

Clustering events is one the basic approaches to study them in details and understand their contrasts and similarities as a group. This gives them an idea of how the events behave as a cluster and what are their connections with the others in various dimensions and platforms. Among those researchers who exploit analysis to cluster and model events are Akcora and Ferhatosmanoglu (Ferhatosmanoglu, Akcora,, Bayir,, & Demirbas, 2010) who presented an efficient way to determine public opinion on a temporal dimension, and extracted important news about the events effectively, along with an application where users can review news and explore the related articles on Web. They used Emotion Corpus Based Method that is based on vector space model for calculating document similarity and tweet count per every interval.

Authors of (Becker, Iter, Naaman, & Gravano, 2012) described a query-oriented solution for extracting social media documents for planned events across various social media platforms. Their work takes an important step in the procedure of organizing social media information for events, towards improved browsing and search for event media. Using a combination of precision-oriented and recall oriented query methods, they presented how to automatically and effectively affiliate social media documents with planned events from various sources. Importantly, they demonstrated how social media documents from one social media site could be utilized to enhance document retrieval on another social media site, thus contributing to the variety of information that we can gather for planned events.

3.2 Text Mining

Authors of (Suttona , Gibsonb , Phillipsb , Spiro, Leage, & Johnson, 2015) explored the elements to predict the extent of retransmission for official hazard communications propagated through Twitter. Using data from events involving five different hazards, they found out three types of attributes—local network properties, message content, and message style—that jointly amplify and attenuate the retransmission of official communications under imminent threat. They found that the use of an agreed-upon hashtag and the number of users following an official account positively influence message retransmission, as does messages describing hazard influence or emphasizing cohesion among users.

Diakopoulos and Shamma (Diakopoulos & Shamma , Characterizing Debate Performance via Aggregated Twitter Sentiment, 2010) looked at the first U.S. presidential debate in 2008, in conjunction with aggregated ratings of message sentiment from Twitter. We begin to develop an analytical methodology and visual representations that could help a journalist or public affairs person better understand the temporal dynamics of sentiment in reaction to the debate video. They also showed that interesting events can be detected by looking at anomalies in the pulse of the sentiment signal and controversial topics can be identified by looking at correlated sentiment responses.

In (Sinha , Choudhury , & Agrawal , 2014) authors has a live coverage of an event and used it as a base criterion to check the textual feed about the same event. They used Twitter as the textual data source. In their study they try to find out the possibility of that correlation with using the aggregated sentiments from the set of tweets of the Roger Federer and Novak Nole semi finals match at Wimbledon 2012.

In (Diakopoulos, Naaman,, & Kivran-Swaine, Social Media Visual Analytics for Journalistic Inquiry, 2010) they presented a visual analytic tool, designed to help journalists and media professionals extract news value from large-scale aggregations of social media content around broadcast events. They discuss present the text analysis techniques used to enable the presentation, and provide details on the visual and interaction design. They provide an exploratory evaluation based on a user study in which journalists interacted with the system to explore and report on a dataset of over one hundred thousand twitter messages collected during the U.S. State of the Union presidential address in 2010.

The paper (Hea , Zha, & Li, 2013) also examines case studies which use text mining to analyze unstructured textual content from Facebook and Twitter sites of the three largest pizza chains: Pizza Hut, Domino's Pizza and Papa John's Pizza. They checked the stream of data and user engagements during time. The outcomes showed the value of social media competitive analysis and text mining as a practical technique to derive business value from the massive amount of social media data.

3.3 Topic analysis

Topic analysis is the study of deriving important and valuable highlights of the data stream, it can be done both by mining texts or using predefined algorithms or lexicons. In the Analysis section of the book (Morstatter , Kumar , & Liu , 2013), authors discussed the automatic discovery of topics in the text through "topic modeling" with latent Dirichlet allocation (LDA) along with the pre-processing pipeline of Lowercase-Tokenize-Stopword Removal-Stemming-Vectorization. Every topic in LDA is a collection of words. Each topic contains all of the words in the corpus with a probability of the word belonging to that topic. Figure 9 shows the modeling pipeline of this method.

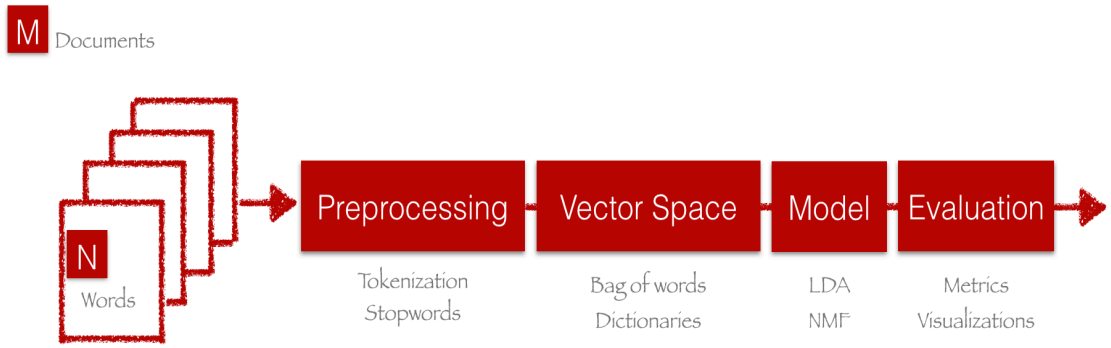

Figure 9: Topic Modeling Pipeline (chdoig.github.io/pygotham-topic-modeling)

In the paper (Choudhary , Agrawal , Patwary , Narayanan , Palsetia , & Lee , 2011) the authors addressed the problem of finding the top trend by modeling and classifying the topic in Twitter. They classified Twitter Topics into 18 categories and applied 2 approaches for topic classification; the Bag-of-Words method for text classification and network-based classification. In the text-based method, they construct word vectors to categorize the topics with a Naive Bayes Multinomial classifier. In network-based classification approach, they identified top 5 similar topics based on the number of mutual influential users.

3.4 Quantitative measurements

Most of the analysis in general is based on the quantitative approaches when dealing with social media, because the results are solid and in this way the risk of biased information and other effecting features are avoided. Shamma and Kennedy and Churchill (Churchill , Kennedy , & Shamma , 2009) discovered that the structure of Twitter traffic could provide insights into segmentation and entity detection. The authors examined Twitter volume over time and observed areas of high and low activity, spikes and pits which are clearly visible in the traffic volume, they also created the network graph of all the users and their tag relations from the Debate based on their quantitative measurements.

Authors in (Fonseca , Salvador , & Nogueira , 2015) proposed a framework to gather social network events and store their data in a relational database system for later analysis. A graphical user interface was developed to allow flexible access to stored information, according to the type of event, thus facilitating the analysis of users behaviors. With computing statistical models over the obtained data, it is possible to define ”normal or typical” usage profiles and detect possible deviations that can be indicative of a compromised user account.

3.5 User Engagement Analysis

Measuring user engagement is one of the most important methods used noticeably in marketing and many other fields. On of the main techniques is tracking influencers, which is one of the most important branches in analysis. Users interactions and their commitment to post are crucial in finding the transmission of data through network. Influential users are very important in the analysis phase as they play an important role in social media; therefore they are considered key players in many analysis studies ranging from marketing to political assessments.

Authors of “Twitter Data Analytics” book (Morstatter , Kumar , & Liu , 2013) focused on two key aspects of Twitter data for data analysis: networks and text. They employed the concepts of Degree, Eigenvector and Betweenness Centrality to answer questions such as “who is important in the network”, “Who is the most influential?” and “Who controls the flow of information?” These centralities are defined on the graph of network defined on vertices and edges based on users and their relations on Twitter.

The paper of (Bruns & Stieglitz , 2013) focused on analyzing engagements based on three metrics. They include user metric that is defining uses reactions to the hashtags as retweets or mentions and etc., Temporal metric that is checking active users and the amount of tweet sent per period of time, and the combined metric, which is determining currently active users along with the data being transmitted by them.

In paper (Cha, Haddadi , Benevenuto , & Gummadi , 2010) authors utilized a large set of data gathered from Twitter, and they presented an in-depth comparison of three metrics of influence: indegree, retweets, and mentions. Considering these measures, they probed the dynamics of user influence with respect to topics and time. They made many interesting observations. For instance, popular users with high indegree were not necessarily influential like causing retweets or mentions; most influential users have notable influence over many topics; influence is not obtained accidentally but with effort in limiting tweets to single topics.

As the studies concerning clustering and correlation, in this thesis we also used a dataset of a program that contains categories of live events from Twitter and Instagram. And we performed the general temporal analysis between the clusters and the two platforms to see how the data streams are correlated. Since the events are not being held at the same time, the periodical measures demonstrate the overlaps and synchronizations as well. The main program addresses many different subjects have diversities and similarities, so by gathering the information about the trends and highlights of the social media traffic, we can determine the tendency of the audience. Like the studies regarding text mining and topic modeling we also look into the topics that are most discussed and propagated in all categories. After applying the quantitative measurements on top expressions, we gather the mutual topics among the events, and the extent of their propagation in each one. Therefore by using a correlation coefficient we determined the extensity of their connection between events. This helps finding the cross relations between the clusters. Moreover this kind of programs attract people with various tastes, so to see how these people committed to their category of interest, we presented the groups based on the reactions they received from the users on social media.

Chapter 4

Social Monitoring for Live Events

This chapter discusses how to use social media analysis for tracking real life events. As the concept of social media analysis is a broad term, our study mainly focuses on the analysis level of social intelligence, with respect to the technologies and the types of analysis being discussed in the previous chapter.

4.1 Objective

The primary challenge of this study is to instruct a metric system that is both comprehensive and scalable to deal with a dataset of a program that contains live events and is extracted from different social media platforms. In this proposed scenario the dataset addresses a main program that includes subsections. The application of such programs varies from a system that analyzes the city information, artistic or cultural gatherings or any cross topic trends. Figure 10 shows the metamodel of such settings.

Different people from various backgrounds can organize live events; therefore they have different priorities and needs. They want to attract a wide range of various audiences from social media platforms and engage them in committing to their program. In order to meet the needs of such customers, analysis should consider many different features to extract proper online data. The events don't share the same time period, yet they may overlap. They each belong to a particular category, yet there are some cross-relations between them. The idea is to be able to recognize the data pattern of live events and extract information about their temporal stream of occurrence, their ability to attract users, the highlight of their content and their connections with each other.

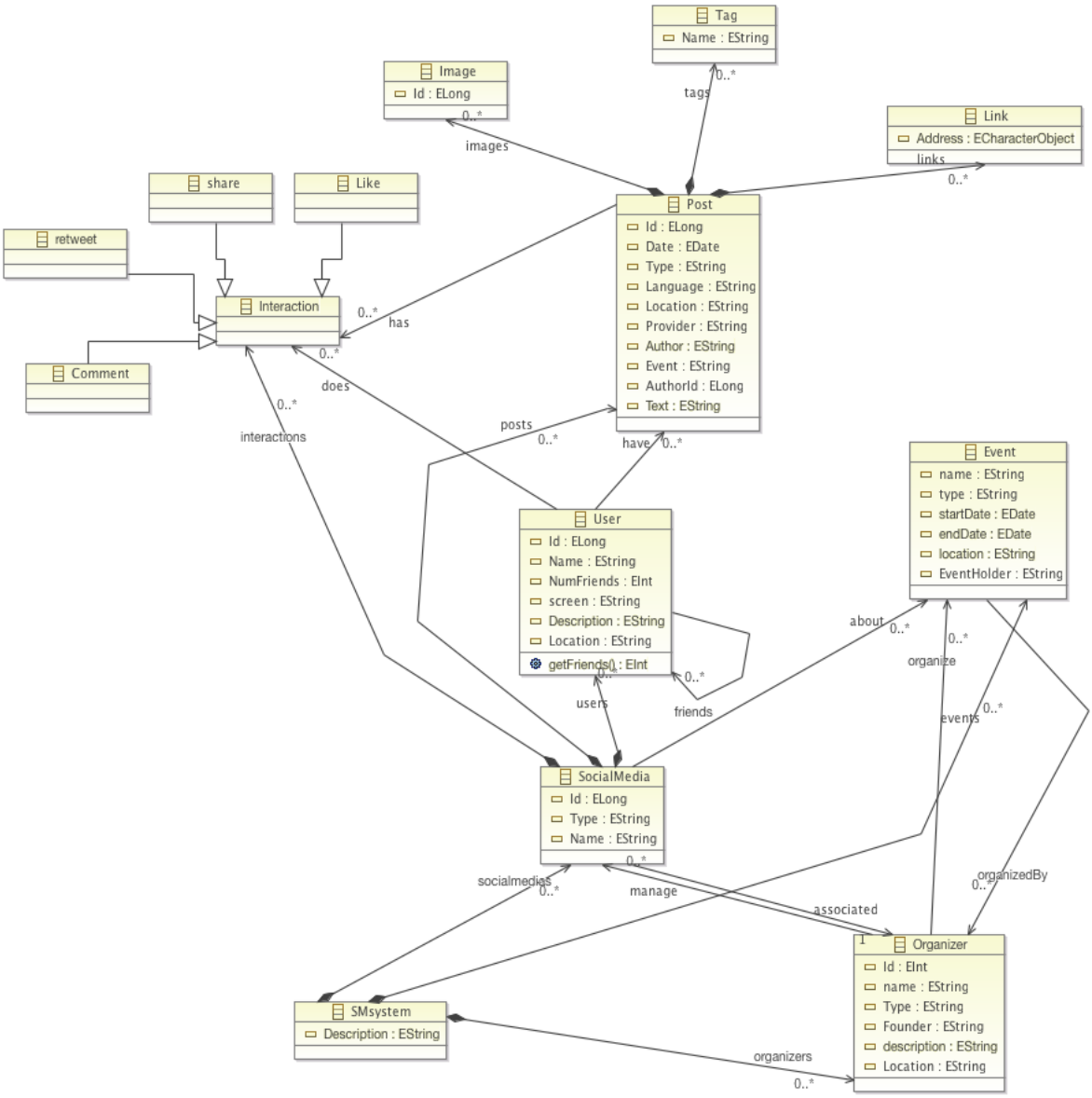

Figure 10: Model of our dataset

The model describes the potential setting of the program and displays the possible elements involved in such setting. There are important entities that have to be considered like Users, Events, Organizers, Posts and etc.

An example of how the entities can behave is presented in figure below.

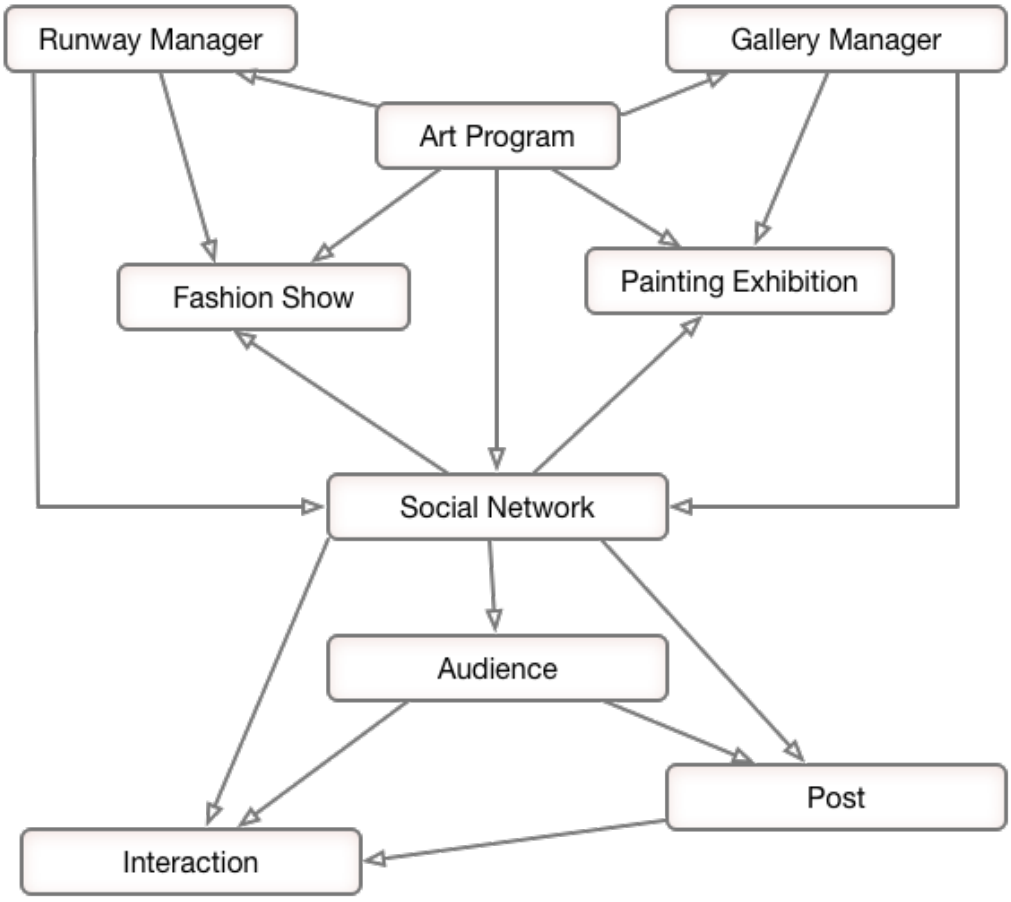

Figure 11: Instance of the system

Dealing with such a dataset from different trends and subjects can be quit confusing, both from the perspective of analysis and the organizers of the events. To better have an understanding on how to deal with this massive pile of data, the features and purpose of the study have to be clarified.

4.1.1 Organizer

There are several targets that organizers are aiming for during the program, but mostly the priority of an organizer is to increase the visibility of their program through out the process. This can be achieved by studying each element of the program and their relations and effectiveness on each other.

For each event the concerns are to improve the quality of that specific event and engage potential users, which are more likely to be interested in participating in them. They look forward to cutting the expenses by making sure that enough people are aware of the information they need before the opening and as the event is being held. For each event they need some insights and feedbacks on how the process is going forward to better improve it in real time, and as the events don't exactly take place at the same time or the same venue, then the importance of such information is much more significant to them to trace the associations between those.

An organizer tends to benefit from the idea of migrating from traditional advertisement to proper use of social media. In this way they have access to a much more updated source of information with less costs. By tracking the process of their project, the top trends and the voice of the audience through the lens of social networks they have the capability of making the right decisions.

Moreover, in case of a chain of events, the organizers of the entire program want to benefit from the word of mouth, meaning that by finding the most effective kind of topics, and most active people on social media, they can accelerate the procedure of spreading useful information about their activity. By this strategy they absorb a variety of audience, make the events more efficient and also diminish the costs of advertisement.

4.1.2 Analyst

When the desired features and milestones are set, the analysis stage will have to come up with some questions that better direct the flow in the process of analysis and result in more realistic outcomes, with respect to the limitations.

Basically in each analysis study, the main concerns are Time and Money. And all the decisions and methods that will be chosen should be in accordance with achieving best use of those two elements. In order to make best of the assets, one has to decide what are the minimum requirements in the project – what are the goals – whether there is a need for analytical tools or not – to what extend an analysis will help achieve that goal – is there a need to consider an analytic tool for the matter – is there another methodology (except spending on commercial tools) that can perform all that is needed, for instance content analysis, stream analysis, brand monitoring and etc. When dealing with a number of events that have to be analyzed both individually and as a whole, the need to have a proper answer to all the above questions are even of much more importance because there are new metrics that have to be defined.

As a result, there will be an indigence of acquiring information that is valid and relevant to study, along with availability of the data source at any time. In a real life event, data modification and flexibility of storing them are of those issues that need to be considered. These changes should reflect in the analysis stage and this makes it more challenging to set a default blueprint of the analysis procedure. The analysis process also requires the capability to be extended to a bigger scale, in order to accommodate the other elements of the study like other events. Also, in the presentation level, the visual functionality should be redefined, because the target is being analyzed in more than one dimension.

4.2 Approach

There are many analysis approaches to gain insight from social media data, and once the requirements have been defined, the analysis techniques can be chosen. The majority of types of analysis we discussed in the last chapter are methods that focus on one specific classification or measurement; the dataset is also based on one trend, which makes it less efficient for our scenario.

That is why, this thesis study suggests non-relational database for data storage, analysis techniques both in individual and categorized level for events, and use of programming libraries for the presentation stage. By storing the data from social media APIs into database, there is the opportunity to modify and insert new data along the period of the program. The dataset can be broken down to categories to better exploit them in the analysis stage.

The analysis stage is the one that is most notable to our study and it focuses on the cross related events. Specifications of this set of events are such as, having their own time period and the possibility of overlaps, diversifications of each plus particular associations among all. So here it requires a kind of analysis that helps in finding out how is the flow of event and it ‘s relevance to real life, also in a broader perspective, how those events are correlated and behave with respect to each other and to what extend their collaboration influences on the target audience.

To perform analysis on such a dataset, there are different methodologies that have been done before, yet in other domains with distinct specifications, that can be exploited and extended to meet the present goal of the project. As mentioned in the previous chapter, we also adopted some of the customary methods of analysis, and some other that are mainly applied to this version of information.

The types of analysis that is applicable in this setting can be categorized as:

- *Stream Analysis:* along with the application of this method to the whole dataset, it can be applied to each subsection of the program, and also between numbers of events having the same or partially the same time periods. This type of analysis also associates with anomaly detection, meaning that with the right visualization method, it can show the correlation to real life incidents as well. When the data is extracted from different media platforms, this analysis can be applied to help demonstrate the comparison of the flows.

- *Text Mining:* dealing with a micro-blogging network and textual formats of posts, makes it easier to dig into the tendency of the audience and their feedback towards the trends and issues. We can use such techniques to extract the most top expressions that users of social media use to express themselves about a subject. It is also expandable to another level that contains a series of related subjects to understand their similarities and contrasts.

- *Effective Influencers:* most effective influencers in each subsection of the program can be extracted by the same strategies that have been implemented so far. In addition the users, which are the common audience of more than one event, are considered to be of a great value in the sense that they bring more insights to the study because they can spread their ideas and also share the ones of others, which helps develop a flow of information among events. This being said, by defining the a set of common features (having the same location or interests, being a member of a specific trend or group, sharing the same links and etc.) between users and exploring through the ones that hold those features, the target users with those characteristics can be inferred.

- *Topic Analysis:* As much as this field of analysis can assist in finding out the top interests of the target crowd, by comparing the result of the highlights of several events, it will be possible to find the relevance between those topics. In addition the topics that are most common in the whole program can give a better insight to the organizer of what do the different users of different events have in common and the probable issues that they are focused on. The topics in common between two events help develop a connection stream to find out the similarities of the two and in a wider perspective, to manage them accordingly.

- *Correlation Analysis:* usually this type of analysis is the extension of quantitative measurements to detect anomalies, which depicts the divergence of the online trends to predefined patterns. In a broader view, this approach can be applied to the users and topics that share the same features to achieve the level of correlation between those data. This method is very promising to increase the level of accuracy with the pre-fetched data.

- *Quantitative Analysis*: based on the same measures defined for each singular event we can broaden it to include a comprehensive metric that is executed on the accumulation of data.

In this scenario the organizer employs a more flexible approach to data extraction and analysis, by extending the level of previous methodologies to multi-section procedure. In addition, certain downsides of the analytical tools are avoided, such as giving the permission to an external party to access the data, the extra cost of finding and applying tools and dealing with the limitations of commercial tools - being one dimensional.

Chapter 5

Implementation of the Approach

The research in the field of extracting data from social medias is very broad, and a complete review of all methodologies would be impossible. The input data has already been provided through open APIs. The focus here is specifically on storage, analysis and visualization of Twitter data. This chapter focuses on the specification of the system used to perform the analysis. The first section describes mongoDB methodologies to data storage and queries to retrieve information and a brief definition of each. The second section elaborates the procedure in which was used to filter information from Twitter data and its advantages over the other approaches of mongoDB. And at last there is a brief overview on the visualization part of analysis.

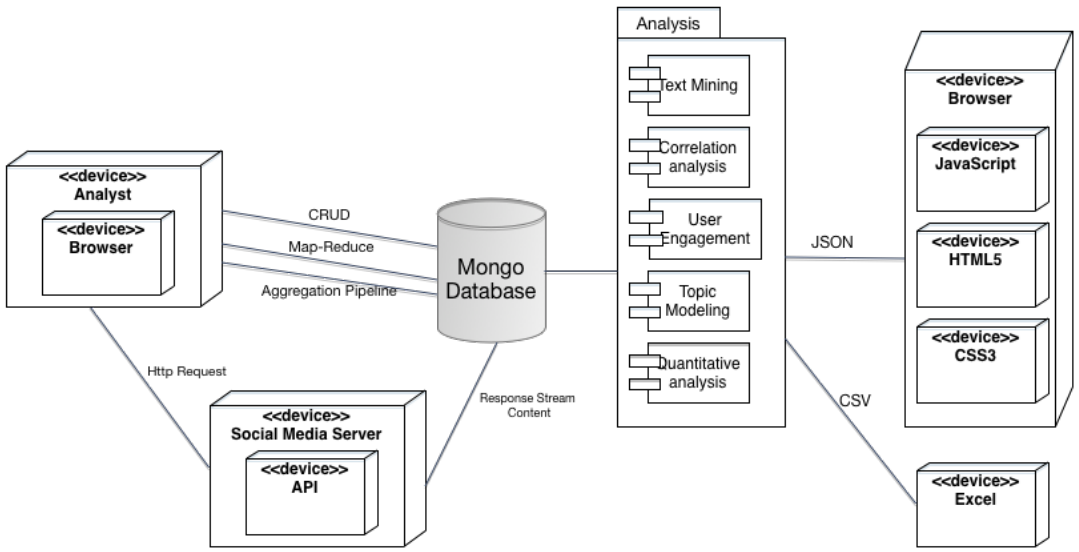

Figure 12: Deployment Diagram of the Process

5.1 MongoDB Approaches

MongoDB is a cross-platform, document-oriented database that provides high performance, simple accessibility, and easy scalability. MongoDB works on the concept of collection and document. Database is a physical container for collections. Each database gets its own set of files on the file system. A single MongoDB server typically has multiple databases. Collection is a group of MongoDB documents. It is the equivalent of an RDBMS table. A collection exists within a single database. Collections do not enforce a schema. Documents within a collection can have different fields. Typically, all documents in a collection have similar or related purpose. A document is a set of key-value pairs. Documents have dynamic schema. Dynamic schema means that documents in the same collection do not need to have the same set of fields or structure, and common fields in a collection's documents may hold different types of data.

MongoDB offers basic semantics for reading and modifying data. CRUD operations are the foundation of all interactions in database and like in figure 13; it is used for creating, reading, updating and deleting data.

Figure 13: MongoDB CRUD operations

For basic query operations, MongoDB CRUD operations provide a db.collection.find() method. The method accepts both the query criteria and projections and returns a cursor to the matching documents. You can optionally modify the query to impose limits, skips, and sort orders. However this process doesn't have the same functionality, as the other methodologies in navigation flexibility or speed, like map-reduce or aggregation framework.

MongoDB also provides a data processing paradigm called Map-Reduce, which is generally used to process large and complex data sets. The map-reduce function first queries the collection, then maps the result documents to emit key-value pairs that is then reduced based on the keys that have multiple values. The Map-Reduce operation uses a temporary collection during processing so it can be run periodically over the same target collection without affecting intermediate states. This mode is useful when generating statistical output collections on a regular basis.

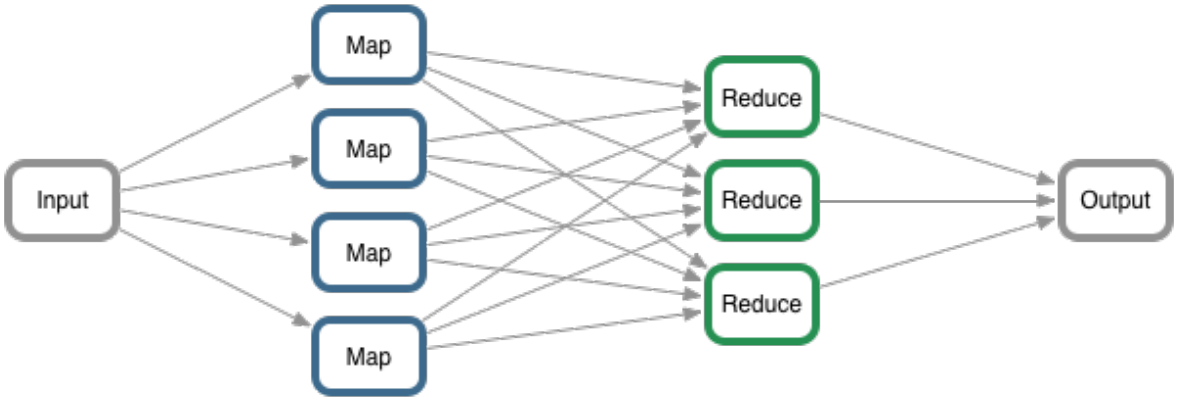

Figure 14: Map-Reduce structure

Although map-reduce is a very powerful tool but it has a considerable overkill for simple aggregation tasks like summation, calculating averages, grouping or reshaping. It is hard to program and debug with high level of complexity. Therefore in our study we applied the aggregation pipeline that is fairly discussed in the next part.

5.2 Aggregation Pipeline

MongoDB also offers another method called Aggregation Framework. It is based on the concept of processing data through a pipeline. The pipeline receives documents as input data and transforms them through different stages into an aggregated result.

Pipeline stages take expressions as operands; each defines the transformation being made on the documents. The expressions only perform modifications on the current document at any specific stage. The result of each stage is then considered an input for the proceeding one.

Almost all the functionalities in relational databases have been implemented in MongoDB except JOIN – which the concept is somehow covered in the mongo new version 3.2 as a left join between two collections - plus many more facilities to better store and extract data in documents. In the figure below, there is a simple example of how pipeline works.

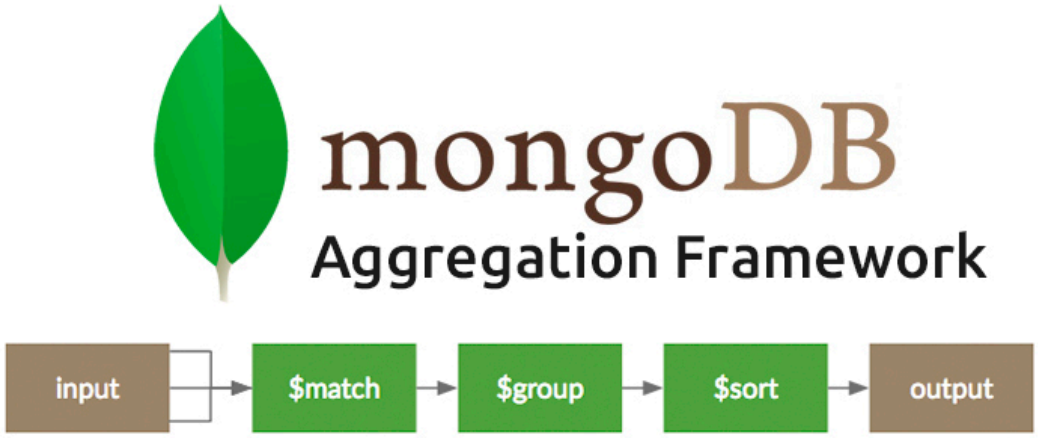

Figure 15: Mongo DB aggregation framework, credits to docs.mongodb.org

In our analysis we used the aggregation framework, because of its robustness and speed. Also because when there are different trends to follow, there are many constraints to take into consideration and aggregation pipeline was the one with the most powerful grouping approach and flexibility in using indexes.
In figure 16 there is a sample query that extracts the tags that are common between two events from the dataset of Twitter, along with the number of appearances in both events.

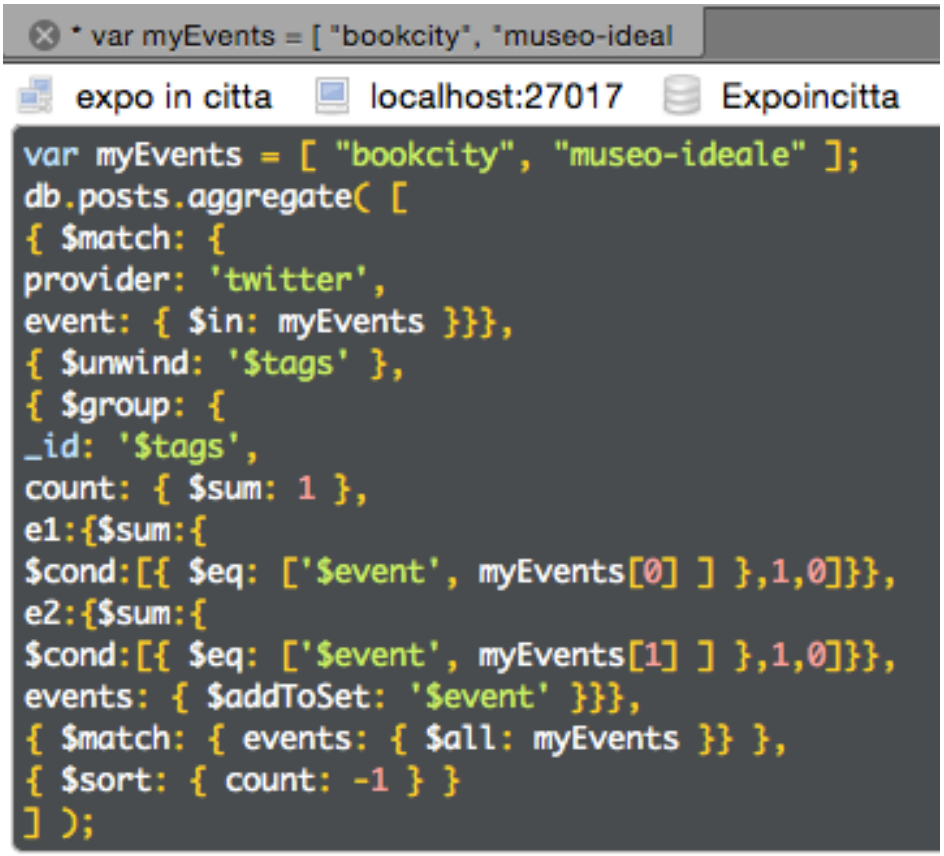

Figure 16: Aggregation Pipeline Example

And the result is in the format of JSON :

"result" : [
{ "_id" : "Milano",
"count" : 597.0000000000000000,
"e1" : 358.0000000000000000,
"e2" : 239.0000000000000000,
"events" : ["museo-ideale", "bookcity"] }, …

5.3 Data Correlation

Among those measures that are defined to check the correspondence of a series of data to a predefined criterion, we chose Pearson Function. This function is a subset of continuous probability distributions and shows the extent of linear relationship between two sets of data. The syntax of the function is to receive two arrays of data as input streams, which are either numbers or references that contain numbers. And it results in Pearson product correlation coefficient that outputs an index ranging from -1 to 1.

The formula of the Pearson product is:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

In which the X and Y are the means of the datasets (array1 and array2).

Basically it can be implemented in an excel worksheet. For instance, by extraction of common features from two events and storing them as an array of numbers we have the initial data that shows the quantitative relevance between those events. Then in order to see how these features correlate, we can also obtain the number of their occurrence in each of the sources, by mining the relevant data stored in the collection, like the example in figure 16 that shows the number of occurrences for common elements both in the general accumulative context and individually in each specific event.

Where the index moves toward “1”, it is safe to say that the measurements of elements from the defined events are following the same orientation and the two arrays are correlated. Whereas indexes close to “-1” shows less connectivity between the elements.

5.4 Data Visualization

For the data visualization stage we benefited from JavaScript libraries. From the many libraries that are being used recently we chose D3 (Data Driven Documents). It is an open source JavaScript library that helps in binding random data to DOM (Document Object Model) and implements data driven transformations. The advantages of this library are minimum overhead, speed, flexibility, large data support and dynamic manipulations.
It is a tool for loading data into the browser and generating DOM elements based them. This library among all is most famous for generating SVG graphics, that is a vector image format supported by web browsers.

After being done with the analysis level, the results are mainly in JSON. This is why using the scripting language are the optimal choice in this field. Simply by binding the JSON, D3 library and HTML elements, the result is generated in the browser and can be modified easily. D3 supports many visualization models as in figure 17.

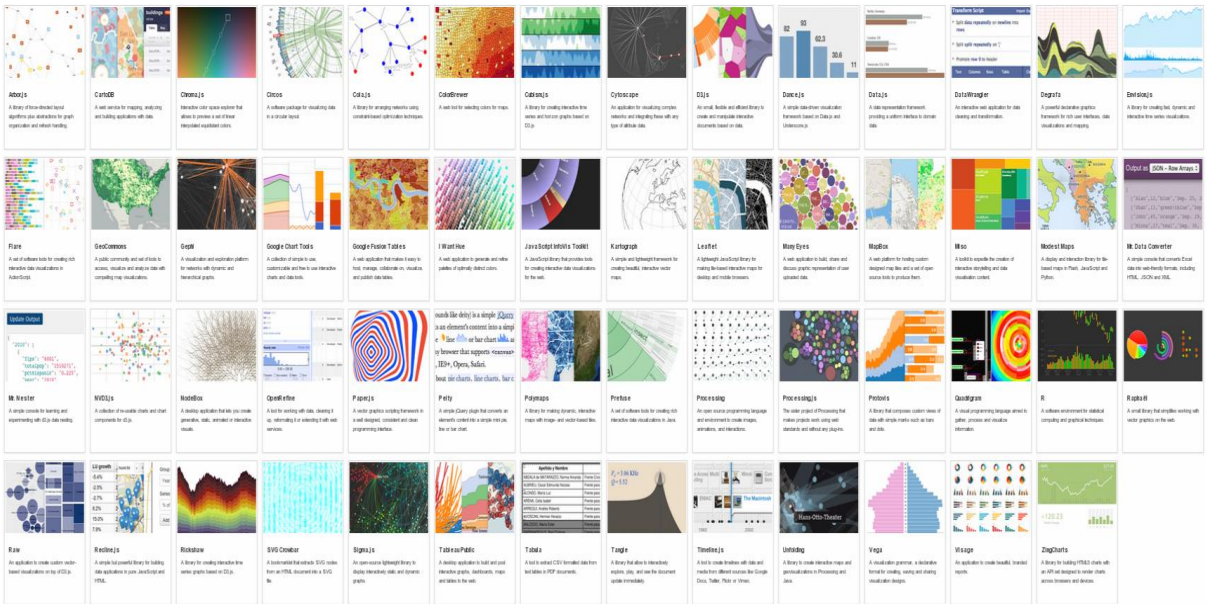

Figure 17: Visualization models of D3.js

Chapter 6

Experiments and Discussion

This Chapter is dedicated to elaborate what we experimented through the study, to present the following outcomes and to discuss the perceivable ideas from the results. In the first section the dataset that was used to apply the analysis is explained. The second section demonstrates the results of the analysis stage with the help of visual stage as was described in the previous chapter, and finally the significant findings and issues that are brought up during the process.

6.1 ExpoInCitta Dataset

Along with developing Information and Communication Technology (ICT), Smart City projects have become of great importance. These projects give an insight on the urban level to better find practical solutions for managing city assets and improve quality and performance of services. As a subset of such projects, our analysis is focused on the dataset from a program called "EXPOINCITTA" that took place in Milan 2015 as the same time as the EXPO. The program includes 13 diverse events, which have some common features and differences as well. The time period of their occurrence varies, and in some cases overlap. Each event belongs to a category defined as in table 3 with their location and time period.
The categories are:

1. *Art*, with reference to exhibitions specially art exhibition like museums fairs.
2. *Performance*, with reference to concerts, live shows and fashion events.
3. *Science*, with reference to scientific gatherings or seminars.
4. *Media*, with reference to events with informational or cultural nature.

Category	Event	Date	Venue
Performance	Vogue Fashion Night Out	22 September	Various
	Piano City Milano	22-24 May	Various
	Estathè Market Sound	8 July	Mercati Generali
	Radio Italia Live	28 May	Piazza Duomo
Art	Museo Ideale	15 May – 13 September	Museo del Novecento
	Aqua Shock	3 September – 1 November	Palazzo della Ragione Fotografia
	Juan Muñoz "Double Bind & Around"	9 April – 30 August	Hangar Bicocca
	Leonardo da Vinci	16 April – 19 July	Palazzo Reale
Media	La mia Basilicata	26 October	Teatro del Verme
	Milano Film Festival	10-20 September	Various
	Bookcity	22-25 October	Various
Science	Wired Next Fest	21-24 May	Giardini Indro Montanelli
	Spinosaurus	6 June – 10 January 2016	Palazzo Dugnani

Table 3: Expoincitta Dataset

The data was collected through querying Twitter and Instagram APIs; the units of analysis are the tweets from Twitter and the photos from Instagram. However most parts of the analysis metrics are focused on those extracted from Twitter APIs and we use data from Instagram for the sake of comparison.

As discussed in the previous chapter the data was imported into the Mongo database in a JSON input file. The relevant tweets and photos have been downloaded by keywords. Generally the total number of 225351 records was stored for the year 2015 that include 94112 tweets from Twitter and 131239 posts from instagram.

The figure 18 is a sample of a tweet and it's attributes that are used in the analysis.

Key	Value	Type
(1) ObjectId("563202384687949f78e57b05")	{ 12 fields }	Object
_id	ObjectId("563202384687949f78e57b05")	ObjectId
provider	twitter	String
id	639435764942667776	String
event	double-bind-around	String
text	"I sette palazzi celesti" #anselmkiefer #hangarbic...	String
date	2015-09-03 13:52:21.000Z	Date
location	{ 2 fields }	Object
author	gabrielesiani	String
authorId	234014213	String
tags	Array [3]	Array
lang	it	String
raw	{ 25 fields }	Object

Figure 18: Expoincitta sample data in MongoDB

The elements of the tweet have their own data type, which are sometimes "Null" except for the ID keys that are essential to identify the tweet and user. The "_id" attribute is the unique key that the tweet can be recognized with. And we can access the user's information with the key "Author" that is the real name of the user, along with the his unique "AuthorId".

Other attributes of the data stored in each record are:

- "Provider", which in this case is either Twitter or Instagram.
- "Event" specifies the name of the corresponding event.
- "Date" specifies the timestamp of the tweet.
- "Text" is the whole context of the tweet as the way it was posted.
- User can activate "Location" while tweeting and shows his longitude and latitude, usually with a predefined name.
- "Tags" is an array of Hashtags mentioned in tweet.
- "Lang" defines the language of the post.
- "Raw" is an object that contains all information about the tweet and the attributes in details, and is shown in figure 19. By employing the information inside this object, possible connections between twitter elements can be inferred.

Key	Value	Type
raw	{ 25 fields }	Object
created_at	Thu Sep 03 12:37:45 +0000 2015	String
id	639416992823398400.000000	Double
id_str	639416992823398400	String
text	CASINO #damianortega #hangarbicocca @ Hang…	String
source	<a href="http://instagram.com" rel="nofollow">In…	String
truncated	false	Boolean
in_reply_to_status_id	null	Null
in_reply_to_status_id_str	null	Null
in_reply_to_user_id	null	Null
in_reply_to_user_id_str	null	Null
in_reply_to_screen_name	null	Null
user	{ 41 fields }	Object
geo	{ 2 fields }	Object
coordinates	{ 2 fields }	Object
place	{ 10 fields }	Object
contributors	null	Null
is_quote_status	false	Boolean
retweet_count	0	Int32
favorite_count	1	Int32
entities	{ 4 fields }	Object
favorited	false	Boolean
retweeted	false	Boolean
possibly_sensitive	false	Boolean
possibly_sensitive_appealable	false	Boolean
lang	it	String

Figure 19: RAW attribute of a sample tweet in MongoDB

6.2 Reports of Analysis

This section presents the result of the social media analytics process that was applied to ExpoinCitta dataset on MongoDB database. It is divided into sub sections that present the results in different stages of the analysis. Each one corresponds to the outcomes from various methodologies that were applied and discussed in the previous chapters.

6.2.1 General Results

As the program took place more or less at the same time of EXPO, it is safe to say that it was to some extend under the influence of the visiting tourists and gatherings in city of Milan. This makes Expoincitta to be a program of diverse tastes and styles, where many people come and share their views on diverse matters. This quality of such international programs makes any observer to wonder about how the events were held and how they were received by audience. What were the main topics that attracted most people and many more questions.

Obviously there are some parts of the program that are more appealing to the audience and attract much more media attention. So as the first step of the analysis and to get a better impression of the classification and the variety of sub sections we refer to figure 20 that shows the amount of reflection that Expoincitta had on Twitter and Instagram. It depicts the categories and sub events due to their diversity is size, which is calculated by the number of posts regarding each event on social medias in 2015. This will give us a general insight on how much attention each event and each category received during the program.

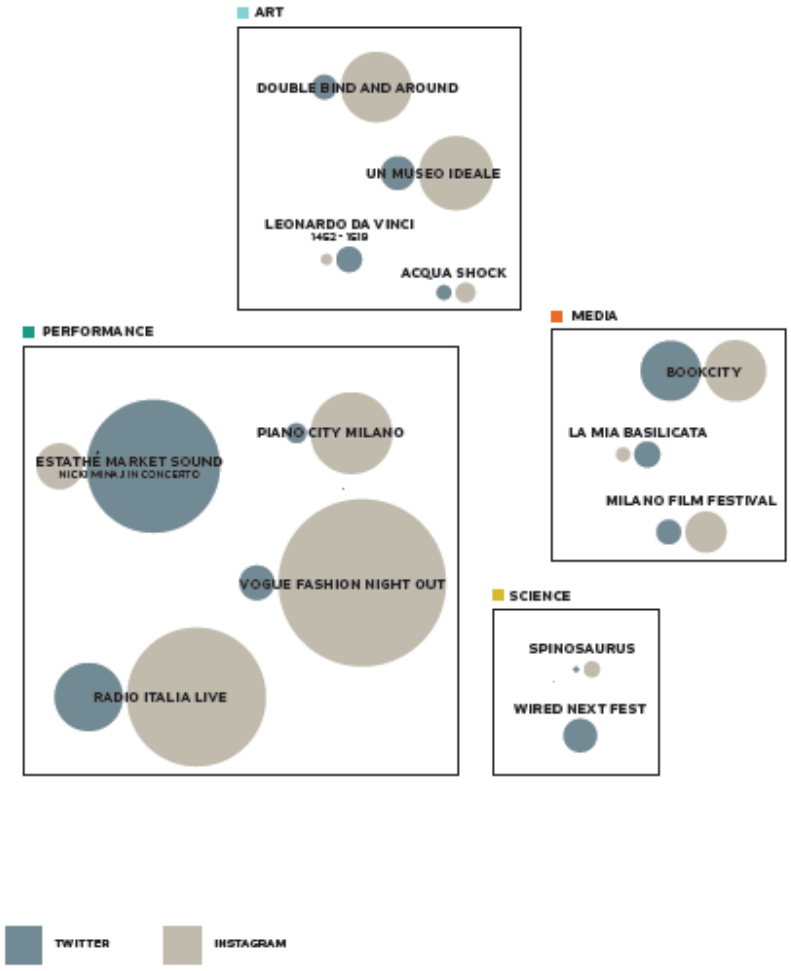

Figure 20: Diversification of Events on Social Media by Category

Figure 20 represents the volume of tweets and photos for each event, represented by a circle whose area is directly relevant to the sum of the tweets and photos. It shows the diversity both between the types of events, and also within the same classification. The category with greater resonance on social media is Performance; instead the less reflected category on social media was the scientific one. In the Performance category events stimulate more sharing of images like in the case of Vogue Fashion Night Out and Radio Italia Live. Whereas events like Estathe Market Sound motivated sharing in text message formats. Interestingly, cultural events, that might seem more distant from the social media world, as Double Blind and Around (Art) and Book City (Media), also received a good resonance on social networks.

The Expoincitta program lasted 9 months from April until November. Instagram and Twitter were the lenses that we used to do the analysis through them. To filter the data, the keywords were used to extract the relevant posts about a specific event in the period of their occurrence. As a result the dataset of Twitter is focused from the start to end point of each event, whereas the dataset from Instagram includes a continuous stream during the whole program.

The peak point in the volume of posts from Twitter dataset happened in May, July and September and the peaks for Instagram took place only in September. In July, people attending Expoincitta program discussed it the most on Twitter with the maximum number of 52342. However, in September the resonance of Instagram was the highest with the maximum number of 36165 posts.

As we saw before, events received different amount of attention in different occasions, and here as in figure 21, it can be also inferred that such attention is received in different intervals.

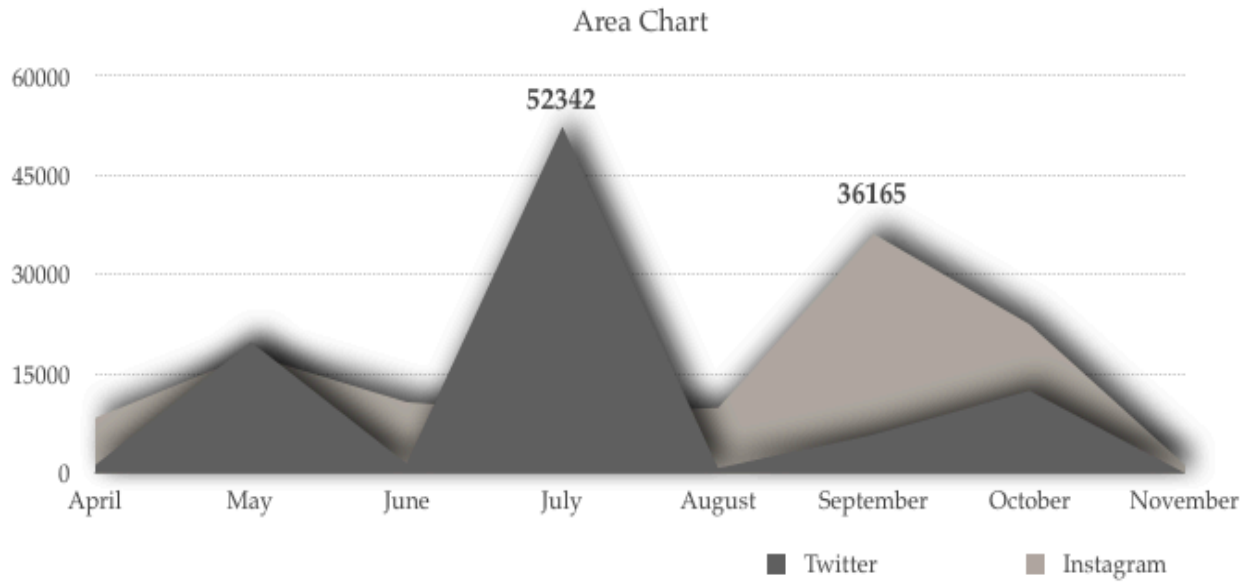

Figure 21: 2015 Expoincitta Data Stream on Instagram vs. Twitter

Figure 22 follows the temporal stream of Expoincitta during 2015. For each month of the program it demonstrates the volume of posts on Twitter and Instagram and relevant proportions for all events.

There are events that happen during a period of time and it is noticeable from the figure below, yet there are events that are limited to a specific date. "Estathe Market Sound" is the one whom the audience posted the most about, and it took place in July, which made the overall measures to increase in that month. On the other hand the dispersal of events like "Museo Ideale" or "Double Bind Around" are marked through time, and "Radio Italia Live" is the one that is reflected on both networks more or less equally on May. This presentation of events renders a further insight in case of temporal comparison with respect to the two platforms.

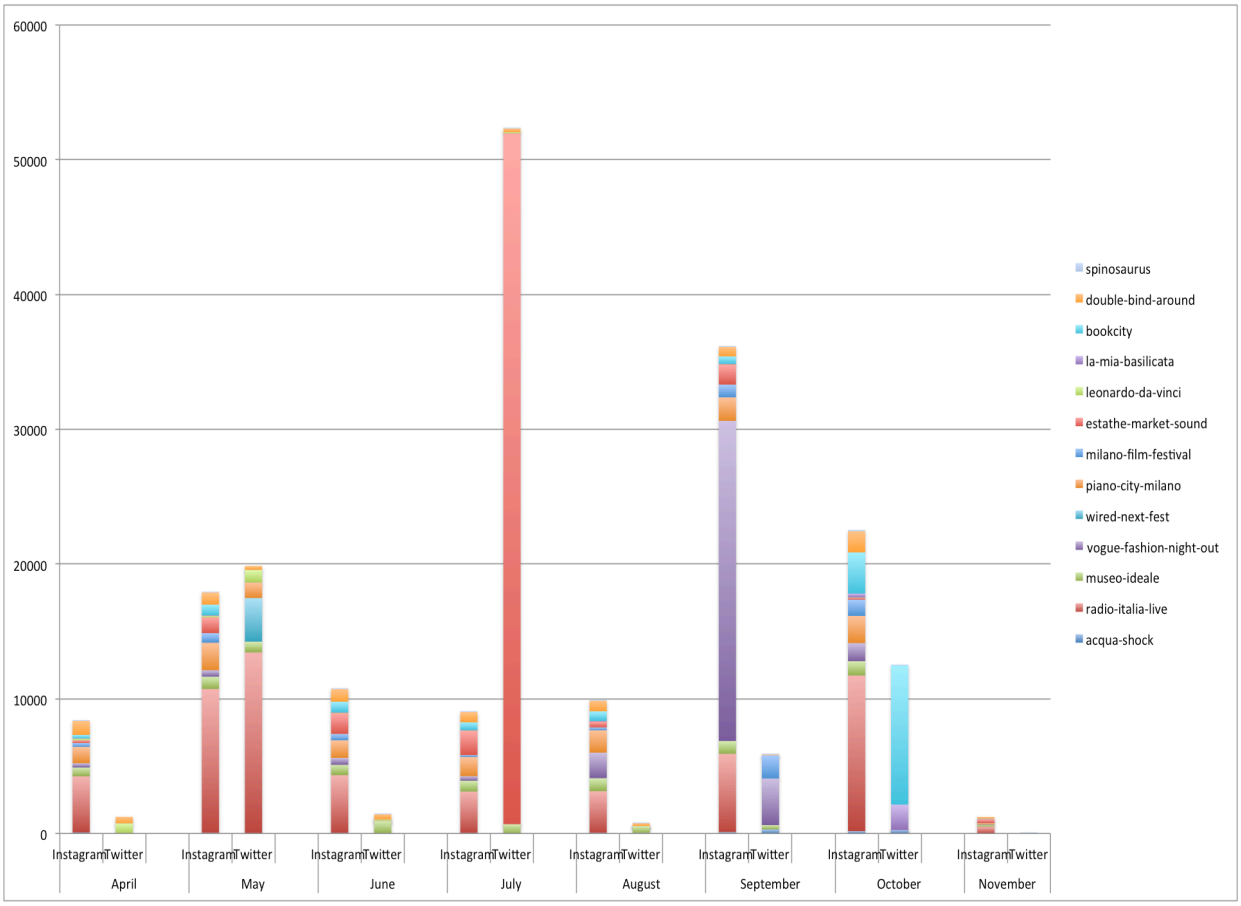

Figure 22: Event Reflection in each month on social media

At the basic level, the information from the events is retrieved quantitatively and in relevance to their timestamps. We applied the analysis on all 13 events involved in this study but mentioning all of them would be cumbersome and not to the point. Among all, we chose to track "Milano Film Festival".

This event was held for 10 days, from10th to 20th September. People posted about it both on Twitter and Instagram. Figure 23 shows the number of posts for each day of the event. Clearly for this film festival, the number of tweets is higher than the number of posts through out the 10 days. Also, for those who posted on Twitter about the event, there seems to be something that attracts their attention on 15th and 18th, as on Instagram, more photos are shared on 12th and 14th.

Usually the peaks and well of a graph conveys a phenomenon that has happened in real life. For this event, there was a meeting on 18th with Christian Braad Thomsen, a movie director. And this might be considered the cause of the distinct increase in posts, since by checking the content of tweets published on twitter on that day; most of the posts include hashtags of the director's name.

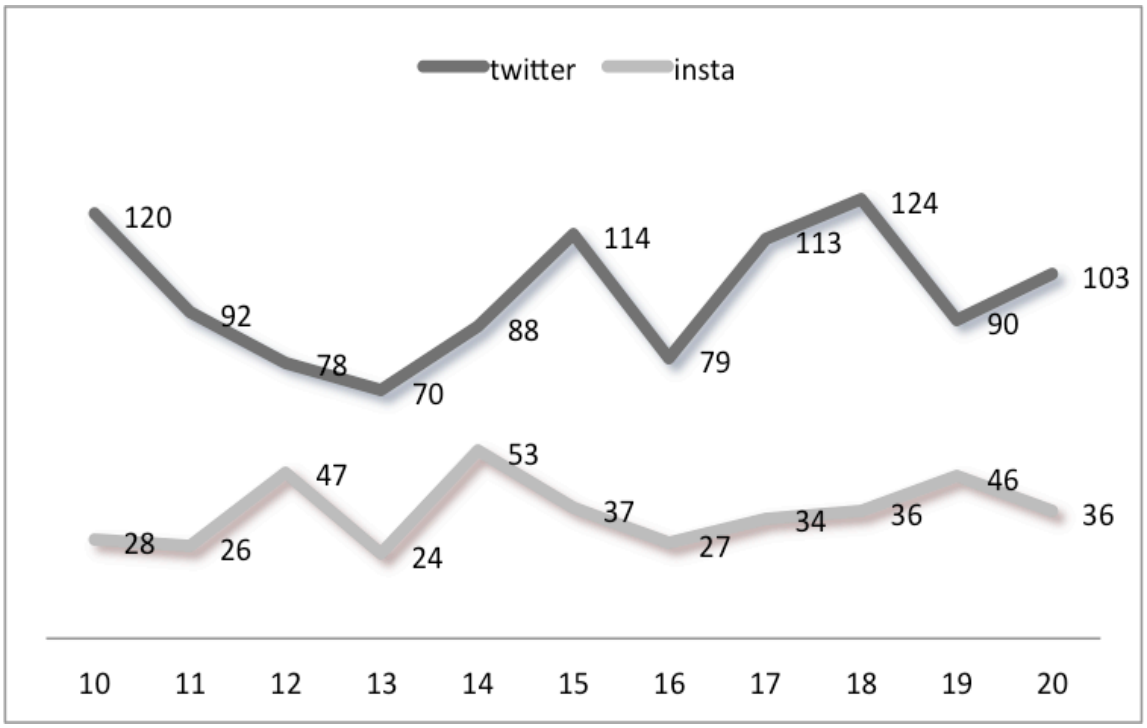

Figure 23: Milano Film Festival in September 2015 on social media

6.2.2 User Specific Results

Studying Users is one the most important branches in social media studies and this thesis as well. At the very basic level, the quantitative measures can show the extent of user attractions of any subject. This can lead the rest of analysis. In this section we have an overview on how much each event and their categories were successful in absorbing their own audience on Twitter.

The number of Twitter users who posted about each event during the program of Expoincitta is given in the table 4 at the end of this chapter. The most user-attracting event would be "Estathe Market Sound" who was mentioned in the tweets of 25429 users. This event belongs to the category of Performance that holds the most number of tweets as well. Evidently from Figure 24, events that belonged to the Performance category attracted 82 percent of the audience of Expoincitta program on Twitter. Whereas some cultural events like "Vogue Fashion Night out" that had a similar volume of posts per event as the others average volumes, had a noticeable range of users.

The least number of users belonged to the category of science. Also the event that received much less attention from users is "Spinosaurus" which is also considered the event that was least talked about. In the category of Media, most people tweeted for the "Bookcity" event, as in Art category; the most talked about event was "Double Bind Around". So far, the most appealing group of events, on social media platform based on number of tweets (69332) and number of users involved (33167), is the Performance category.

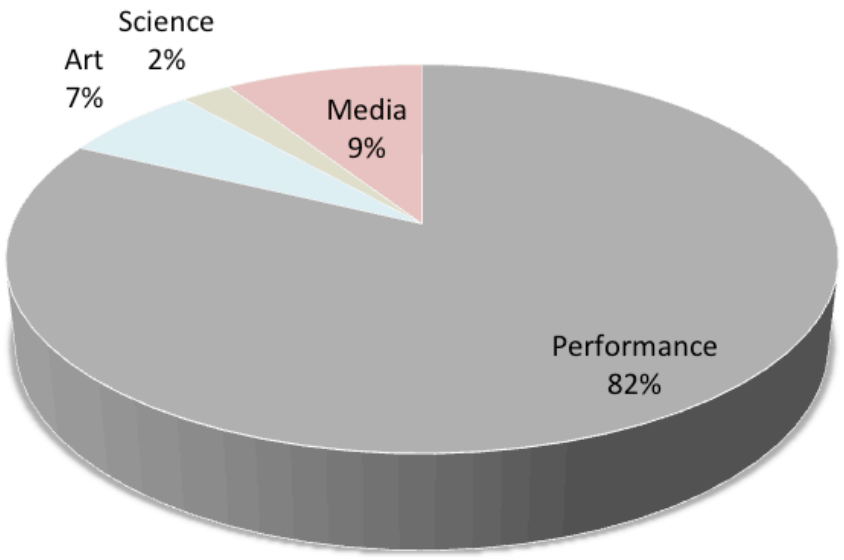

Figure 24: Dispersion of Twitter Users for categories of Expoincitta events

To control the propagation of data through Twitter, we considered two features that enable users to interact and share ideas on this network. As discussed before there are many ways on social media to estimate the feedbacks of users and control their commitment to the subject. Here, we focused on two basic behaviors defined on Twitter, which is favoring a tweet or sharing it as a retweet.

By measuring the number of favorites given for tweets regarding each event in each category, we came up with the figure that shows the popularity of events and also their relevant categories, as shown in figure 26. The most popular event was "Estathe Market Sound" which was also the one that people talked about the most, also the same for Performance category. The categories of media and art, which were different by 2 percent in population, almost got the same amount of likes from users. The figure highlights a particular twist, where events with a notable difference in volume of posts receive the amount of attention that are not expected and are not corresponding to the previous measure.
For instance "Museo Ideale" with 746 users, received almost the same attention in number of likes, as "Radio Italia Live" with 5039 users and "Vogue fashion Night Out" with 2206.

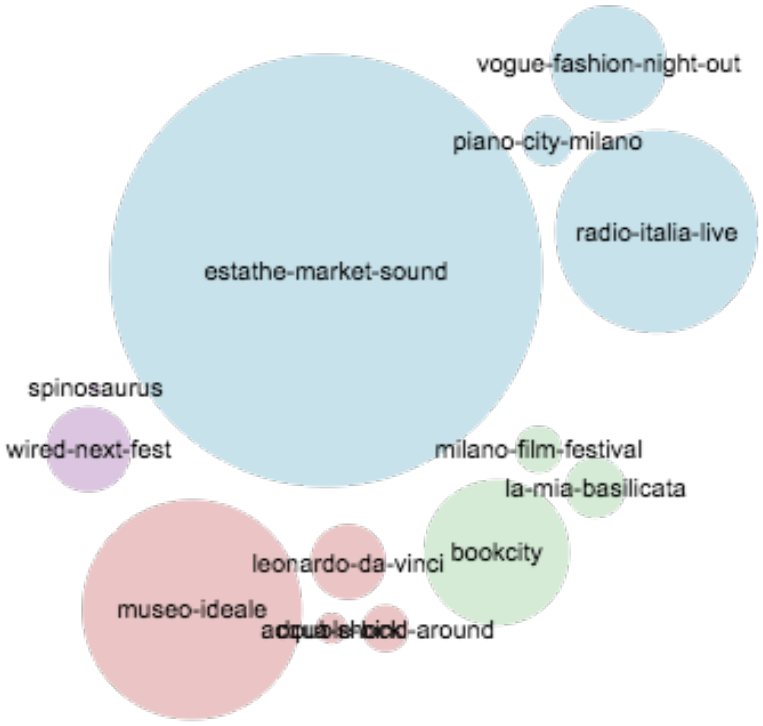

Figure 25: Retweet propagation of events on Twitter

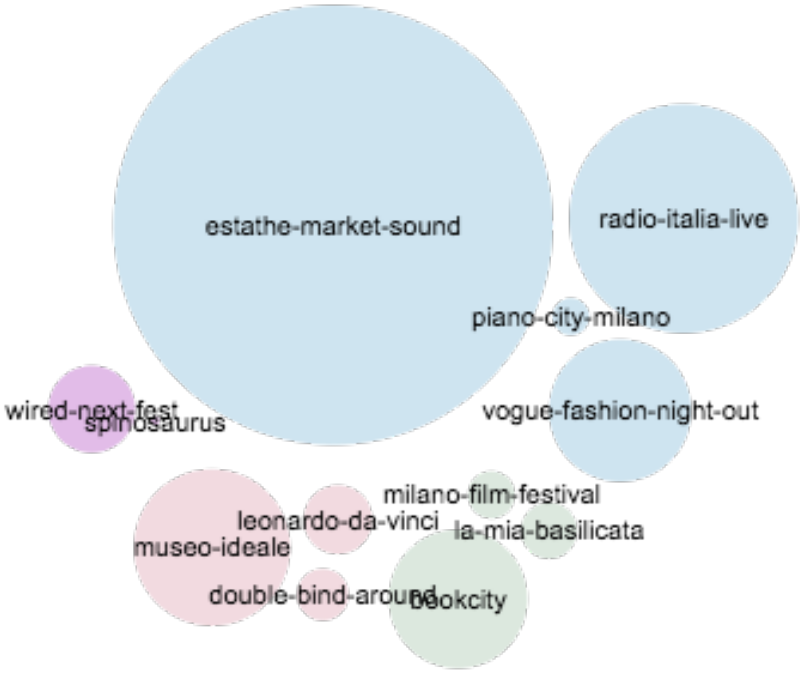

Figure 26: Popularity of tweets of different events

Another measure that was applied for assessing interaction was the number of posts that was shared through the network of Twitter; they were relevant to each event and their category. Figure 26 shows the scale of events with respect to the volume of retweets per event. The total number of retweets for Performance category was 189590, 44245 for Art, and 5633 for Science and 20832 for Media.

As expected from the volume of users involved in Performance category, there propagation of tweets is higher for this group of events and especially for "Estathe Market Sound" in figure 26, and the least-shared tweets belong to the scientific events of the program and the "Spinosaurus" event in particular. There are some events from different categories that are almost equally shared on Twitter, for example "Museo Ideale" from category of Art, "Bookcity" from Media and "Vogue fashion Night Out" from Performance. The interesting point from the figure 26 is that "Museo Ideale" which contains half of the audience of the other two events, is partially greater in retweet counts. And comparing the previous figure of popularity, "Museo Ideale" is less appealing to the audience when it comes to share its posts. On the other hand events that users were partially willing to like on Twitter like "Aqua Shock" turned to be not so much of importance in measuring it's tweets propagation. Other kept their volume in both measures that means that they were almost equally liked and transmitted through Twitter.

In comparing the categories of events, evidently the posts about events belonging to Media category are less shared than the ones from Art cluster. This is however in contrast with the population of these clusters. Media is by 2 percent higher than Art in the volume of posts, yet in the popularity measures the overall attention of Art events is again more than Media.

These comparisons were between the recent measurement and the volume of users involved in each event. Events like "Museo Ideale" and "Vogue Fashion Night Out" are both equal in post and retweet volume. However, by comparing to the size of the event, it is obvious that events like "Bookcity" and "Radio Italia Live" are different in propagation measure whereas they both have the same amount of posts published about them.

6.2.3 Content Specific Results

Delving into the content of social media is one of the metrics that gives us a better understanding of opinions of the crowd. In this section we dealt with the Hashtags that were used to reference the events. Overall in the Performance category, people used 8994 Hashtags and 4152 in Media, 3062 in Art and 1012 in science. The number of hashtags used in the Performance group is higher but that doesn't necessarily correlate with the relevance of the topics or contents used, as people use these expression to express whatever they think of something.

So to explore more into the specifics of the contents in our program, we derived the top expression that were used and their frequency during the events. Figure 27 shows the top ten tags of each cluster of the program, where the volume of the nodes corresponds to their frequency. Some hashtags directly represent the events they were used for, like "bookcity", "wfn", "museoIdeale" and etc.

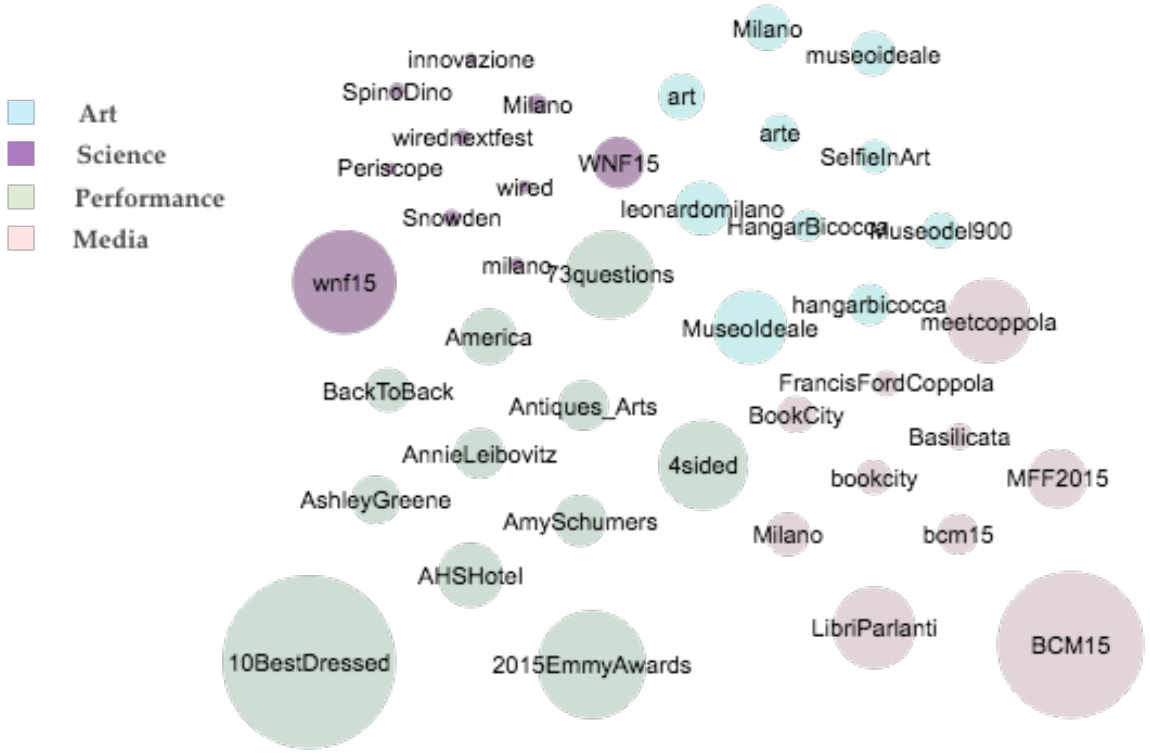

Figure 27: Popular Topics from events in Expoincitta

6.2.4 Common Grounds

One of the most important aspects of our study is to address the similarities and difference between the groups of events that were present in the Expoincitta program. And to fulfill such endeavor we performed the analysis based on their relative clustering. Yet there are many more cross relations between these elements that are revealed only when we pay attention to the correlation between them.
Figure 28 shows the extent of relations between events of different categories based on the common users that posted in the two specific events. For example people who posted about "Bookcity" also posted in many other different events. But the audience of " Spinosaurus" is barely connected to the other nodes.

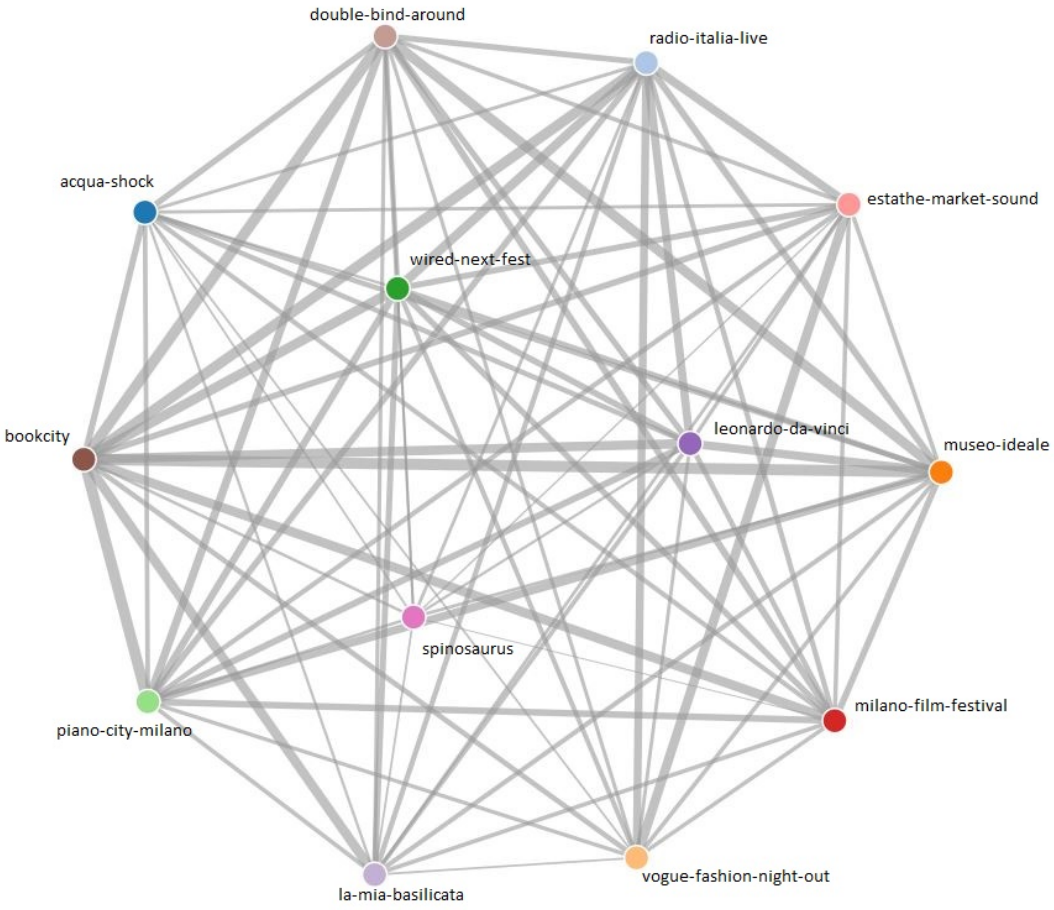

Figure 28: Users Correlation between Events

The other significant measure that gives us a great insight on how the events are content related, is the correlation between the contents that were used in each two events. Figure 29 shows the relevance between events of the program based on the number of their mutual hashtags. This represents a network, where events are depicted as nodes and edges shows the relation between two nodes based on the number of similar expressions. Clearly the events of the same category are more content-related; also the nodes of Performance and Art have the strongest connection that means the majority of expressions that were used in the two groups are mutual.

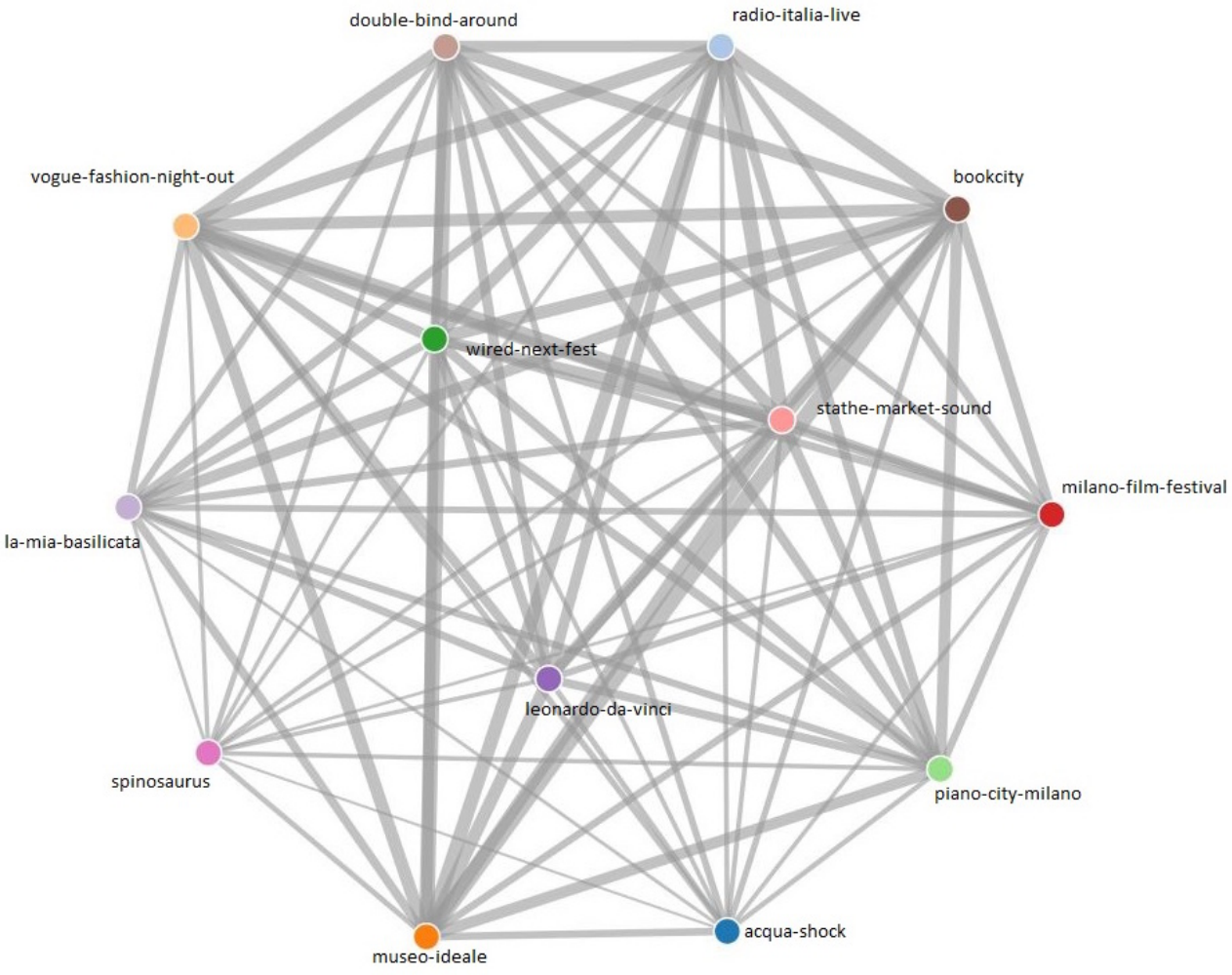

Figure 29: Tag Correlation between Events

To better understand the concept of relations between events based on their content, we took the previous measure a step forward where we check the mentions of those common tags among each two event with applying the Pearson coefficient to check their correlations. The events that belong to the same category show strong connection based on their similarity in shared expressions so we applied this method to events from different categories to gain a different perspective.

For instance, we chose the "Radio Italia Live" from Performance category and "Museo Ideale" from Art; these two have 105 tags in common which shows a solid connection. Then we checked if those common tags had the same volume of mentions by applying the Pearson coefficient and the result was -0.006 that does not make sense. So we defined a limit to avoid considering the irrelevant tags and applying the Pearson coefficient to the rest. The limit eliminates the tags, which their difference in propagation is more than 100, for example if one expression is published 2 times in one event and 200 times in another one. After considering the limit, we used Pearson correlation analysis over 101 mutual hashtags. And the result is 0.85, which as expected represents the correlation of the contents of the two events.
"Milano Film Festival" and "Vogue Fashion Night Out" are from Media and Performance categories respectively and share 41 common tags. After limiting the data, we applied the Pearson correlation analysis on 40 hashtags and the result was 0.61.
To see how the Pearson coefficient reflects the correlation when there are few similar tags, we chose "Wired Next Fest" from Science category and "Aqua Shock" from Art, which have 12 expressions in common and still 12 after applying the limit, and the Pearson coefficient was 0.55. This number reveals that although these two events have a few mutual hashtags but the dispersion of them are moderately equal in each event.
In another try to compare Media and Art category we chose "Museo Ideale" and "Bookcity" which are the two popular events of their groups and share 149 tags, which then decreases to 143 that bring the Pearson coefficient to 0.45. These two events may have a lot of topics in common but clearly the magnitude of such conversations is not balanced between the two.
These examples and some others are showed in the tables 4 and 5. There we examined some vents from each category.

	Performance	Art	Media	Science
Performance	Estathe Market Sound Piano City Milano 65	Radio Italia Live Museo Ideale 105	Vogue Fashion Night Out Milano Film Festival 41	Radio Italia Live Spinosaurus 10
Art		Leonardo da Vinci Double Bind Around 61	Museo Ideale Bookcity 149	Aqua Shock Wired Next Fest 12
Media			Piano City Milano La mia Basilicata 30	Wired Next Fest Bookcity 95
Science				Wired Next Fest Spinosaurus 10

Table 4: Samples of events across categories and the volume of their mutual tags

Event/Category		Limited tags	Pearson Coefficient
Estathe Market Sound (Performance)	Piano City Milano (Performance)	64	0.45
Radio Italia Live (Performance)	Museo Ideale (Art)	101	0.85
Vogue Fashion Night Out (Performance)	Milano Film Festival (Media)	40	0.61
Radio Italia Live (Performance)	Spinosaurus (Science)	8	0.75
Leonardo da Vinci (Art)	Double Bind Around (Art)	60	0.79
Museo Ideale (Art)	Bookcity (Media)	143	0.45
Aqua Shock (Art)	Wired Next Fest (Science)	12	0.55
Piano City Milano (Media)	La mia Basilicata (Media)	23	0.63
Wired Next Fest (Science)	Bookcity (Media)	94	0.38
Wired Next Fest (Science)	Spinosaurus (Science)	10	0.98

Table 5: Pearson coefficient analysis for events based on limited tags

The previous analysis was based on the hashtags that were used mutually in events. In addition it is important to probe into these similarities and reach them in a textual level. Obviously there are many hashtags that were used in all categories, but the ones that are vital to the study are those that have occurred several times and in several occasions. So we chose the top ten mutual hashtags between each two events and used them as a scale to create the figure 30. It is a textual representation of mutual tags in all events. The texts are the hashtags used to address the events and their size is proportional to their frequency on Twitter.

Obviously the most frequent hashtags in all events is "milano", that represent the main venue of the Expoincitta program. Simply by viewing this presentation, one can guess about the main topics that were discussed mostly in Expoincitta events, even in case he doesn't have any information about the program in advance.

Figure 30: Diversity of top common tags between events

There are many topics being discussed in all events and only a portion of them is mutually used in the program. So in order to better grasp the volume of topics discussed in each event and the portion that is considered mutual, we came up with the figure 31, which is a representation of all topics people mostly talked about on Twitter.

This representation is a model of topics that corresponds to the four groups of events involved in Ecpoincitta. A color is assigned to the topics of each category, and those that are common in between them have a mixed combination of the source category colors. Clearly the most frequent hashtags are those that are particularly relevant to their own category like the names of events. But among all, "milano" is the one that is being repeated in all categories with a notable frequency, as we discussed it was the most frequent mutual tag in the program.

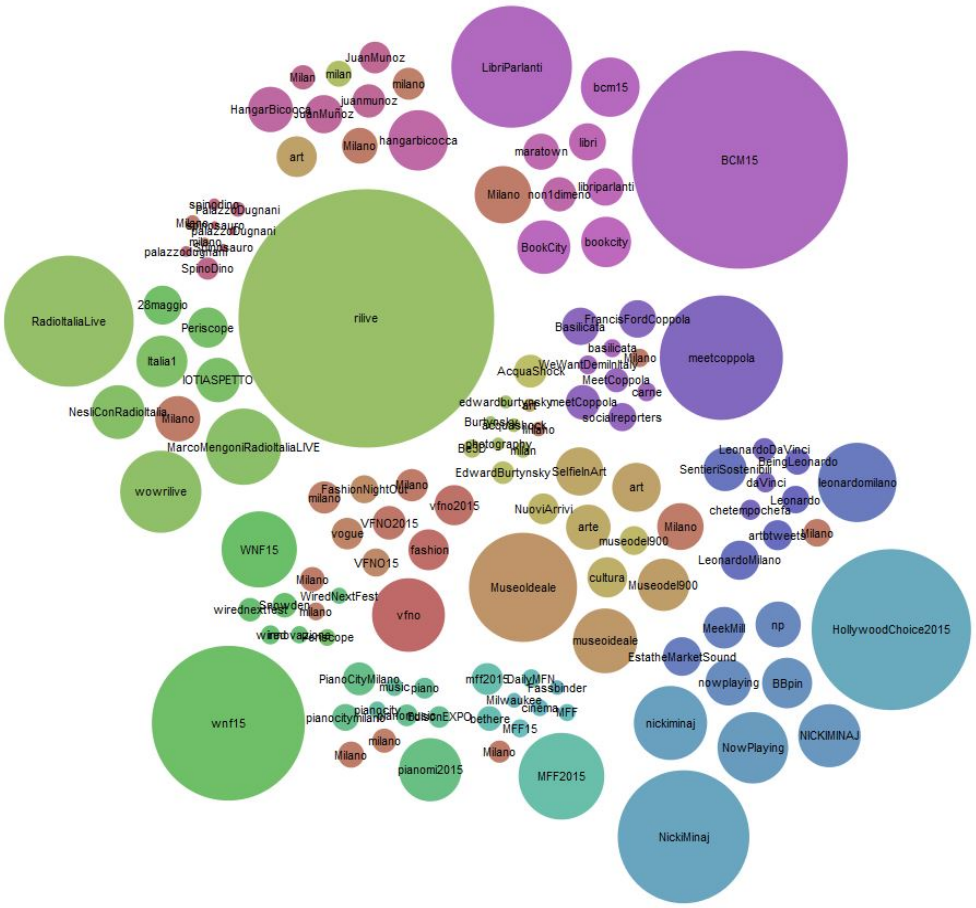

Figure 31: Common Tags and their Frequency among Events

6.3 Discussion of Results

The evaluation was divided into categories for the sake of presenting a better frame to comprehend the goals of the study. However, data intelligence is the art of binding together the result of analytical foundations. Therefore this section is dedicated to bring together the insights that were specified by the defined metrics.

As it was discussed in the general form of the evaluation, the category of Performance was greater in volume of posts and also the number of users involved. However the propagation of such volumes is not equal for all events of this category. So only by the size of a cluster, we cannot firmly state that all events belonging to that are necessarily exceeding in user engagement.

When comparing two social media networks, we observed that for the categories of Performance and Art, users tend to post on Instagram more than Twitter, and for categories of Media and Science the portion of tweets exceeded the posts on Instagram. As a result, users are more willing to share photos for concerts and artistic events, whereas in scientific and cultural incidents, they prefer to express themselves on a textual platform. In addition the stream of posting also differs from Twitter to Instagram. In all events the maximum number of reactions on social media was observed during their occurrence, yet the stream of data is stronger before the event for Twitter and after the event period for Instagram. This conveys that users tend to post about something before it happens and share a photo of it after it is finished.

The highs and lows in the stream graph of data during an event are directly influenced by the incidents in real life. As we saw in the example of "Milano Film Festival" where there is a schedule planed for everyday by it's organizer on their website, the fluctuations that occurs along the period of this event is easily traceable. This helps making sense of the anomalies found in the events data streams on social media networks.

There are many elements involved in analyzing the behaviors of users on social media, like social and sociological factors that cannot be counted in the study. This makes user analysis a complex term that might have contrasts with our expectations.

When analyzing data about Twitter users, we found out that the most appealing events to Expoincitta audience belonged to the category of Performance. In all the measures like number of participations, being favorable and the volume of data transmitted through the network. On the other hand the scientific events were the ones that got the least attention on social media.
Yet in terms of popularity, events and categories that were different in number of users almost got the same attention. So this defies the theory that by attracting more audience on social media, the popularity of that field increases.
The posts that were transmitted the most by users on Twitter belong to the category of Performance that has the most posts and the most number of users. Still in comparing the relevance between the volume of posts and tweet propagation volume, rises some paradox. Therefore there are events that are not much received from the audience but their posts were retweeted more than their popular peers.

The feature that was studied to retrieve information about the topics shared on Twitter is Hashtag. By exploiting the top most frequent topics in all categories, it is evident what were the most favorable content for users in each category. In this case if we combine the temporal stream and topic extraction, it is fairly easy to trace the subjects that people were concerned about during the period of the event. By representing the textual form of such outputs, it is almost simple to guess the concept of the program even without any prior knowledge about it.

Furthermore, in this thesis the relations between events are derived from their correlation in measuring mutual users and topics. This resulted in the fact that the events of each group share a solid connection with each other. And the events that come from the same social or cultural concept share most similarities as well. The figure 32 and 33 offers the number of these mutual elements between events.

Events	Performance				Art				Media			Science	
	Estathe Market Sound	Radio Italia Live	Vogue Fashion Night out	Piano City Milano	Acqua Shock	Museo Ideale	Double Bind Around	Leonardo da Vinci	Milano Film Festival	La mia Basilicata	Bookcity	Wired Next Fest	Spinosaurus
Number of Posts	51276	13441	3464	1151	614	3273	1728	1869	1728	1891	10327	3237	113
Number of Users	25429	5039	2206	493	277	746	811	742	584	460	2654	869	61
Number of hashtags	4061	2210	2053	670	212	1325	966	559	641	511	3000	921	91
Number of Retweets	145470	31815	10379	1926	709	37545	1661	4330	1646	2772	16414	5585	48
Number of Likes	207027	56020	21605	1449	804	26155	2814	5009	2327	3335	20498	8059	79

Table 6: General measures on Twitter

commonTags	acqua-sho	radio-itali	museo-ide	vogue-fas	wired-nex	piano-city	milano-fil	estathe-m	leonardo-	la-mia-bas	bookcity	double-bi	spinosaurus
acqua-shock	0	13	31	21	12	15	11	13	18	7	22	26	4
radio-italia-live		0	105	106	87	62	44	135	66	47	122	76	10
museo-ideale			0	92	69	72	31	85	91	44	149	131	17
vogue-fashion-night-out				0	68	46	41	176	52	37	86	94	11
wired-next-fest					0	55	35	75	52	39	95	48	10
piano-city-milano						0	34	65	42	30	74	55	12
milano-film-festival							0	42	27	24	59	29	6
estathe-market-sound								0	44	36	81	82	11
leonardo-da-vinci									0	36	82	61	11
la-mia-basilicata										0	80	29	7
bookcity											0	90	12
double-bind-around												0	18
spinosaurus													0

Figure 32: Similarity measurements based on common tags between events

commonUsers	acqua-sho	radio-itali	museo-ide	vogue-fas	wired-nex	piano-city	milano-fil	estathe-m	leonardo-	la-mia-bas	bookcity	double-bi	spinosaurus
acqua-shock	0	8	18	3	5	14	15	8	18	5	27	24	3
radio-italia-live		0	31	41	55	27	18	49	44	18	72	23	7
museo-ideale			0	12	20	43	27	12	46	15	82	66	7
vogue-fashion-night-out				0	15	10	11	58	8	3	18	9	0
wired-next-fest					0	34	24	24	26	24	64	19	3
piano-city-milano						0	29	16	29	12	80	43	6
milano-film-festival							0	12	18	12	58	31	1
estathe-market-sound								0	10	7	26	12	2
leonardo-da-vinci									0	13	57	24	5
la-mia-basilicata										0	60	10	3
bookcity											0	76	6
double-bind-around												0	5
spinosaurus													0

Figure 33: Similarity measurements based on common users between events

Chapter 7

Conclusions

This thesis is focused on employing social media data to monitor a series of live events to extract insights about the behaviors of users and highlights of the events. We proposed customized analysis techniques to apply to such dataset.

We collected the data relevant to a program of series of events that happened almost at the same time of EXPO 2015 in Milan. We started by extracting data from two social media APIs, Instagram and Twitter. We chose a noSQL database like MongoDB to store data with JSON format, because of its flexibility and scalability in dealing with massive amount of information. Then the Aggregation Pipeline was mainly used to excerpt the meaningful information.
Following that, by considering the objectives we set for this kind of dataset, we applied different metrics in analysis stage. We used quantitative measurements to study the cluster of events and their behavior through the time period of the program. For each cluster we applied user engagement measures, we extracted the most influential users and also the popularity of its events based on the users reaction to posts, which showed the level of commitment of users to each category of the program. Mining the textual posts from users of all categories helped in performing topic analysis and finding the top trends of data between events. Furthermore we detected correlations between clusters and events by finding the similarities between their users and contents. This gave us a better insight on how events behaved with respect to one another.
Finally we modeled the outcomes of the analysis stage with the help of JavaScript Libraries, in order to have a better, more flexible visualization experience.

The challenging part of the process it is to maintain the validity of data with respect to many elements involved in each stage. The outcome of the process depends relatively to the proper keywords and temporal range that were applied to gather data from social media. In the process of filtering data, some parts of the implementation have to be handled manually as MongoDB is not an automated analysis tool. So to perform some measurements, the results were extracted as CSV file and stored in Microsoft Excel. During the analysis stage, it should be noted that the data is user generated, so there are many factors that affect the flow of information, which cannot be inspected fully. For example as the users choose the content to post about something, they may use any topic that comes to mind, and this can be influenced by social, sociological, political and many other factors. Also as an endeavor to filter the data properly we choose to define a variable as a limit in the analysis, like in the Pearson coefficient analysis, we risk loosing the rest of data and yet the accuracy is not guaranteed. This is why most of our study relies on the quantitative measures in order to have a solid outcome.

As a future work we envision the application of this approach in various fields that tend to harness the power of social media to carry out their objectives on a similar dataset. The analysis methods that were used in this study are fairly applicable to any settings that include sub sections which have characteristics like events for instance they correspond to live matters; they have some diversities and mutual features. In marketing industry it will boost the process of receiving feedback and decision making from different trends based on the accumulative outcomes from users. Sociology can study the interactions between groups of people based on their behavior toward a subject. Politicians can benefit from the insights of the analysis in conducting elections where the people's opinions can lead to a great change.

Bibliography

Zubiaga, A., Liakata, M., Procter, R., Bontcheva, K., & Tolmie, P. (2015). Crowdsourcing the Annotation of Rumourous Conversations in Social Media. NY: World Wide Web Pages.

Veltman, N. (2013). Retrieved from schoolofdata: http://schoolofdata.org/

Asur , S., & Huberman , B. A. (2010). Predicting the Future With Social Media . *IEEE/WIC/ACM* .

Becker, H., Iter, D., Naaman, M., & Gravano, L. (2012). Identifying content for planned events across social media sites. *Web search and data mining.* ACM.

Bolioli, A., Salamino, F., & Porzionato, V. (2013). Social Media Monitoring in Real Life with Blogmeter Platform . *ESSEM.*

Bruns , A., & Stieglitz , S. (2013). Towards More Systematic Twitter Analysis: Metrics for Tweeting Activities. *Social Research Methodology* .

Churchill , E. F., Kennedy , L., & Shamma , D. A. (2009). Tweet the Debates. *SIGMM.*

Cha, M., Haddadi , H., Benevenuto , F., & Gummadi , K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *Weblogs and Social Media* .

Choudhary , A., Agrawal , A., Patwary , M. A., Narayanan , R., Palsetia , D., & Lee , K. (2011). Twitter Trending Topic Classification . *Data Mining Workshops* .

Diakopoulos , N., & Shamma , D. (2010). Characterizing Debate Performance via Aggregated Twitter Sentiment. *SIGCHI.*

Diakopoulos, N., Naaman,, M., & Kivran-Swaine, F. (2010). Social Media Visual Analytics for Journalistic Inquiry. *IEEE Symposium.*

Fan, W., & Gordon, M. D. (2014). Unveiling the Power of Social Media Analytics. *COMMUNICATIONS OF THE ACM* .

Ferhatosmanoglu, H., Akcora,, C. G., Bayir,, M. A., & Demirbas, M. (2010). Identifying Breakpoints in Public Opinion. *First Workshop on Social Media Analytics* .

Fonseca , H., Salvador , P., & Nogueira , A. (2015). Tracking Social Networks Events. *nternet Monitoring and Protectio* . Aveiro: DETI-University of Aveiro .

Grimes, S. (2008). Retrieved from breakthroughanalysis: http://breakthroughanalysis.com

Haldar , N. A.-H., Abulaish , M., & Azam , N. (2015). Twitter Data Mining for Events Classification and Analysis . *Soft Computing and Machine Intelligence* .

Hea , W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry . *International Journal of Information Management* .

Leskovec, J. (2011). Social Media Analytics. *ACM SIGKDD conference* . Stanford University .

MongoDB inc. (2015). *MongoDB: Bringing Online Big Data to Business Intelligence & Analytics.* white paper, newYork.

Morstatter , F., Kumar , S., & Liu , H. (2013). *Twitter Data Analytics.*

Suttona , J., Gibsonb , B., Phillipsb , N. E., Spiro, E. S., Leage, C., & Johnson, B. (2015). A cross-hazard analysis of terse message retransmission on Twitter.

Sinha , P., Choudhury , A. D., & Agrawal , A. K. (2014). Sentiment Analysis of Wimbledon Tweets. *Making Sense of Microposts* .

Sint, R., Schaffert, S., Stroka, S., & Ferst, R. *Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis.* Siemens AG , Austria.

Thelwall, M., Buckley,, K., & Paltoglou, G. (2011). Sentiment in Twitter Events. *American Society for Information Science and Technology,* .