FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Social Media Text Processing and Semantic Analysis for Smart Cities

João Filipe Figueiredo Pereira

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Rosaldo José Fernandes Rossetti

Co-supervisor: Pedro dos Santos Saleiro da Cruz

June 23, 2017

Social Media Text Processing and Semantic Analysis for Smart Cities

João Filipe Figueiredo Pereira

Mestrado Integrado em Engenharia Informática e Computação

June 23, 2017

Abstract

With the rise of Social Media, people obtain and share information almost instantly on a 24/7 basis. Many research areas have tried to extract valuable insights from these large volumes of freely available user generated content. The research areas of intelligent transportation systems and smart cities are no exception. However, extracting meaningful and actionable knowledge from user generated content is a complex endeavor. First, each social media service has its own data collection specificities and constraints, second the volume of messages/posts produced can be overwhelming for automatic processing and mining, and last but not the least, social media texts are usually short, informal, with a lot of abbreviations, jargon, slang and idioms.

In this thesis, we try to tackle some of the aforementioned challenges with the goal of extracting knowledge from social media streams that might be useful in the context of intelligent transportation systems and smart cities. We designed and developed a framework for collection, processing and mining of geo-located Tweets. More specifically, it provides functionalities for parallel collection of geo-located tweets from multiple pre-defined bounding boxes (cities or regions), including filtering of non complying tweets, text pre-processing for Portuguese and English language, topic modeling, and transportation-specific text classifiers, as well as, aggregation and data visualization.

We performed empirical studies and implemented illustrative examples for 5 cities: Rio de Janeiro, São Paulo, New York City, London and Melbourne, comprising a total of more than X millions tweets in a period of 3 months. The topic modeling and text classifiers were evaluated with manual labeled data specifically created for this work. Both software and gold standard data will be made publicly available to foster further developments from the research community.

Resumo

Devido à ascensão das Redes Sociais, as pessoas obtêm e partilham informação quase que instantaneamente 24/7. Muitas áreas de investigação tentaram extrair informações importantes destes grandes volumes de conteúdo, gerado por utilizadores, e livremente disponíveis. As áreas de invetigação de sistemas inteligentes de transportes e de cidades inteligentes (*smart cities*) não são excepção. Contudo, extrair conhecimento acionável e significativo de conteúdo gerado por utilizadores exige um esforço complexo. Primeiro, cada serviço de social media possui as suas próprias especificidades e restrições para o método de recolha dos dados; em segundo lugar, o volume de mensagens produzidas pode ser esmagador para o processamento automático e prospeção; e por último, não menos importante, os textos das redes sociais são, geralmente, curtos, informais, com muitas abreviações, jargões, gírias e expressões idiomáticas.

Nesta dissertação, tentamos abordar alguns dos desafios acima mencionados com o objectivo de extrair conhecimento de mensagens das redes sociais que possam ser úteis no contexto de sistemas inteligentes de transportes e cidades inteligentes (*smart cities*). Nós idealizamos e desenvolvemos uma *framework* para a recolha de dados, processamento e prospeção de Tweets geo-localizados. Mais especificamente, a *framework* fornece funcionalidades para a recolha paralela de tweets geo-localizados de *bounding-boxes* (cidades ou regiões), incluindo filtragem de tweets não preenchidos, pré-processamento de texto para a língua portuguesa e inglesa, modelagem de tópicos e classificadores de texto específicos para transportes, bem como, agregação e visualização de dados.

Realizamos estudos empíricos e implementamos exemplos ilustrativos para 5 cidades: Rio de Janeiro, São Paulo, Nova York, Londres e Melbourne, perfazendo um total de mais de X milhões de tweets em um período de 3 meses. O modelo de tópicos e os classificadores de texto foram avaliados com dados manualmente anotados e criados especificamente para este trabalho. Tanto os dados quanto o software criados serão disponibilizados publicamente para promover novos desenvolvimentos da comunidade de investigação.

Acknowledgements

João Pereira

"You should be glad that bridge fell down.
I was planning to build thirteen more to that same design"

Isambard Kingdom Brunel

Contents

List of Figures

LIST OF FIGURES

List of Tables

Abbreviations

SC	Smart City
SM	Smart Mobility
ITS	Intelligent Transportation System
ICT	Information and Communication Technology
SMA	Social Media Analytics
HTTP	Hypertext Transfer Protocol
TSL	Transport Security Layer
POS	Part-of-speech
BoW	Bag-of-words
VSM	Vector Space Model
LDA	Latent Dirichlet Allocation
CRF	Conditional Random Fields
HHM	Hidden Markov Model
ABSA	Aspect-based Sentiment Analysis
SSWE	Sentiment Specific Word Embeddings
ML	Machine Learning
SVM	Support Vector Machines
NB	Naïve Bayes
ME	Maximum Entropy
RF	Random Forests
DL	Deep Learning
MAE	Mean Absolute Error
OLS	Ordinary Least Squares
LR	Logistic Regression

Chapter 1

Introduction

1.1 Context and Motivation

In the last few years, the rise of Web 2.0, seen as the evolution of conventional Web services into collaborative and social platforms [Chi08], conducted to an excessive amount of User Generated Content [KH10] (UGC) being placed *online* by the population. Due to this emergency of web-content, the research community has been exploring it in order to extract added-value information regarding a large diversity of domains, such as opinion mining, human behavior and respective activity patterns, political issues, social communication (e.g. news websites). Social media platforms, more specifically, social media content (SMC), a type of UGC, has been targeted by several scientific researches focused mostly in the text mining area. Although the application of SMC in the previous mentioned domains, the *smart cities* [BAG+12] and, in particular, the transportation [GTGMK+14] domain are under a smooth growth, meaning that a large path is still unexplored allowing new opportunities and challenges for the research community to reach its full potential [MSLG15].

Availability and authenticity are some of the social media content advantages considering that such information do not require additional costs regarding its exploration, is, *a priori*, generated by humans, transcending a certain level of credibility and, lastly, due to the availability of tools provided by social media platforms, we can store the data and perform off-line analysis [KMN+17]. Twitter is considered a MicroBlog, a type of social network, which content is similar to SMS-like messages, characteristic of a 140-characters length, and the 11th most visited website in the

world [1]. This platform has already proved its value and potential in domains ranging from news detection [SST+09] to real-time traffic sensing [CSR10] being for this reason one of the most explored sources of data during the conduction of research studies.

Mining Twitter data is although the availability and free cost, a laborious and time-consuming process due to the restrictions and difficulties present in its content. The informal language, the existence of slang, abbreviations, jargons and the short length of the message are some of the problems when analyzing this data. Harvesting tweets automatically and, at the same time, extracting valuable information for the target domains delineated in this dissertation makes the task even more complex. However, by surpassing the previous mentioned problems, the extracted information may be of extremely importance and useful to the final stakeholders, namely *smart cities* and transportation entities, during decision-making policies to improve their services.

1.2 Problem Statement

The problem around this dissertation is focused in the analysis of a continuous flow of social media streams provided by Twitter. To analyse such streams, multiple steps composed in an iterative process are needed in order to filter out non-related content and proceed with extraction of information about a specific scenario. Here, since the target scenarios are associated to *smart cities* and transportation domains, data related to it must be explored and analysed. To the best of our knowledge, there are no public datasets related to these domains and the creation of a gold standard dataset constitutes a complex endeavor, which is, for this reason, an obstacle to surpass in this dissertation. The extraction of information from social media content is another overwhelming task since it is necessary the application of several NLP methods in order to minimize/extinct its peculiarly problems. Hence, the main problem can be divided in five distinct sub-problems:

1. **Data collection method for various locations**
   Choosing a method to collect data that provides a large range of valuable information for different cities constitutes the first sub-problem.

2. **Content filtering**
   It is necessary the guarantee of that all information is fully related to the target scenario in the analysis, as well as removing messages which does not brought additional information (for instance, tweets only composed by *emoticons*) or are not related to the end-users expectations, i.e. if we are targeting content from a specific city, we must assure that such content is indeed posted when users were there.

3. **Identification of topics in Twitter messages**
   The identification of topics in Twitter messages is a very important point in the analyses of the *smart cities* context. This task allows the identification of what is been talked about recently and also where the conversation topics are geographically distributed.

[1] http://www.alexa.com/siteinfo/twitter.com

4. **Travel-related classification**
   In order to produce valuable information for the transportation services, we need to analyse the content of a message and verify if it is truly related with the domain in study. Hence, discriminate travel-related tweets is one of the sub-problems that must be tackled.

5. **Data aggregation and visualization**
   The aggregation of the results provide by all other tasks is needed. This aggregation task may be continuously calculating the results in order to make the user experience easier and smooth without taking to much response time by the data visualization UI. The graphical visualizations should be of easy interpretation by the end-user and having this in mind some qualitative and quantitative indicators may be presented.

1.3 Goals and Expected Contributions

Following the previous mentioned problem in Section 1.2, the main goal of this dissertation passes through the development of a prototype framework based on the concept of analysis. Such framework demands a solution for each of the aforementioned sub-problems, and for that reason modularity is needed in the design and implementation of the final tool. Its usability will be directed to companies or even ordinary users and should be able to provide relevant information about a specific real-world scenario under the *smart cities* and transportation fields. The framework should be capable of automatically processing social media texts, more specifically, general topic detection and characterization of travel-related tweets. The following list summarizes the crucial goals behind this dissertation:

- Extraction of valuable information from Social Media Content to the Transportation and *Smart Cities* domains;
- Designing and implementation of a framework capable of automatize the analysis process;
- Application, when possible, of recent advances in the area of text analysis;

In terms of expected contributions, we hope that such generated information through the framework data analytics may be relevant both to ordinary users of a particular service and to the responsible entities in order to improve decision-making policies.

1.4 Publications

In this section several scientific contributions performed during the period of this dissertation are mentioned:

- João Pereira, Arian Pasquali, Pedro Saleiro and Rosaldo J. F. Rossetti. Transportation in Social Media: an automatic classifier for travel-related tweets. In *Portuguese Conference on Artificial Intelligence* (EPIA), 2017. In Press.

- João Pereira, Arian Pasquali, Pedro Saleiro and Rosaldo J. F. Rossetti. Classifying Travel-related Tweets using Word Embeddings. In *International Conference on Information and Knowledge Management* (CIKM), 2017. Under review.
- João Pereira, Arian Pasquali, Pedro Saleiro and Rosaldo J. F. Rossetti. Characterizing Geo-located Tweets in Brazilian Megacities. In *IEEE International Summer School on Smart Cities* (IEEE S3C), 2017. Under review.

1.5 Dissertation Structure

The effort applied to this dissertation generated a great diversity of points and due to that the remainder of the document is structured into four chapters. Section 3 starts with a brief conceptualization in the Smart Cities and Intelligent Transportation System domains, as well as previous related works using social media content as its basis. The proposed framework is referenced in Section 4, being each its composing modules depth described. Experiments performed to test each module of the framework are reported in Section 6. Completing this document, conclusions, future work and a few final remarks are exposed in chapter 7.

Chapter 2

Background and Literature Review

This section aims to analyse and reflect about some works and topics that will be relevant to fully understand the problem. The study of solutions found by other authors can simplify the difficult task that is the analysis of social media data. Hence, this section has been divided into several parts in order to perceive not only the environment in which the problem is located but also the most important points to be studied in order to build our the final product. Respectively to the problem scope its important to know what is a smart city and how the transportation system can contribute to this meaning. Since our product is a framework which goal is to extract information about social media data, i.e. texts from the Twitter, its interesting obtain the knowledge about extraction tools, such as the Twitter APIs, in order to have an idea how to construct our crawler module. The meaning of text mining and how the information present in texts can be extracted through different kinds of techniques, regarding disambiguation, filtering and modeling. Finally, it is also important to analyze several works about sentiment analysis in order to know the different methodologies and which are the most advantageous for this problem.

2.1 Smart Cities and Intelligent Transportation Systems

Chapter 3

Old Background and Literature Review

This section aims to analyse and reflect about some works and topics that will be relevant to fully understand the problem. The study of solutions found by other authors can simplify the difficult task that is the analysis of social media data. Hence, this section has been divided into several parts in order to perceive not only the environment in which the problem is located but also the most important points to be studied in order to build our the final product. Respectively to the problem scope its important to know what is a smart city and how the transportation system can contribute to this meaning. Since our product is a framework which goal is to extract information about social media data, i.e. texts from the Twitter, its interesting obtain the knowledge about extraction tools, such as the Twitter APIs, in order to have an idea how to construct our crawler module. The meaning of text mining and how the information present in texts can be extracted through different kinds of techniques, regarding disambiguation, filtering and modeling. Finally, it is also important to analyze several works about sentiment analysis in order to know the different methodologies and which are the most advantageous for this problem.

3.1 Smart Cities and Intelligent Transportation Systems

The Smart City concept appeared thanks to the continuous growth of a city's population which has contributed to an aggressive urbanization [URS16]. In the last few years, several definitions for its meaning have emerged, but the ideal one is not yet fully known. Angelidou in [Ang15] defines Smart City as a "conceptual urban development model on the basis of the utilization of human, collective, and technological capital for the development of urban agglomerations" and enhance as its primary key the knowledge and the innovation economy. In her work, there is an identification of four forces that model the concept of a Smart City and two of them are very important to enhance: *technology push*, where new products and solutions are introduced in the market regarding the fast advance in science and technology; *demand pull*, where solutions and problems are developed in order to respond to the society demands, like the continuous growth of the population [Ang15].

The development environment in a city tagged with the concept "smart" is another key factor to reach the success. Komninos focus the importance of collective sources of innovation to the improvement of life quality in cities. The globalization of innovation networks are the responsible for the emergency of another types of environments, such as "global "innovation clusters and i-hubs, intelligent agglomerations, intelligent technology districts and intelligent clusters, living labs" making possible the experimentation of products or services by the population in order to identify problems or even to analyse the behavior of the people regarding what they have experimented [Kom09].

The transportation system is inherently connected to the progress of a city, since people on a daily-basis uses the several ways of transportation, i.e. bus, private cars, metropolitan, etc, to go to their jobs and make their own life. This system is also influenced by the problem of the population growth being relevant the need of finding solutions to minimize or even raze it [CD15]. Hence, "a smart city should be focused on its actions to become smart", coming up the concept of innovation [URS16].

To understand what are *Intelligent Transportation Systems*, it is crucial introduce the meaning of Smart Mobility. SM is a combination of comprehensive and smarter traffic service with smart technology, enabling several intelligent traffic systems which provide control in the signals regarding the traffic volume, information about smooth traffic flows, times of bus, train, subway and flight arrivals and their routes [CL15]. The majority of *Intelligent Transportation Systems* are expressed through smart applications where the transportation and traffic management has became more efficient and practicable, allowing the users to access important information about the transportation systems in order to make correct decisions about what they want to use in their cities [CD15]. ICT-based infrastructures are the main support for Smart Cities when the focus are ITS, since through that is possible to pilot the activities operations and its management over a long period of time [URS16].

Nowadays, cities are exploring some initiatives of sensing to support the development of technological projects. Areas such as utilities management (where, for example, is monitored the

consumption level of power, water and gas), traffic management (using vibration sensors to measure the traffic flows on bridges, or even the full capacity of a parking lot), environment awareness (using video cameras to monitor the population behaviour and sensors to measure the level of air pollution) make use of physical sensors, i.e. some devices that can capture information to study and improve the quality of life in a daily basis [DSGD15]. R. Szabo et al. [SFI+13] and D. Doran et al. [DSGD15] report the highly economic cost that this kind of sensing needs, since it's require the maintenance and replacement of this devices, as well as a tracking infrastructure store and treat the information collected. Hence, a new form of sensing has emerged - Crowd Sensing - to offer the cities several ways to improve their services by exploring the participation of the population in the social networks where there are a publicly share of citizen's opinions and thoughts regarding some problems [RMM+12]. This type of sensing consists in *human-generated* data provided by the population through the use of mobile devices and the social networks platforms. Such data can be further used to extract some analytics regarding specific services in a city, namely the urban transportation system [RMM+12]. Based on all this, social media can be seen as a good source of data to extract valuable information in order to direct it to the smartness evolution process of a city [SFI+13].

Several works have already been developed and presented taking into account these two large areas, Transportation Services and Smart Cities, using social media as source of information. G. Anastasi et al. [AAB+13] proposed a framework which objective was the promotion of flexible transportation systems usage, i.e. encouraging people to share transport or to opt for the use of bicycles in order to minimize infrastructural and environmental problems. Their tool takes advantages of the crowd sensing techniques by exploring social media streams to predict accidents or traffic congestion and alert the users of their service about this type of events. W. Liu et al. [LAR12] have made a study in three different transportation modes (private cars, public transportations and bicyclists) using theirs channels on Twitter to estimate a percentage of the majority gender that uses this services in the city of Toronto. They have extract all the channel's tweets appealing only to the *non-protected* followers and applied an already developed classification model to label each tweet with its creator gender: male or female.

T. Ludwig et al. [LSP15] proposed a tool capable of collect and display social media streams in order to help the integration and coordination of volunteers in actions performed by emergency services to prevent engagement in dangerous areas. Their tool present to the end-users map visualization of a city where they could identify public calls of the emergency services to accept or deny them.

In conclusion to everything that has been analyzed in this sub-section, it's possible to verify that the cities are increasingly opting for technological opportunities that involve crowd sensing, once this type of exploration brings a considerable reduction of costs and the information that is collected may contribute to the extraction of value from data generated by the population itself.

3.2 Social Media Analytics

In the last few years social networks have made impact on the business communications, since users assume the role of costumers through the publication of content on this networks, rising the levels of interaction between users and businesses entities [URS16]. A proof of that is the amount of information produced since 2011 which is equivalent to a number over than 90% of the available data online [SIN13]. Facebook, Twitter and other social networking sites are nowadays used as business tools by companies aiming the efficient use of digital marketing techniques to publicize their products [RL14]. Besides the business field, the population turn into this new communication technologies in a intensely way, where they publicy share real-life events, their opinions about certain topic, their on-time feelings in the network through a simple message [DDLM15]. Social Media Analytics can be describe as a type of digital analytics to study the people interaction with others, or their opinion about companies, its products and services through the social media data. This study provides important information to "analysts, brands, agencies or vendors", and its analysis could facilitate the generation of economic value to many organizations [Phi12]. To achieve the main goals of the SMA, the companies focus their effort in the development and evaluation of frameworks, to make possible an easy collection, analyse, summarization and visualization of processed social media data. Hence, the companies can establish specific points about what to improve in their products [ZCLL10]. To create a significantly value regarding the SMA, J. Philips in [Phi12] enhance some important factors: users permissions, the listening of real-time information, the search mechanism, the data access and integration, and others, before the choice of a tool that allows the information collection. Besides the tool, is also important have an idea of what is need to explore because the use of a wrong technique of SMA could have bad business impact for the company. The majority of SMA techniques focus on modeling in order to understand the large range of data collected and support techniques, such as sentiment analysis, trend analysis and topic modeling, are the most commonly used [FG13].

3.2.1 Twitter

Twitter is a social network where people freely micro-blogging about any topic and, like any other social network, makes possible the connection between users around the world [SFD+10]. This social network has faced an exponential growth since its inception, and nowadays its users, which surpass 200 millions, produce around 500 millions tweets daily, performing a massive bunch of information that could be an ideal testbed for research projects on big data [LIR15, GSZS14]. N. Banerjee et al. [BCJ+12] classify Twitter as a micro-blogging service presenting three attributes that justify such characterization.

- ***Limited Context Information*:** The length of a Twitter message never exceeds 140 characters, which, in cases of knowledge extraction, since the amount of information is short, the final results could not be the expected and none knowledge contribute is obtained.

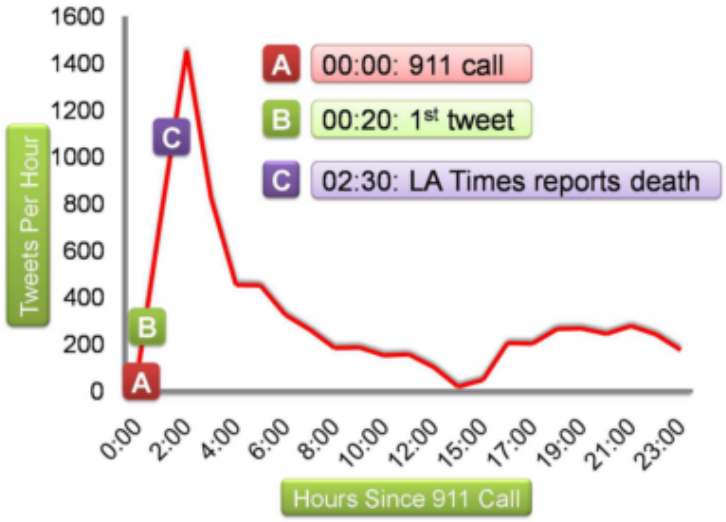


Figure 3.1: Traffic in tweets per hour relating to Michael Jackson's death by J. Sankaranarayanan [SST+09]

- ***Richness of Exchange*:** Symbolizes the great diversity of posts that exists in the micro-blog network. People talk about their daily activities, have conversations with each other and shares their thoughts (moods or opinions) about a certain event or topic.

- ***High Dimensionality*:** The informality and ambiguity present in the messages and the expanse vocabulary present in any language makes a large dimension of data. The informality of the messages can be seen as the presence of spelling errors and abbreviations in its content, while the ambiguity can be the presence of words with multiple meanings.

Twitter has not only evolved in terms of usability, but also in the purpose of its use, i.e. Twitter is not only used as personal diary for people but also represents a source of information for reporters and journalists to find potential news about real-time events [SFD+10]. A good example is given by J. Sankaranarayanan et al. [SST+09] in the period of Michael Jackson's death. The first tweet about the incident was posted 20 minutes after the call to the 911 emergency service and, nearly, two hours before the first communication on the news as it's possible to verify in the figure 3.1.

One of the advantages of Twitter compared to other social networks, for example, Facebook, is the easier way to access its users originated data. While Facebook does not provide private information about its users unless there are permissions to do so, or the content shared is present in public pages or groups, Twitter allows the collection of all tweets from channels or directly from people in order to be analysed in any kind of project because the user's accounts are usually public [MSLG15, SDX13].

Although this freedom in the collection of data, Twitter has also a an ethical perspective and a regulation must be accomplish by developers or researchers. The TOS (Twitter Terms of Service)

was created in order to make known what can be done with the data and protect the users' rights [KC13].

3.3 Text Mining

Text mining is a derived field from Data mining and aims to extract valuable information from unstructured textual data[HZL13]. The reason why this technology is nowadays so much explored is because of the massive amount of information that is stored in text documents, such as "text files, HTML files, chat messages and emails" and it's required an automated technique that make possible the identification, extraction, management, integration and the knowledge exploration of information from texts in a efficiently and systematically way [HZL13]. On the other hand, the social media applications also have contributed to the growth of text mining usage where companies have seen a potential path to improve their business model and increase the economic value relatively its competitors.

A. Stavrianou et al. [SAN07] identify text mining as an interdisciplinary field since this technology takes advantages from Data mining techniques and combines several methodologies from similar research areas, such as Categorization, Information Extraction, Information Retrieval, Topic Tracking and Concept Linkage. A common problem related to text mining is its similarity with Information Retrieval and Information Extraction which leads people to a non-differentiation between this technologies. The difference between Information Retrieval and Text mining is established in their final goal, while IR aims to find and retrieve documents that match a certain part of a text or some keywords (e.g. Google Search Engine[1]), TM tries to discovery unknown patterns in texts that can be interpreted and explain some facts or truths contained in the lexical [SAN07, HZL13, HNP05]. Regarding the Information Extraction, the differentiation can be seen in the data specificity and structure. IE focus on the extraction of expected information from structured data and precocious relations, while the information returned by TM techniques should be unsuspected and unexpected with the data holding an unstructured format [SAN07].

The motivation behind text mining holds on the benefit that other fields of research could take from a use of its techniques. Information Retrieval systems can improve their precision since its basis is the identification of semantic relations. Several areas can explore this methodology to find inconsistencies in relational databases and make the integration, update and querying tasks easier [SAN07].

Text mining shares some of the issues presented by the Natural Language Processing field. Once texts are usually performed by humans some associated problems can appear, such as spelling mistakes, wrong phrasal construction, slang among other. Before the "mining" of a text, it's important to apply some pre-processing steps in order eliminating noisy data from the primary analysis process. A. Stavrianou et al. cite this issues very well in they work and it can be seen in Table 3.1.

[1] https://google.com

Table 3.1: Text Mining Issues by A. Stavrianou [SAN07]

Issue	**Details**
Stop list	Should we take into account stop words?
Stemming	Should we reduce the words to their stems?
Noisy Data	Should the text be clear of noisy data?
Word Sense Disambiguation	Should we clarify the meaning of words in a text?
Tagging	What about data annotation and/or part of speech characteristics?
Collocations	What about compound or technical terms?
Grammar / Syntax	Should we make a syntatic or grammatical analysis? What about data dependency, anaphoric problems or scope ambiguity?
Tokenization	Should we tokenize by words or phrases and if so, how?
Text Representation	Which terms are important? Words or phrases? Nouns or adjectives? Which text model should we use? What about word order, context, and background knowledge?
Automated Learning	Should we use categorization? Which similarity measures should be applied?

The removal of words from the text can sometimes not be desirable because some sentences can lose its information or even leads to a different meaning compared with its original form. The generation of a stop list words should be a supervised task as long as little words could induce distinct results in the text classification [Ril95].

Stemming is a task that depends mostly from the language of the text than its domain [SAN07] and the main goal of this technique is to reduce a word to its root to help in the calculus of distances between texts or even keywords or phrases.

The noisy data is derived from spelling mistakes, acronyms and abbreviations in texts and to solve this, a conversion of this terms should be done to keep a valid integrity of the data. The most commonly solution approaches involve text edit distances (Levenshtein Distance[2]) and phonetic distances measures between known words and the mispelling ones to achieve good corrections [BDF+13]

Word Sense Disambiguation focus on solving the meaning ambiguity present in words. Other similar field to WSD is Name Entity Disambiguation (NED) where the disambiguation target are named-entities mentions, while WSD focus on common words. WordNet[3] is a resource very used to extinguish this ambiguity [CSMA16]. There are two types of disambiguation, the supervised, where the task is support by a dictionary or a thesaurus [SAN07], and the unsupervised one, where the different meanings of a word are unknown and normally learning algorithms with training examples are used to achieve good results in the disambiguation task [Yar95].

[2] https://en.wikipedia.org/wiki/Levenshtein_distance
[3] https://wordnet.princeton.edu/

Tagging can be describe as the process of labeling each term of the text with a part-of-speech tag, i.e. classify each word as a noun, verb, adjective, etc [HNP05]. Collacation are very important in text mining, since this task consist in group two or more words to give the correct meaning content in the text. Collacations are usually made before the WSD task since some compound technical terms have different meaning from the individual words which composed it [SAN07].

Tokenization serves to pick up all the terms presented in a text document and to achieve this it's necessary the split of the document into a stream of words implying the removal of the punctuation marks and non-text characters [HNP05]. some authors also see tokenization as a text representation form since one of the most used models to represent texts is *Bag-of-words* (BoW). This model broke down texts into words and stores it in a vector being also presented the word frequency occurrence in the text. Hence, each word may represent a feature [SFD+10]. Another commonly used model to represent texts is Vector Space Models that represent all the documents in a multi-dimensional space where documents are converted to vectors and each vector may be seen as a feature. This model provides some advantages since the documents can be compared with each other by performing some specific vector operations [HNP05].

The purpose of this section is to provide the reader the definition of what is text mining, and also the identification of basic operations and steps that are necessary for the preprocessing of this type of unstructured data - texts.

3.4 Information Extraction

Information Extraction is an important field of text mining and its main goal is finding structured information from semi-structured or unstructured texts. This kind of information can range from the identification of entities, such as people, organizations and places names, to a relation between this concepts. In the sentence, *"At 1976, Apple was founded by Steve Jobs and his friends"*, its possible extract information about who were the founders of Apple and what was the year of its foundation. Problems regarding the analysis of the sentence can be located on the words "Apple" and "his", and how could a machine know that "Apple" is a reference of a technological company and not a reference to a fruit, or even, that the word "his" establish a connection between "Steve Jobs" and "friends". The exploration of this kind of information constitutes a potential measure to improve computer systems, such as search engines and database management [AZ12].

When the target analysis is social media data, i.e. texts from social networks, the filtering process of the data is a crucial step. It's desirable that only related-topic information should be collected in order to avoid the existence of noisy objects in the data set [URS16, SRP+13].

Information Extraction presents a set of components that can tackle this problems, such as Named Entity Recognition (NER), Named Entity Disambiguation (NED) and among others.

3.4.1 Name Entity Recognition and Name Entity Disambiguation

NER and NED are two distinct tasks that sometimes could cause some ambiguity in its purpose.

Named Entity Recognition is seen as a sub-task of Information Extraction aiming the correct labeling of words in a text in order to have knowledge of its types. The accuracy retrieve by this task is important when further steps depends on it, e.g. Relation Extraction [AZ12]. Gazetteers are commonly used in this task since they provide pre-processed lists of organizations, days, places and person names which can be matching against some terms we want to recognize. In some cases, this "tools" are not enough to solve the problem because their domain is very limited and some terms can not exist, implying the use of external knowledge to fill this lack [RR09].

There are two main approaches to conduct this task: Ruled-based and Statistical Learning. The Ruled-based approach focus on the definition of a set of features for each token in the text to further comparing the text with a bunch of rules. The rules are composed by patterns that should trigger some labeling action in a sequence of tokens. The definition of this rules usually requires human expertise [AZ12]. The Statistical Learning approach, also known as Statistical Machine Learning, treats the text as a sequence of observations which are represented by a vector of features. The final goal of this approach is the assignment of a label y_i to each observation x_i. The mapping process usually follows a BIO notation, firstly introduced to text chunking by [RM99], where the entity name could be at the beginning (B) or inside (I) of the observation and never outside (O). Patawar et al. [PP15] enhance three types of methods to this approach.

Supervised methods, where it's characteristic the existence of a labeled training data set to train a model and then classify a set of test to measure the model performance and accuracy. Hidden Markov Model was used in [BSW99] to recognize and classify some text. In [Iso01], Decision Trees were combined with a simple rule generator to prove that this method could achieve similar results as Maximum-Entropy-based methods used in [Bor99]. Support Vector Machines were used by H. Isozaki et al. [IK02] where they have proved that SVM models could also achieve good results for NER tasks if some analysis was carefully made in the *kernel functions* and filtering methods were applied, like the removal of useless features.

Semi-supervised methods used different amount of training data, i.e. labeled examples, and test data. The test data is usually a bigger amount facing the training data. The methodology commonly applied to this approach involves the use of "bootstrapping" which is an iterative process of training the model with progressive supervised increases of the training data set until the performance starts to decrease [IK10].

The last type are the Unsupervised Methods, where a large amount of labeled data is necessary and it's difficult to have this requirement. This need bases on the huge number of features required to this kind of methods. To fill this lack, a very frequent method used is the clustering where the formation of groups is made using the similarity present in the texts domain [PP15].

Y. Li et al. [LWH+13] define Named Entity Disambiguation as a "process of associating an entity name mentioned in a text to an entry, representing that entity, in a knowledge base". In the last few years, NED has been a target for a considerable number of research projects. The majority methods implemented to tackle this task are focused in three main features. Y. Li et al. [LWH+13] enhance *entity popularity* as a statistical one where there is a "assumption that the most prominent entity for a given entity mention is the most probable underlying entity for that

mention". At this feature, a link between the term and its Wikipedia page reference is established. The *context similarity* is another feature and aims the complementarity of the *entity popularity* feature. This feature centers on similarity measures between the text in analysis and the content text of the Wikipedia page. Y. Li reveals that this feature is word-dependency since it's necessary that both texts shares identical words in order to produce expected results. *Topical Coherence* is the third feature and solves the emerged problem of the second one. This feature uses the Wikipedia cross-page links mechanism in order to look up for related-topics of other entities and makes a connection with the target entity in the disambiguation problem. Through this process the domain text is expanded, decreasing the word-dependency problem appeared in the second feature.

D. Spina et al. [SGA13] present two different approaches to solve the problem of ambiguity presented in texts. The first one is *entity linking* and consists in the establishment of an association between the mention in the text and a entity present in a Knowledge Base. Three steps are needed to perform the linking of the entity name to the knowledge base:

- **Query Expansion:** Mining the Wikipedia structure and solving co-references in the document in order to enrich the query;
- **Candidate Generation:** Construction of a list of candidate entities from the knowledge base according to the information presented in the query;
- **Candidate Ranking:** It is the phase in which is computed some similarity measures between the query and the entities, in order to rank the candidates and select the best one.

Another approach to tackle this problem of disambiguation of entities presented by D. Spina et al. [SGA13] is the *document enrichment by linking to Wikipedia Articles*. Similar to the *entity linking*, this approach is also composed by three different steps and takes advantages from text representation models, such as Bag-of-words or Vector Space Models, for example.

- **Mention or surface form representation:** There a context definition of the target mention to disambiguate. Text representation models are build with all set of entities that are ambiguous and the unambiguous ones which are already resolved with linking to Wikipedia pages.
- **Candidates entities retrieval and representation:** All the candidate entities referenced by the knowledge base (e.g. Wikipedia or Freebase) and the content on the page are converted to text representation models. After this, there is extraction of some features that can range from page categories of the candidate entities or even syntactic features.
- **Best candidate selection:** The computing of similarity and distance functions between the two text representation models, produced in the two previous steps, is made to select the best candidate.

Some works related to NED were made in the context of micro-blogging services, such as Twitter.

At 2010, Ferragina et al. [FS10] developed a system capable of identity entities in short texts as, for example, tweets. Their system take advantage of the hyperlink mechanism of Wikipedia, extracting related links between pages and the anchors texts in the links. By detecting some senses present in the anchors, they try to disambiguate the ambiguous ones through a collective agreement function, i.e. a voting classification. They used the unambiguous senses to boost the selection of the ambiguous ones and have trying some pruning in the anchors set to improve the performance of the system.

Meji et al. [MWdR12], similar to Ferragina et al., also explored anchors texts in Wikipedia articles. The authors have used a supervised machine learning approach to conceive a list of candidates to disambiguate each mention present in their tweets. Their strategy focus on the identification of some patterns in each tweet, such as n-grams, to further matching it with the anchor texts of the Wikipedia articles, taking also in account the hyperlink mechanism of this Knowledge Base.

Considering this works it's possible conclude that Wikipedia is a potential source to explore in order to solve mentions of entities that could lead to a ambiguous problem.

3.4.2 Content Filtering

The content filtering is one of the most important tasks to analyze micro-blog data (e.g. Twitter). The main goal of content filtering is the classification of Twitter posts which contains an entity name, assuming the existence of a relation between the name and the content in order to erase ambiguity in the dataset [Spi14]. Recently, some contests related with Online Reputation Monitoring (ORM) have explored this task of filtering content. The WePS-3 [4] and the RepLab 2012 tackle unknown-entity scenario approaches, while the RepLab 2013 [5] focus on the known-entity scenario approach.

In the WePS-3, the LSIR [YMA10] research group has build a system where a profile identify each of the companies mention. The Wordnet [6] and the company web-page were used to extract a bunch of keywords related with the company. Combining this previously set of keywords with some manually defined they have created the profiles to the companies in analyse. They used this profiles to extract specific features from "tweets" and added to a set where there already was some generic features. This information was further used to classify the tweets with "related/unrelated" labels.

Regarding the ITC-UT [YMO+10] research group, they have, firstly, made a prediction of the company name class according to the related-tweet ratio. After this step, a distinct heuristic was found to each of the classes, using basically part-of-speech tagging and the named entity label of the company. Their approach was a two-step classification task.

[4] http://nlp.uned.es/weps/
[5] http://nlp.uned.es/replab2013/
[6] https://wordnet.princeton.edu/

SINAN [CVSO10] system used an approach of ruled based heuristics, specially the existence of the entity name both on tweets and external resources, such as Wikipedia, DBPedia [7] and the company web-page.

The RepLab 2012 follows an identically problem as the WePS-3, the unknown-entity scenario. Some research teams follow the same approach as S. Yerva et al. [YMA10] where the use of profiles describing each company mention to correctly filter the content. DAEDALUS [VRLSM+12] and OXYME [Kap12] tackle a manually exploration, such as the development of dictionaries and rules sets to the detection and the classification task, and the selection of feedback terms about the entities, respectively. The automatically methods were explored mainly with external resources. CIRGDISCO [QYOP15] and ILPS [PdRS12] used the Wikipedia, while BMEDIA [CBRB12] combined it with Freebase to extract related and unrelated concepts. CIRGDISCO proposed a two-step algorithm to solve the filtering task. The first step involves the extraction of the entity related-terms from the Wikipedia and further calculus of the IDF (Inverse Document Frequency) score for each term founded. The second step focus on the idea of concept term score propagation, i.e. to propagate the labels of the high-precision classified tweets to the remaining, in order to increase the recall measure. ILPS tackle the filtering task by using semanticising, where two probabilities are verified: Link Probability and Commonness. The first one represents the probability that an n-gram is linked to an Wikipedia page, while the second is the probability of an n-gram is linked to a certain concept. The ILPS group also used list aggregation and disambiguation techniques to carry out this task.

At RepLab 2013, the filtering task was in a known-entity scenario where the data provided to the groups consists of a collection of tweets about 61 entity names in two distinct languages, English and Spanish. Saleiro et al. [SRP+13] have devolved POPSTAR which was the system that, using a supervised learning, has obtained the best results classifying the tweets as related or non-related with the entities. Their group has explored internal features (RepLab Metadata, probabilities in the text, keyword similarities) and external features, such as Web Similarity (between tweet text and the Wikipedia page text) and Freebase scores relatively to the position of the target entity in the retrieved list.

The second best score in the filtering task at RepLab 2013 was obtained by V. Hangya et al. [HF13] where their system made usage of text normalization methods, combining the textual features with topic distribution features retrieved by a LDA (Latent Dirichlet Allocation) model. The resulting features were further used in a maximum entropy classifier to perform the filtering task. The LIA [CBB+13] group has used k-Nearest-Neighbour (kNN) algorithm with a set of discriminant features based on similarity measures. They have used Bag-of-Words representation, combining TF-IDF (Term Frequency-Inverse Document Frequency) with Gini purity criteria, for the tweets collection and calculated the Jacard similarity measure.

This kind of methodologies will be a huge step to validate our dataset since it's important to have only related-topic tweets to analyze the people's feelings, opinions about a correct entity instead unrelated ones.

[7] http://wiki.dbpedia.org/

3.4.3 Topic Modeling

The emergence of topic modeling techniques was due to the people's chase of a better understanding of the available information in document corpora. Topic models provide the discovery of certain patterns in a collection of texts and enhance specific words/terms that have a direct relation to the content information [MSBX13]. There are many studies that were conducted in order to prove that is possible to extract coherent topics from micro-blogging data using the LDA (Latent Dirichlet Allocation) model [MSBX13, HD10, ZJW+11]. LDA models are difficult to apply to micro-blogging texts because of the characteristics present in this kind of text: short, mixture of contextual clues (URLs, tags, name mentions with the '@'), informal language with many misspelling, acronyms and abbreviations [MSBX13]. L. Hong et al. [HD10] describe Latent Dirichlet Allocation as "an unsupervised machine learning technique which identifies latent topic information in large document collections". This technique uses "bag-of-words" to each document which are represented by a probability distribution over some topics, and each topic is, in turn, represented by a probability distribution over a number of words.

R. Mehrotra et al. [MSBX13] have explored the improvement of the standard LDA model using several pooling schemes of tweets, i.e. aggregating tweets by some characteristics present in its content. Their polling schemes characterization range from basic scheme: where each tweet is treated as a single document; author-wise pooling: aggregating the tweets according its author; burst-score wise pooling: tweets are aggregated by the scores obtained from the execution of a burst detection algorithm; temporal pooling: pools are formed by tweets posted at the same hour; *hashtag*-based polling: the tweets are grouped according to its *hashtag* (#) reference, and if there are more than one reference then the tweet is added to each of the groups. The authors evaluate the resulting clusters through some metrics, such as the *purity*: verifying the average of the corrected labeled tweets inside the clustering; *normalized mutual information* (NMI): its the calculus of the matching results between the clustering and the category labels; and finally the *pointwise mutual information* (PMI): measure of the statistical independence between two words regarding the close proximity. Their approaches also were studied combining similarity tag assignment (TF and TF-IDF) and the best presented results were performed by *hashtag*-based polling with TF-similarity tag assignment regarding the purity and the NMI metrics, while the best PMI metric was obtained by the simple *hashtag*-based polling method.

L. Hong et al. [HD10] also explored the LDA models through a set of schemes formed by them. Their schemes diverge between user-based and term-based groups, where the user-based are agglomerations of messages from the same user while the term-based groups are formed by messages that have the same term in the content. In their approach they have also used Author-Topic Model which is an extension of the LDA model but the main difference is that in the LDA, each document is associated with a multinomial distribution over T topics while in the AT model the association is made to the author instead the document. They used JS Divergence to study the similarity between the performed schemes. The main goal of their work was not the topic modeling but they proved that this sub-task can improve performances of classification, namely

when the messages are group by the same user.

W. Zhao et al. [ZJW+11] proposed another extension of the LDA model and named it Twitter-LDA [8]. Their model follows the idea that each tweet is about some topic, so instead of grouping the tweets into schemes and than extract some topic, they tackle each tweet as a singular problem and extract the target of the content. In their work, the evaluation of the model was made by comparing its effectiveness against the standard LDA model and the Author-topic model. The Twitter-LDA results, obtained from a small set of topics in a preliminary test, have surpass the performance of the others models (standard LDA and Author-topic models).

The last model should be a good start to face the problem of topic modeling in our work, since it's open-source tool and it's available in GitHub.

3.5 Sentiment Analysis

Sentiment Analysis is a task of NLP (Natural Language Processing) and aims the finding of the polarity in opinions, sentiments of people about a specific topic contained in a document or even the overall sentiment present in it. Research done in this area has grown at an impressive pace and this is due to the value that this type of analysis can provide to the business world. "Marketing managers, PR firms, campaign managers, politicians, and even equity investors and online shoppers are the direct beneficiaries of sentiment analysis technology" since the retrieved information can favor and make easier the decision-making process [Fel13]. This task is composed by several distinct problems and there are two main approaches to tackle it: supervised [SGS16, KWM11] and unsupervised [MSLG15, AMB+13]. Feldman in their work [Fel13] enhaces the several types of problems found in the sentiment analysis task. One of them, the document-level sentiment analysis focus on the determination of the sentiment polarity of opinions expressed by the author in his document. Another problem that is widely explored is the sentence-level sentiment analysis which is a deeper version of the previous. A document may have multiple opinions about a specific entity and in order to extract the polarity value about it, a phrase-level split is required. Some countermeasures must be taken into account in the polarization of phrases since the sarcasm component can be present in the content and it's very difficult to treat correctly this. There is another problem in this field named aspect-based sentiment analysis where the sentiment polarity should be directed to the aspects/topics contained in the document. In the following subsections a deep description and studied solution of this problem will be presented.

3.5.1 Lexicon-based vs Machine-learning based

Sentiment analysis in Twitter can be divided in three different approaches relatively to the sentiment classification: lexicon-based, machine-learning based or even a hybrid approach between the previous two. In the first place it's necessary to talk about what features are relevant or not

[8] https://github.com/minghui/Twitter-LDA

in order to tackle this problem. Aggarwal et al. [AZ12] in his work refer some of the common features used in this problem:

- ***Term presence and frequency:*** groups of words, named *n-grams*, and the frequency they occur in the document;
- ***POS Tag***: the existence of adjectives can be relevant indicators to determine the opinion polarization;
- ***Opinion words and phrases:*** words that usually transmits some polarity, such as *good and bad*, or even whole phrases that don't have this type of words, e.g. "cost me an arm and a leg";
- ***Negation***: the existence of negative words that may change the opinion orientation, such as "I don't like apples" which means the same as *hate*.

After the features engineering process, it may be necessary to select only a few ones to apply in the classification task. W. Medhat et al. [MHK14] mentioned in their work some of the most used methods in this particular step:

- ***Lexicon-based:*** It's necessary human annotation. Starts with a small set of seed words and then a bootstrapping methodology is applied to expand the lexicon domain through the discovery of synonyms in external resources;
- ***Point-wise Mutual Information***: It's a statistical method where the co-occurrence level between a given word w and a class c is computed in order to see if a feature is or not correlated with the class. The formula of calculus is given in the equation 3.1,

$$M_c(w) = \log(\frac{F(w).p_c(w)}{F(w).P_c}) = \log(\frac{p_c(w)}{P_c}) \tag{3.1}$$

  where $F(w).P_c$ is the expected co-occurrence level and $F(w).p_c(w)$ is the true value of the co-occurrence.

- ***Chi-square***: It's another statistical method used to measure the correlation between the features and the classes 3.2,

$$X_c^2 = \frac{n.F(w)^2.(p_c(w) - P_c)^2}{F(w).(1 - F(w)).P_c.(1 - P_c)} \tag{3.2}$$

  where n is the total number of documents that composed the collection, $p_c(w)$ represents the conditional probability of class c in the documents containing the word w, P_c are the fraction of documents that contain the class c and $F(w)$ are the documents fraction that contain the word w.

- ***Latent Semantic Indexing***: It's an unsupervised method that aims the reduction of the original set of features into a new ones through transformation techniques like PCA (Principal Component Analysis)

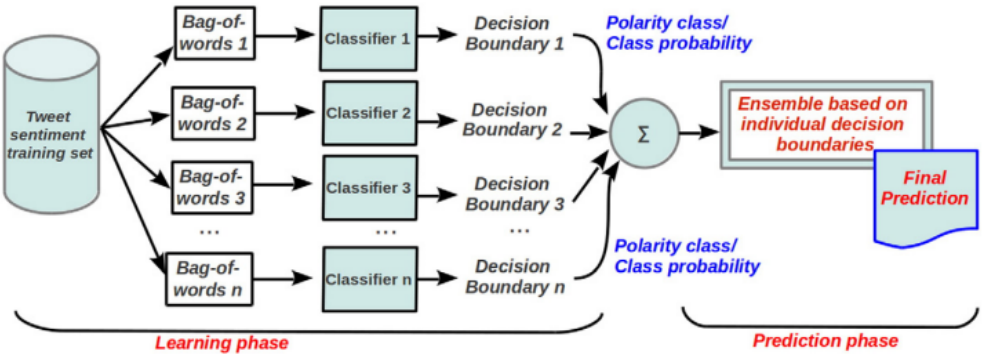
After the conclusion of this step of features selection, the sentiment classification is conducted and there are a very high number of techniques that can be applied.

A. Giachanou et al. [GC16] divided the **machine-learning approaches** into three different categories: supervised learning, the classifiers ensembles and deep learning.

Supervised learning methods focuses on the training of classification models with a manually labeled dataset, also named training dataset, and various features extracted from the Twitter messages in order to submit a test dataset under the model and have an automatic prediction regarding the polarity of the sentiment (positive, negative or neutral) in the message. There are many types of classifiers, such as Naïve Bayes (NB), Maximum Entropy (ME), Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF) or even Conditional Random Fields (CRF). In the last few years, many studies of Twitter sentiment analysis using supervised learning were conducted. At 2009, Go et al. [GBH09] tackle a problem of binary classification about Twitter sentiment analysis using SVM, NB and ME algorithms and distant supervision to produce their classifier models. Their dataset was composed by 1.6 million twitter messages and they don't have the problem of imbalanced classes. As features they used POS tags, unigrams and bigrams and they also have in consideration the existence of negation in the messages. The best performance they have obtained was with the Naïve Bayes model and as final conclusions they said that using POS tags as feature doesn't improve the final accuracy. Hamdan et al. [HBB13] made some experiments using SVM and NB models in Twitter messages but their set of features was different from the previous work mentioned. The authors use DBPedia to extract concepts, WordNet to extract adjectives, Senti-WordNet to extract the sentiment score of some words and also have in consideration the existence of emoticons in the message. As final results they concluded that the SVM model performance surpasses the NB model between 2%-4%, and for that they have used the harmonic mean F-measure as evaluation metric. Saleiro et al. [SGS16] work with Twitter messages to study the political opinions about the five political leaders during the Portuguese bailout between 2011 and 2014. The features they have used were composed by sentiment aggregate functions applying them to a non-linear regression model using the Random Forests algorithm and to a linear regression model using the Ordinary Least Squares (OLS) algorithm. The authors have grouped the dataset per month in order to see what was the monthly variation. In the validation process, a 10-fold cross validation was used and regarding to the evaluation metric they explore the Mean Absolute Error (MAE).

The Classifiers Ensembles approach is based on the combination of several classifiers to improve the performance of the classification task. The workflow of this approach can be verified in the figure 3.2. da Silva et al. [dSHJ14] have explored the sentiment analysis in Twitter using a combination between Random Forests, Support Vector Machines, Multinomial Naïve Bayes and Linear Regression. The final classification decision in this kind of approach is usually made by majority voting between the models. In their work, da Silva et al. decided to calculate the average

Figure 3.2: The workflow of the classifiers ensembles by da Silva et al. [dSHJ14]



between all classification probabilities given by the models and applied that resulting value to the decision task.

Hassan et al. [HAZ13] have explored a different approach regarding the classifiers ensemble. They proposed a framework that was composed by seven different classification algorithms and used bootstrapping to the sample input in order to provide a small portion to each model. Their set of features had several types: semantic, POS tags, sentiment scores (SentiWordNet) and n-grams. Information Gain criteria was used to the feature selection process. By using bootstrapping in their framework, the problem of imbalance classes was reduced which could be a great advantage to explore in our approach.

Deep learning (DL) is the third and last method in the supervised learning approaches. DL is a recent field in the area of Machine Learning which may imply a scarcity in its addressed use to the sentiment analysis on Twitter [GC16]. The studies conducted in this method took advantages from the SemEval-2013 and SemEval-2014 datasets. Tang et al. [TWY+14] developed three different neural networks models to learn sentiment specific word embeddings (SSWE). The results provide by the models were later used as features to classify the polarity of sentiment in the dataset messages. The authors have combined the obtained features with others like sentiment lexicons [TWY+14], emoticons, negation, n-grams, punctuation, clusters, etc [TWQ+14]. The evaluation metric used to measure the performance of their classification models was the F-measure. The best result obtained to the SemEval-2013 dataset was 86.58% while for the SemEval-2014 dataset was 87.61%.

There are a high diversity in the methods to apply supervised learning, i.e. application of machine-learning algorithms, to automatic classify the sentiment polarity either on tweets or opinion reviews. The problem in its use focuses on the features engineering which is a hard task and the learning algorithms effectiveness depends on its selection. The bad choice of some features may cause that the final results obtained are not the most desirables.

Contrary to the machine-learning approaches, the **lexicon-based approaches** doesn't depend on training data and features to classify the sentiment as positive, negative or neutral. In this

approach the final sentiment classification is given by measuring the sentiment score of each term using external resources, such as dictionaries with a large number of previously evaluated terms (SentiWordNet [9], SenticNet [10], LIWC [11]). At 2010, Thelwall et al. [TBP+10] developed a lexicon-based algorithm, named SentiStrength [12], capable of detecting the sentiment value in messages that usually have informal language, such as tweets. The algorithm have access to 298 positive terms and 465 negatives and to a list of emoticons, negations and booster words to increase or decrease the sentiment value of derived words. The authors have compared their algorithm with machine-learning approaches and the results were very interesting since in terms of accuracy as evaluation metric, SentiStrength has surpass the others.

C. Musto et al. [MSLG15] have developed a domain-agnostic framework to produce some social media analytics regarding some events that happen in Italy in the last years. They evaluate the sentiment present in each tweet using lexicon-based approaches. The external resources used were SenticNet and SentiWordNet. The authors have split each tweet message by cues, such as punctuations and conjunctions, creating two or more micro-phrases. After this step, each micro-phrase is classified according the scores of the terms present in the resources. The sentiment polarity of the original message is obtained by summing its related micro-phrases. They also studied an emphasized approach where the Part-of-speech (POS) category of each term has a weight. Adverbs, verbs and adjectives received a value greater than 1, while for the remaining categories the value was 1.

L. Allisio et al. [AMB+13] proposed a framework, named Felicittà, in order to measure the hapiness level in the Italian territory. The study was made on geotagged tweets and it was used the resources MultiWordNet [13] and WordNet-Affect [14]. All the emoticons presented in the tweets were replaced by its meaningful words. The approach used by the authors consists in for each tweet term, a search is computed in the MultiwordNet dictionary to find all the meanings the word can have. After this step, each meaning found is associated with the sentiment score present in the WordNet-Affect corpus. The sum of all meanings is calculated, assigning a value of -1, 0 or 1 to the term. The tweet final classification is done by calculating the mean polarity of all terms and comparing with a heuristic constant defined by the authors.

The lexicon-based approaches are simpler to implement compared with the machine-learning approaches. They also presented disadvantages as the need of a continuously update of the word lists (lexicon sentiment dictionaries) because the conversation themes on Twitter are always changing which may result in the absence of words in the lists, and consequently their scores [GC16]. For this reason, the missing words are not considerate to the sentiment polarity classification and the results may not be so reliables.

[9] http://sentiwordnet.isti.cnr.it/
[10] http://sentic.net/
[11] http://liwc.wpengine.com/
[12] http://sentistrength.wlv.ac.uk/
[13] http://multiwordnet.fbk.eu/english/home.php
[14] http://wndomains.fbk.eu/wnaffect.html

The last approach for sentiment analysis is a mixture of the two previously presented, a **hybrid approach**. This kind of approach was explored by Ghiassi et al. [GSZ13] where they used machine-learning algorithms (SVM and Dynamic Artificial Neural Networks - DAN2) with a n-gram analysis. The collection of tweets was about Justin Bieber and as features to the classifiers, the authors choose emoticons, tweets that have positive and negative words, e.g. *happy or sad*, and also synonyms of this words. The model DAN2 proved be the best in the classification task. A. Kumar et al. [KS12] mixed a log-linear regression model with lexicon-based methods. Firstly, they have made pre-processing to the tweets collection by removing the URL references, replacing emoticons with their score value, calculate the percentage of caps in the message and also the sentiment orientation of the adjectives, verbs and adverbs. The overall sentiment of the tweet message was computed by the linear equation of the model, which was enough to prove the efficiency of the approach explored by the author relatively to the polarity of a tweet.

The main advantage of the hybrid approach establishes in the no need to manually classify the dataset for its use in machine learning methods. By applying lexicon-based methods we can have a labeled dataset ready to be use in ML classifiers. A disadvantage on this approach is high computational power need to bear out both approaches at the same time [GC16].

3.5.2 Aspect-based Sentiment Analysis

Aspect-based sentiment analysis (ABSA) is the most difficult problem to solve regarding the field of sentiment analysis. This approach focuses on the recognition of aspects in the messages and consequently on their sentiment polarity classification.

In particular to the aspect extraction, an overview was already done in the subsection 3.4.3 where the topic-based approach was mentioned and described using the LDA model as the most used model. There are three more approaches that we can follow to discover relevant aspects in a document: frequency-based [HL04], ruled-based [GWS13] and supervised learning [JG10, JHS09].

The frequency-based approach focus on finding some nouns or noun phrases from a large corpus using the occurrence frequency as the main requirement. M. Hu et al. [HL04] used this approach in order to summarize some customer reviews regarding a set of products. They used POS tagging to found nouns and noun phrases present in the document and association mining to find frequent itemsets because the features that composed the itemset are usually product features. After this step, they submitted the features set to a pruning section in order to remove the meaningless ones.

In the ruled-based approach, S. Gindl et al. [GWS13] have made a study in order to prove that it's possible identify and extract aspects in sentences by propagating the sentiment charge to noun targets following a set of defined rules. They have verified a problem when the sentiment and the aspect mention are in different sentences. Simple propagation rules are not enough to found the aspects. To overcome this, they have defined another rule where if a sentence starts with a pronoun, the target aspect is probably the last noun identify in the previous sentence.

Supervised learning approaches are based in sequential learning methods, such as Conditional Random Fields (CRF) and Hidden Markov Models (HMM). The mentioned methods are similar because both of them attempt to discover patterns relative to an input data set. These types of methods are often used in the aspect extraction task of opinion mining. N. Jakob et al. [JG10] has built a classification model using the CRF algorithm in order to enhance the target of each opinion in the reviews. The domain of their dataset was constituted by four independently categories: movies, web-services, cars and cameras. Since the approach taken by the authors was through machine learning classifiers, they had the need of establishing a set of features to train the model. The features used for the classification vary from the string of each token, the POS tag for each token, the level of dependency between each token and the opinion expressed as well as its word distance, and, finally, the last feature is the opinion sentence itself to allow the CRF algorithm the ability of distinguish when a token is present or not in a sentence that is an opinion. Regarding the system validation, the authors applied also a 10-fold cross-validation to see if the model performance improves or not. As performance metrics to evaluate the system, they used the precision (Equation 6.1),

$$\frac{TP}{TP+FP} \tag{3.3}$$

the recall (Equation 6.2)

$$\frac{TP}{TP+FN} \tag{3.4}$$

and the F-measure (Equation 3.5) that is the harmonic mean between the previous two metrics.

$$2.\frac{precision.recall}{precision+recall} \tag{3.5}$$

W. Jin et al. [JHS09] also worked under the opinion reviews and tried to find relevant opinion targets in the content. They have developed a novel framework based on machine learning techniques. The framework, named OpinionMiner, appeals to a classification model with the HHM algorithm and to a bootstrapping approach in the training step. The bootstrapping divides the main process of training in two sub-process as well as the training dataset into little portions in a randomly way. Each sub-process has his own HHM model and after its training step, the main process only selects the objects which their label is agreed by both classifiers. The bootstrapping process is repeated until no more targets in the objects can be discovered.

Regarding the polarity sentiment classification, the majority of the works studied appeals to one of the approaches described in the section 3.5.1.

3.6 Conclusion

This chapter had the objective of review some basic concepts that may be relevant to contextualize the reader about the problem of performing analysis in social media streams, e.g. Twitter mes-

sages. Hence, the literature studied was divided in several points in order to have a overview about what is already done and what are the approaches that some author proposed to tackle each of the sub-problems that composed the main problem in this dissertation work. After a careful research, it was possible to identify that there are a great diversity of approaches to each sub-problem, whether it be disambiguation, filtering, topic detection or even sentiment analysis.

An important point identified in the literature was the few works done using deep learning to take the problem of sentiment analysis with supervised leaning approaches, since its applicability in the artificial intelligent field has grown at an exponential level in the recent years.

Regardless the task that the authors dealt with, it was possible to identify that the features engineering process and its selection, when their proposed solution used classifier models, is similar. This may be an advantage to the development of the different modules that composed the proposed framework in this dissertation work.

The framework modules will also have classifier models, so the evaluation and validation of it is important. The literature review shows a large set of evaluation metrics to do this step.

In short, it is expected from the reader that this review to the State-of-the-Art has provided a coherent understanding regarding the study of different real scenarios using the social media streams as source of information.

Chapter 4

Framework

In this chapter it is described the details and specificities of the framework proposed in this dissertation. First, we enunciate the necessary requirements to fulfill and achieve the mentioned development. Moreover, it is present the framework architecture design, as well as it inner pipeline. The modules that constitutes such architecture are described afterwards as so the required methodologies and algorithms incorporated in each of its tasks. Finally but not least, we mentioned and explained the different data visualizations available in the framework.

4.1 Requirements

The development of frameworks to the domain of *smart cities* and intelligent transportation systems using human-generated content (e.g. text messages) is a laborious and time-consuming process. The source of the data to fed such system is one of the biggest challenges in this kind of developments, ranging from social media, smart phones and urban sensors. In this dissertation we tackle the problem of exploring social media data since this kind of data have, recently, been seen as a new opportunity and source to mine valuable information to the cities services and corresponding responsible entities [MSLG15].

Social media data are mostly represented by text messages being necessary the application of Natural Language Processing (NLP) methodologies in order to extract information from its content. Such methodologies are usually complex and composed by several different steps (e.g. some related to the syntax of the sentences while others are related to the semantics of its content) before the achievement of the desired results. Social Media streams are no exception, indeed, the analysis of such texts is even more complex since messages are usually short and present lots of informal characteristics.

A framework for the domain of social media content requires, in the first place, a data collection module. Depending on the social network, the data collection module can have different heuristics with respect to the data retrieving. Here, the choice of such heuristics is important and needs to be made according the final users expectations, or at least, according the framework final use case. Towards the application of NLP techniques, a module in charge of preprocessing tasks is required. The main purpose of this module establishes in the performance and robustness of the results obtained by the previously mentioned techniques. NLP techniques can provide different types of information, however in this dissertation the focus is on the classification of travel-related tweets, characterization of the topic associated with a tweet and also travel-mode extraction. Each technique is represented as an independent module whose belongs to the boundary of text analytics. This framework needs to also be capable of processing information regarding the creation date of a tweet, *metadata* and geographic distribution associated to it. For the fast retrieving of this informations to the data visualization view, some aggregations need to be made. This requirement is due to one of the big data demands, the instantly availability of the results. Such demand is important for the framework end-users since it helps in the entities' decision-making process making easier and faster the improvement of its services.

4.2 Architecture Overview

4.3 Data Collection

In Section 4.1, we explain the importance of the decision made to the data collection's heuristics. Twitter allows the developers' community two different tools to collect data, the Search and the Streaming APIs. The Search API is based on the RESTful protocol and only looks up for tweets published in the last 7 days, while the Streaming API creates basic endpoints (independent of the REST protocol) and retrieves up to 1% of the Twitter Firehose [1]. Regarding the proposed and developed framework, we chose the Streaming API due to its free-access for the community and smooth integration in the module implementation. A positive point about the Streaming API is the three available heuristics to the data collection, allowing the retrieval of tweets that match a specific text query (e.g. tweets with the word `bus` or `car`), the retrieval of tweets associated to a variable number of users - being necessary previous knowledge about these users *ids* - or even

[1] Twitter Firehose - is a paid Twitter service that guarantees the delivery of 100% of the tweets matched with certain criteria.

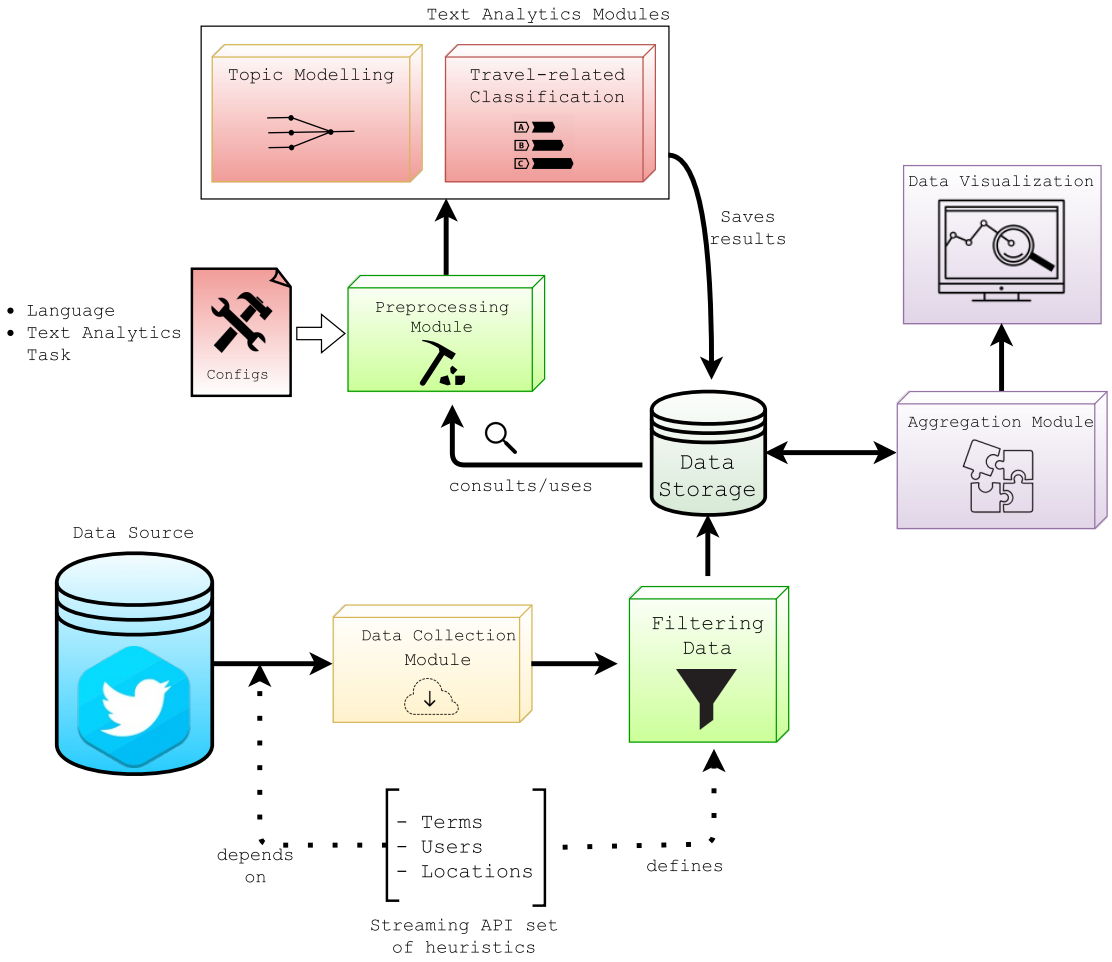


Figure 4.1: Architecture overview of the proposed framework

the retrieval of tweets located inside a bounding-box [MKWP+16]. There are two negative points regarding the Twitter Streaming API: first, Twitter imposes limits in its data exploration, where only 400 words can be tracked, 5,000 users can be followed and 25 different bounding-boxes can be explored[2]; second, the previously mentioned heuristics cannot be used together, i.e. we can not track specific tweets from an user that match with certain words. Although the negative points, we remain with the choice made, of using the Twitter Streaming API as our source of information and limiting the heuristic to the one that retrieves tweets located inside a bounding-box. Our choice is additionally supported by the need of studying cities and exploring the information derived from it. This way, we know, a priori, that if the data collection method is able to retrieve tweets with precise geo-location then this makes our work easier since the exploration of specific regions of a city is already available taking into consideration the information available in tweets.

After the method selection, as well as the selection of its heuristic, we conduct an experiment regarding the amount of tweets being retrieved by one Twitter client for a city. Twitter has into consideration the number of clients used in the data collection process by tracking the IP address of the machine in the network. This constitutes a restriction to explore several cities with the same client since the Streaming API retrieves only 1% of the total overcome. In the experiment, we tested the capacity of a client to retrieve all the tweets posted in New York City and used four different clients for it: one defined with the city bounding-box, and the other three defined with bounding-boxes of three boroughs in the city: Bronx, Brooklyn and Manhattan. Considering the bounding-boxes creation, we took support of an open-source *online* tool coined BoundingBox [3], which is integrated with the Google Maps API.

Results showed that the client defined with the greatest bounding-box, New York City, was able to retrieve 100% of the tweets from the three different boroughs. This experiment is consolidated with the work of F. Morstatter et al. [MPLC13] where it was compared the Streaming API's capacity, regarding geo-located tweets, against the Twitter Firehose. Authors concluded that the percentage of geo-located tweets corresponds to 1-2% of total overcome from Twitter and the Streaming API is able to retrieve almost 90% of it. Hence, we do not need to be concerned about how many bounding-boxes are used in the collection process because if we did we would need to be aware of 90% of the world, which is not the case.

4.4 Data Preprocessing

The extraction of information from text, in particular from social media streams, is an iterative process and requires a segmented and planned pipeline to achieve the final results. In the requirements section (4.1), we mentioned some problems of social media streams as the short length and informality of the text message. The informality problem ranges from the writing style of each

[2] https://dev.twitter.com/streaming/reference/post/statuses/filter (Accessed on 18/06/2017)

[3] http://boundingbox.klokantech.com/ (Accessed on 23/06/2017)

person to the existence of lots of abbreviations, slang, jargons, *emoticons* and bad usage of punctuation signs. The preprocessing module presented in this section has as main goal the submission of the text messages under several operations in order to remove, or at least, reduce this type of informality characteristics and make easier the work of future tasks.

Below, we enumerate and described the different preprocessing methods implemented:

- **Lower casing:** This operation is responsible for the conversion upper case characters to lower representation. The advantages provided by this operation are centered in the analysis of words written in different ways. An representative example is `london` and `London` whose meaning is the same but due to the different case in one letter, its representation/interpretation by text mining techniques may be disparate.

- **Tokenization:** Is the method of dividing each sentence in a list of tokens/words. Since we are dealing with social media content, standard tokenizations techniques available in packages, such as the `tokenize` [4] of NLTK Toolkit for Python, perform poorly and are not capable of dealing with *#hashtags*, *@mentions*, abbreviations, strings of punctuation, *emoticons* and unicode glyphs which are very common in Twitter. Having considered this, we used a Twitter-based tokenization package, coined Twokenize and firstly presented by B. O'Connor et al. [OKA10], which is capable of dealing with these special characteristics of tweets.

- **Punctuation Removal:** Depending on the future task, all signs of punctuation are removed. In this case, every *emoticon* was removed, as well as the symbols # and @ which composed the *hashtags* and user mentions.

- **User mentions and *URLs* Removal:** Following the condition of the above mentioned operation, the removal from the text of this type of content depends of the current task.

- **Stop words Removal:** This operation consists in the removing of the most common words in the language in analysis. We used the standard words of the NLTK Corpus package.

Regarding other fields in a tweet, this module was also in charge of convert the date of creation of a tweet to the city timezone. The field *created_at* in a tweet is given in the Coordinated Universal Timezone (UTC) and in order to have knowledge about the most active local hours and days on Twitter, we used the Python timezone package `pytz` to convert the world timezone to the one desired.

Although the existence of more text preprocessing techniques, in this dissertation we only used the ones previously described since each of them is associated to, at least, one text analytics module whose are described in the following section.

[4] http://www.nltk.org/api/nltk.tokenize.html

4.5 Text Analytics

The extraction of information from texts can vary in several types depending on the task performed to achieve it. In this dissertation, it was developed different types of analysis having in consideration the text messages.

4.5.1 Travel-related Classification

Prima facie, we tried to extract and characterize travel-related tweets from large datasets in order to study the geographical and temporal distributions of such specific content. To be successful in this task we create an automatic text classifier capable of discriminating travel-related tweets from non-related ones. Due to the absence of gold standard datasets in this domain, there was the need of creating a training and testing set of data in order to proceed the experiment and evaluate the performance of the obtained model. Conventional classification tasks in the domain of intelligent transportation systems follow traditional approaches by constructing their group of features using standard bag-of-words techniques. In our experiment, we tried to combine a bag-of-words technique with word embeddings methodologies, producing, for the best of our knowledge, the first travel-related classification model with both type of features.

The word embeddings technique is used by T. Mikolov et al. [MCCD13] in the implementation of a powerful computational method named *word2vec*. This method is capable of learning distributed representations of words, and each word is represented by a distribution of weights across a fixed number of dimensions. Authors have also proved that such representation is robust when encoding syntactic and semantic similarities in the embedding space.

The training objective of the skip-gram model, as defined by T. Mikolov et al. [MYZ13], is to learn the target word representation, maximizing the prediction of its surrounding words given a predefined context window. For instance, to the word w_t, present in a vocabulary, the objective is to maximize the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log P(w_{t+j}|w_t) \tag{4.1}$$

where c is the size of the context window, T is the total number of words in the vocabulary and w_{t+j} is a word in the context window of w_t. After training, a low dimensionality embedding matrix $\mathbf{E}$ encapsulates information about each word in the vocabulary and its use (i.e. the surrounding contexts). For instance, by using the skip-gram model over our datasets we were able to verify that words such as `ônibus` and `busão` are used in the similar contexts, as a mode of transport.

Later on, Q. Le and T. Mikolov [LM14] developed paragraph2vec, an unsupervised learning algorithm operating on pieces of text not necessarily of the same length. The model is similar to *word2vec* but learns distributed representations of sentences, paragraphs or even whole documents instead of words. We used *paragraph2vec* to learn the vector representations of each tweet and tried several configurations in the model hyper-parameterization.

The previous described methods are available in the collection of Python scripts we used in this dissertation, coined `Gensim` [5], presented and lately improved by R. Řehůřek and P. Sojka [RS10].

The overall experiment regarding the travel-related classification of tweets is described and detailed in Section 6.2. Concluded the experiment, we select the best classifier and used it the implementation of the travel-related module allowing the framework to discriminate potential new tweets related to the transportation domain.

4.5.2 Topic Modelling

Further developments towards the enrichment of different information provided by the framework took us to the path of topic modelling techniques for text messages. Topic modelling is a text mining technique which goal is the identification of latent topics in a collection of documents. During the last decade, the research community had been using this technique in a vast range of works aiming the test of its applicability in different domains. Here, we also used topic modelling to characterize the different cities and provide this type of information to the framework's end-users.

Latent Dirichlet Allocation (LDA) is a generative statistical model proposed by D. Blei et al. [BNJ03] that makes possible the discovering of unknown groups and its similarities over a collection of text documents. The model tries to identify what topics are present in a document by observing all the words that composing it, producing as final result a topic distribution.

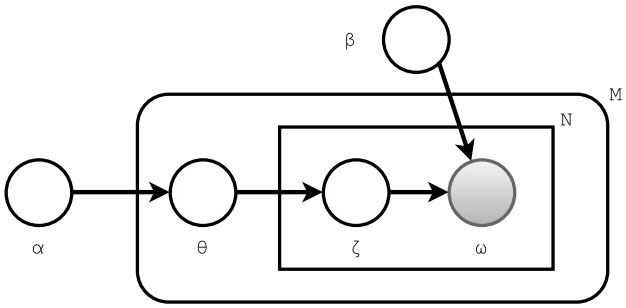


Figure 4.2: Plate Notation of the graphical model representation of Latent Dirichlet Allocation by D.Blei et al. [BNJ03]

In Figure 4.2 it is illustrated the plate notation to the graphical model of LDA. There, we can observe that for a collection of documents M, each one composed by a sequence of N words, the model tries to attribute a per-document topic distribution, using an α dirichlet prior, to a topic-word distribution ξ (associated also with a dirichlet prior β), inducing that each topic's probability θ is focused in a small set of words w which characterize that topic.

The most important advantage this model provides is related to the group of features involved in its training process. Conventional application of this model uses only as features a bag-of-words matrix representation [6], and for this reason the task of topic modelling becomes very simple since

[5] `https://radimrehurek.com/gensim/about.html` (Accessed on 20/06/2017)

[6] Bag-of-words representation matrix is a list of lists, where each entry of the matrix is associated to a sentence of the document and takes the form of a term-frequency vector.

the the frequency of words in the documents are taken into account. Last but not least, LDA model performs two different distributions: (1) distribution of words over topics and (2) distribution of topics over the documents, resulting in the assumption that each document is random mixture of topics, whose in turn are composed by a probabilistic distribution of words.

The cities' characterization provided by our framework centers in the topics being talked about at the time. We conduct an experiment to evaluate if such information could bring added-value for the cities entities and the results although being very promiscuous proved to have potential in certain occasions. The overall experiment is described in Section 6.1 as well as potential improvements to the generated model.

4.5.3 Final Remarks

The previous mentioned text analytics methodologies were implemented as separate modules in the framework since each of them needs different preprocessing operations over the data. A future interesting improvement to the framework, presented in this dissertation, is the incorporation of an extra module of sentiment analysis that should work together with the two already developed, and provide additional information about the services of a smart city, including the transportation domain.

4.6 Data Storage and Aggregation

Besides the few percentage of geo-located tweets provided by Twitter (1-2% of the total Firehose overcome), this data requires, in the first place, large physical space for storage and, secondly, a tool that allows the easy manipulation and quick access of data. Having considered this, we opted for the use MongoDB, an open-source cross-platform document-oriented database, as the database system for our framework. MongoDB allows the storage of JSON-like documents which is the retrieved format of tweets by the Streaming API. Since in this dissertation we developed the framework as a prototype of a system capable of extracting information related to *smart cities* and transportation services, the large physical space to storage data was not a priority.

MongoDB presents, alongside the high performance, availability and scaling, an inner framework that allows the aggregation of data according to specific user-generated queries. Here, we took advantage of such a pipeline in order to produce interesting statistics regarding the processed data. Map-reduce is the processing paradigm behind the aggregating operations allowing high performance even when applied to large volumes of data, as in this particular case where it is necessary to process thousands or millions of tweets in a short period of time.

4.7 Visualization

One of the most laborious and time-consuming tasks in the development of this social media based framework was the selection of data visualizations to illustrate the results provided by the previous

mentioned modules. Due to the amount of data being processed, the generation of data visualization using an atomic implementation is sometimes poorly in terms of response time. Hence, we needed to adopt a different approach in order to solve this non-efficient procedure.

After a long period of research, we found a solution to this problem by creating a set of routines (bash scripts) that are called periodically (hourly) to execute all type of necessary aggregations and update its corresponding data collections in the database. Then, other routine is invoked to generate all type of data visualizations and store its visual representation in HTML files. In the implementation of this module, these files - containing the data visualization - were embedded inside several view pages. `Plotly` [7] is a Python graphing library that has available the saving of the visualizations produced in files with HTML format. Besides that, the library offers an extensive range of graphical representations, such as basic charts (bar charts, scatter plots, etc), scientific charts (heatmaps), financial charts (time series) and maps (choropleth, bubble and line maps), which facilitates the construction and designing of dynamic dashboards. Here, we explore mostly the section of basic charts to build simple representations of the results obtained from the analytics phase and also added top lists about some metadata of the tweets, as so the overall, daily and hourly top *hashtags* and uni-grams.

4.8 Summary

[7] https://plot.ly/python/

Chapter 5

Exploratory Data Analysis

The main goal of this chapter is the devise of relevant analysis taking into consideration the five different collected datasets. Since this dissertation is supported in experiments using real-world data, such analysis is crucial in order to gain better knowledge of the intrinsic characteristics of it. A tweet provides some fields of interest, such as, the text message, date of creation, language, and the *entities*, which are constantly analysed in several data analytics systems. An *entity* is metadata and additional contextual information contained in the tweet and is composed by the *hashtags*, *user mentions*, *urls* and *media* fields. We count the amount of tweets containing this kind of information for all the cities, London, New York, Melbourne, Rio de Janeiro and São Paulo, and projected some data visualizations for different temporal frequencies. The following subsections are divided into three different categories: (1) Geographical Distribution, (2) Temporal Frequencies and (3) Metadata Composition. Additionally, we discuss the results of each city, as well as the main observable differences.

5.1 Geographic Distributions

As previously mentioned, in Section 4.3, we exploit an auxiliary *online* tool to generate the co-ordinates for the bounding-boxes used in the collection process. The visual representation of the each city bounding-box is illustrated in Figure 5.1, as well as its the corresponding coordinates which are presented in Table 5.1.

Table 5.1: Collecting Bounding-boxes Coordinates (South-West and North-East)

City	South-West	North-East
Rio de Janeiro	(-43.7950599, -23.0822288)	(-43.0969042, -22.7460327)
São Paulo	(-46.825514, -24.0082209)	(-46.3650844, -23.3566039)
New York City	(-74.2590899, 40.4773991)	(-73.7002721, 40.9175771)
London	(-0.3514683, 51.3849401)	(0.148271, 51.6723432)
Melbourne	(144.5937418, -38.4338593)	(145.5125288, -37.5112737)

(a) São Paulo (b) London (c) Melbourne

(d) Rio de Janeiro (e) New York

Figure 5.1: Search Bounding-boxes for the data collection

Taking a careful observation into to coordinates used to each bounding-box, we can affirm that Rio de Janeiro present the broadest bounding-box comparatively to the others cities.

In the first attempts to study the geographic distribution in our datasets, we discover that not all tweets had a precise coordinate attached to it. Nonetheless, there were cases where tweets from other cities were collected to our datasets and this phenomenon is not supposed to happen when the collection method is based in geo-located characteristics. By studying the Twitter mobile application, we found out that a user can tag himself in the tweet by two different ways: (1) a user can activate the GPS in the mobile application and associate to the tweet his precisely geo-location; (2) a user can choose a place from a predefined list provide by Twitter and associate the place to the tweet.

The second method of tagging the geo-location to the tweet can arise some conflicts when this kind of tweets is used to perform scientific studies or even development of system to help the cities in the regularization, control and improvement of its services. Having this in consideration, it was necessary to understand how the Twitter Streaming API works and what kind of heuristics follows in order to retrieve this type of tweets. Hence, the documentation [1] enhances two different heuristics:

1. If the coordinates field is populated, the values there will be tested against the bounding-box;
2. If the coordinates field is empty but place is populated, the region defined in place is checked for intersections against the locations bounding-box. Any overlapping areas will yield a

[1] https://dev.twitter.com/streaming/overview/request-parameters#locations (last visited on 17 June, 2017)

positive match.

The first heuristic only happens if a user is able/willing to tag a post with his precise geo-location associated with it; otherwise, the user can tag the post associated with a place and in this case the second heuristic is applied. Each place contained in the previous mentioned list, which is provided by Twitter, is composed by a bounding-box, and if any piece of it overlaps the bounding-box used in the collecting process, then a positive match is yielded and the tweet is retrieved. For example, if a tweet has a place such as Brazil and our filter bounding-box is defined for Rio de Janeiro, all tweets from place Brazil will be in our dataset, regardless the fact some tweets are posted elsewhere, such as in the city of Manaus, very far away from Rio de Janeiro.

This restriction required the development of a external layer which was responsible for the filter of tweets located outside the area of each city. To built this so, it was necessary *a posteriori* information and, thus, we extract the Twitter default bounding-box of each city appealing to the tweets *place* field. Such information was then used as the limit area in order to filter out tweets which *coordinates* field was not populated. These bounding-boxes, the Twitter default ones, are listed in Table 5.2 and its corresponding visualization is the biggest rectangle demonstrated in Figures 5.2 (subfigures 5.2b and 5.2a) and 5.3 (subfigures 5.3a, 5.3b and 5.3c).

Table 5.2: Twitter Default Bounding-boxes Coordinates (South-West and North-East)

City	South-West	North-East
Rio de Janeiro	(-43.795449, -23.08302)	(-43.087707, -22.739823)
São Paulo	(-46.826039, -24.008814)	(-46.365052, -23.356792)
New York City	(-74.255641, 40.495865)	(-73.699793, 40.91533)
London	(-0.510365, 51.286702)	(0.334043, 51.691824)
Melbourne	(144.593742, -38.433859)	(145.512529, -37.511274)

The final volume of tweets located inside and outside the cities correspondent bounding-boxes are presented in Table 5.3. Alongside with the location analysis, the language count was also performed since future experiments only took into consideration tweets with the native language of the city in study and not foreign ones. In the abovementioned table (5.3) it is possible to verify a vast difference regarding the activity on Twitter in Rio de Janeiro. Numbers tell that such activity, with respect to geo-located tweets, is almost two times more than São Paulo, four times London and twenty five times Melbourne. A possible justification for this noticeable difference may be associated to the area of the bounding-box used in the collection process, but, on the other hand, according to some sources related to the demographic measures, for the case Rio De Janeiro *versus* São Paulo, the population volume has an opposite behavior, where São Paulo [2] has almost 12 millions habitants while Rio de Janeiro [3] has 6 million. Having only this amount of information it is impossible, at the moment, formulate a explanation to this phenomenon.

[2] https://cidades.ibge.gov.br/v4/brasil/sp/sao-paulo/panorama (last visited on 17 June, 2017)

[3] https://cidades.ibge.gov.br/v4/brasil/rj/rio-de-janeiro/panorama (last visited on 17 June, 2017)

Table 5.3: Datasets composition after verification of the tweets inside the corresponding bounding-box

City	All	PT/EN		Non-PT/EN		In Bounding-Box		Out Bounding-Box		PT/EN and In Bounding-Box	
		No. tweets	%	No. tweets	%	No. tweets	%	No. tweets	%	No. tweets	%
Rio de Janeiro	18,803,774	15,906,680	84,59%	2,897,094	15,41%	12,976,048	69,01%	5,827,726	30,99%	11,060,136	58,82%
São Paulo	9,319,624	7,203,115	77,29%	2,116,509	22,71%	6,237,427	66,93%	3,082,197	33,07%	4,886,626	52,43%
New York City	8,507,145	7,260,829	85,35%	1,246,316	14,65%	6,972,312	81,96%	1,534,833	18,04%	5,956,355	70,02%
London	5,596,551	4,774,310	85,31%	822,241	14,69%	4,752,918	84,93%	843,633	15,07%	4,040,092	72,19%
Melbourne	789,927	669,435	84,75%	120,492	15,25%	742,946	94,05%	46,981	5,95%	629,424	79,68%

Later, after the filtering process, we tried to understand the volume, as well as the location of each tweet. Through this kind of analysis it was possible to find out that a tweet which *coordinates* field was empty and is, actually, represented with a bounding-box, can also be a specific place, i.e. a place that has a precise coordinate. Not all places were represented by a bounding-box in which each point that composed it are different. An example to that is `Estádio do Maracanã` which although being represented by a bounding-box, all four points are equal. A division was made considering this three types of location - (1) bounding-box with four different points; (2) bounding-box with four equal points; (3) precise coordinate - in order to have a perception of how different specific places and bounding-boxes as so which is the volume of tweets that are related to it.

Table 5.4: Volume of tweets for each type of geo-location

City	Total	Bounding-boxes			Specific Places			Precisely		
		Distinct	No. Tweets	Percentage (%)	Distinct	No. Tweets	Percentage (%)	Distinct	No. Tweets	Percentage (%)
Rio de Janeiro	11060136	297	10237280	92,56%	11159	49440	0,45%	163748	773416	6,99%
São Paulo	4886626	325	4284795	87,68%	7189	21022	0,43%	100028	580809	11,89%
New York City	5956355	328	4210854	70,70%	16078	85204	1,43%	138123	1660297	27,87%
London	4040092	53	3196043	79,11%	8123	53412	1,32%	95317	790637	19,57%
Melbourne	629424	22	523870	83,23%	0	0	0,00%	21826	105554	16,77%

The final counts of the analysis for each identified type of geo-location are presented in Table 5.4. Looking at the numbers it is possible to conclude some facts applicable to all cities. Citizens tend to geo-locate themselves with a location which has variable bounding-box size since more than 70% of the tweets are of this type. Furthermore, only a few percentage of tweets, between 0% and 1.43%, are located in specific places, although the existence of a higher number of distinct specific places comparatively to the bounding-boxes with variable size, with exception of Melbourne that has zero specific places in our dataset. Other interesting point to enhance is the considerable percentage of tweets with precise location (i.e. tweets that people tagged himself using the GPS). The Brazilian cities proved to be less supportive of precisely located tweets, while the English cities were more contributive. The distribution of each type of geo-located tweet is illustrated in Figures 5.2 and 5.3. The variable bounding-boxes are showed in 5.2a, 5.2b, 5.3a, 5.3b and 5.3c proving that our filter method was able to correctly agglomerate places that were, indeed, inside of the Twitter default bounding-boxes. In 5.2c, 5.2d, 5.3d, 5.3e and 5.3f is illustrated the distribution of the specific places found out in our datasets for each city. A particular point identified was the absence of specific places in Melbourne and the limited places in a certain area of London. With a first look at the image of London, there may be doubts about the results concern-

ing the filter method, however the bounding-box used to that process was the same in both cases, and so the only viable explanation for such result is the absence of specific locations for that area in the predefined list of places provided by the Twitter applications. Lastly, in 5.2e, 5.2f, 5.3g, 5.3h and 5.3i is illustrated the distribution of precisely located tweets. Through a careful observation in this distribution it was possible the arising of another doubt relatively to the first aforementioned heuristic of the Twitter Streaming API. There were tweets retrieved that not matched the bounding-box used in the collection process and this fact conducts to uncertainty and mistrust regarding the performance of this type of collection available on Twitter.

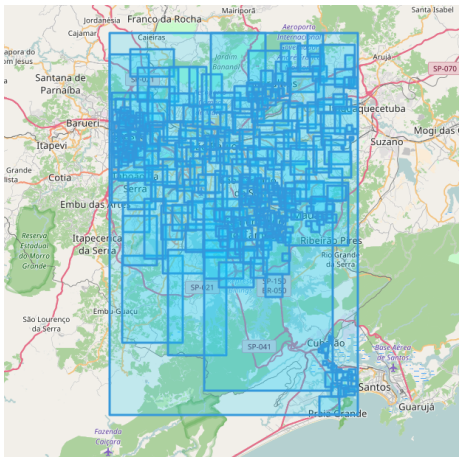
5.2 Temporal Frequencies

Another interesting analysis in our datasets concerns the temporal distribution of the data. The volume of tweets posted per hour, per day, as well as the activity by day-of-the-week or hour-of-the-day are statistics that enable the possibility of finding out patterns or variations which can be correlated to some events or incidents happening in a city.

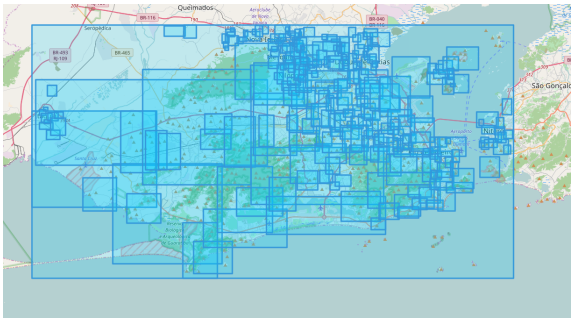
During and after remarkable events, citizens are impelled to share their feelings, opinions or even report their safety and well-being conditions (e.g. in cases of terrorist attack) through mobile applications. This share of information increases the activity of social media platforms, which can be potentially used for the identification of uncommon events. Figure 5.4 illustrates the daily distribution of all cities for the period of collection, three whole months, between 12 March and 12 June, 2017. The Brazilian cities present high level of variation between consecutive days (with the volume varying in a tens of thousands of tweets) and so the task of identifying remarkable events turns out to be much harder. On the other hand, the English speaking cities in our study are very similar, with exception of Melbourne whose activity is very low comparatively to the other cities (New York City and London). In the particular case of London, we can identify an abrupt increase of volume during days 8 and 9 of June. With the support of external sources such as news websites, we learnt about the United Kingdom General Elections 2017 [4] occurred on that period which suggests that an increase of the Twitter activity might be associated with that event.

In order to understand the most active days and hours in Twitter, for all cities under this study, we aggregate the datasets by these attributes and represented the final results in a box plot representation. This type of data visualization allows, in a standardized way, the displaying of distributions of data based on the six different values: (1) minimum and (2) maximum values for each day/hour regarding the activity on Twitter; (3) median value for the each day/hour, (4) first and (5) third quartiles as well as (6) the interquartile range (IQR). Figures 5.5 and 5.6 illustrated this type of data visualization for the whole three months of data collected. Taking into analysis the city of Rio de Janeiro, it was possible to observe and enhance Tuesdays as the day of the week where there is more activity on Twitter. Moreover, Fridays revealed to be the day less active, not only for the city of Rio de Janeiro, but for all remaining cities with exception of Melbourne. Particularly, the activity on Twitter in Melbourne is centered in the weekend days while the other cities the highest

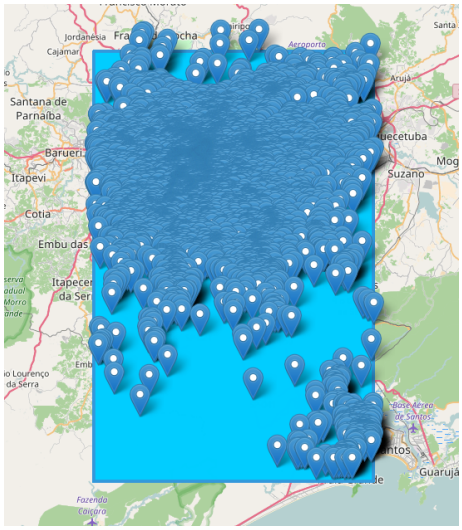
[4] https://www.theguardian.com/politics/general-election-2017 (Accessed on 17/06/2017)



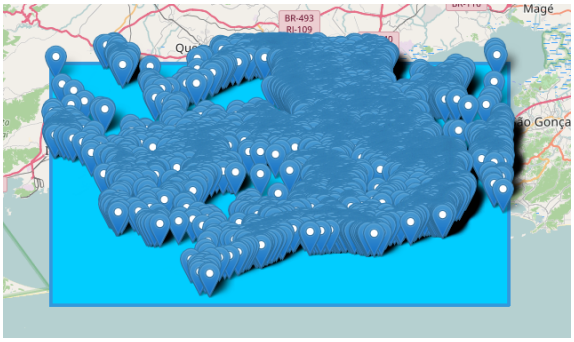
(a)



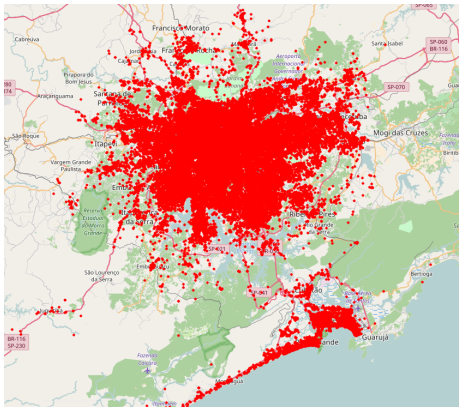
(b)



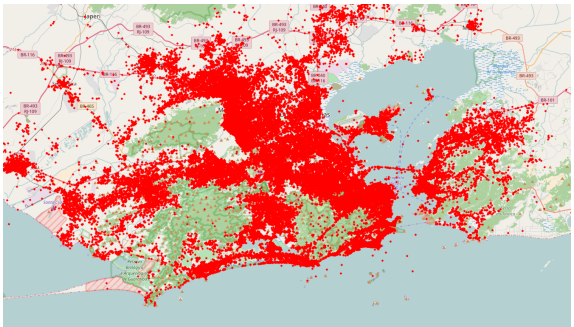
(c)



(d)

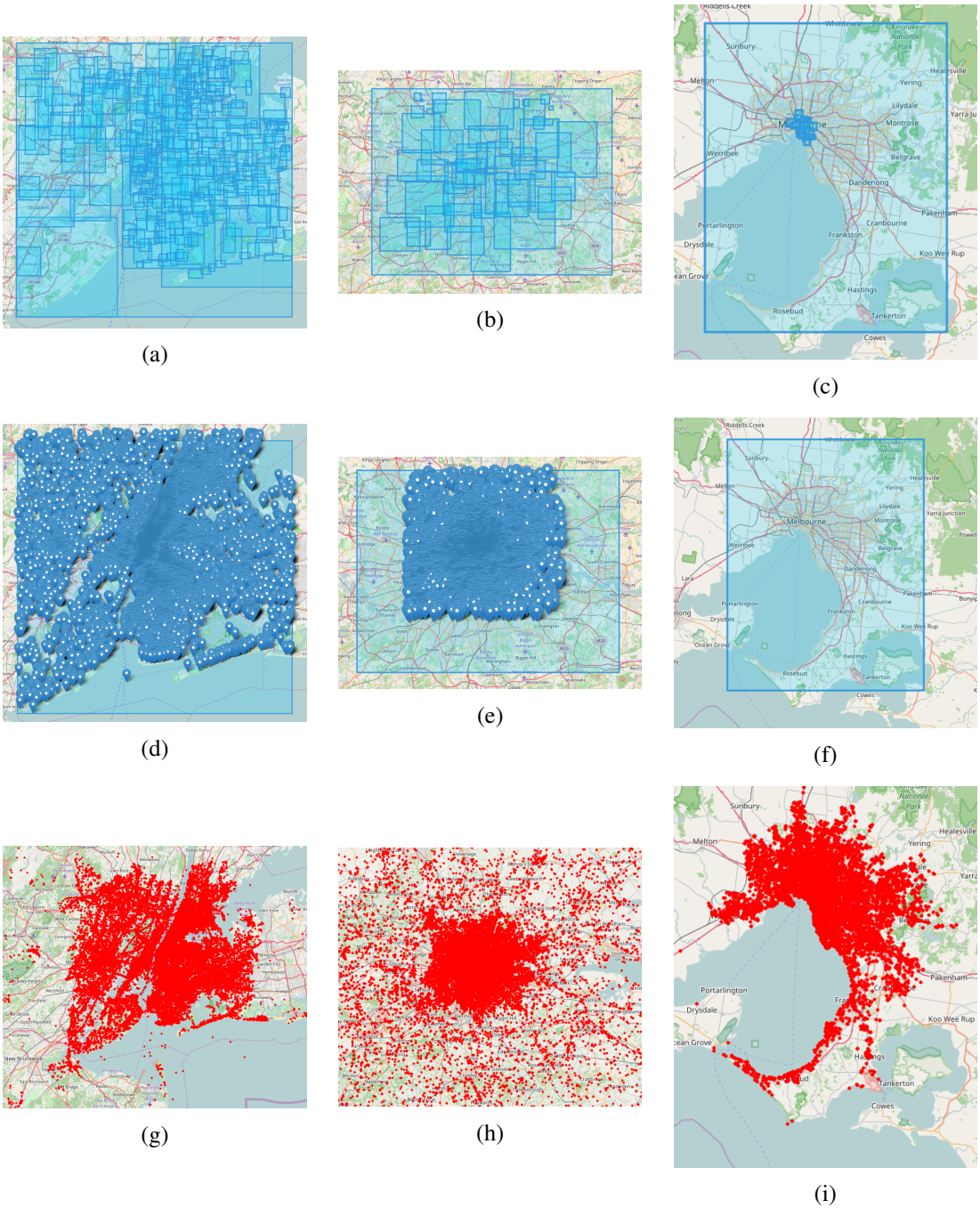


(e)



(f)

Figure 5.2: São Paulo (a, c, e) and Rio de Janeiro (b, d, f) Geographical Distributions: (a, b) Bounding-boxes of places (c, d) Specific places (e, f) Geo-tagged tweets

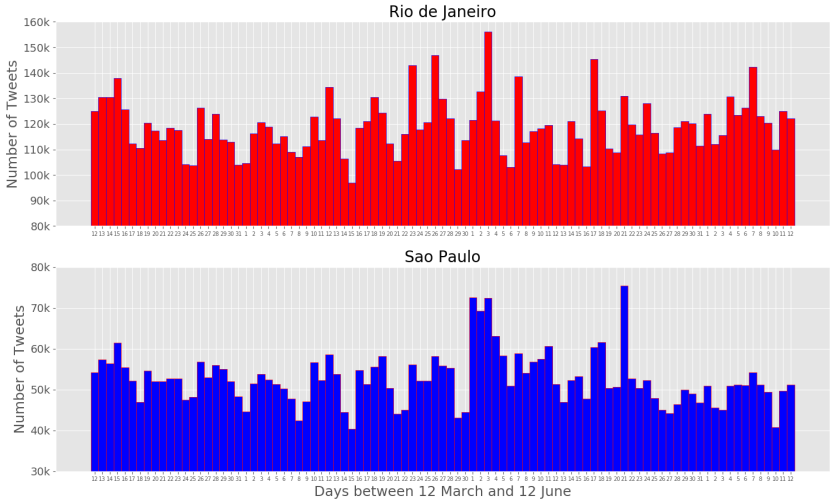


(a) (b) (c)

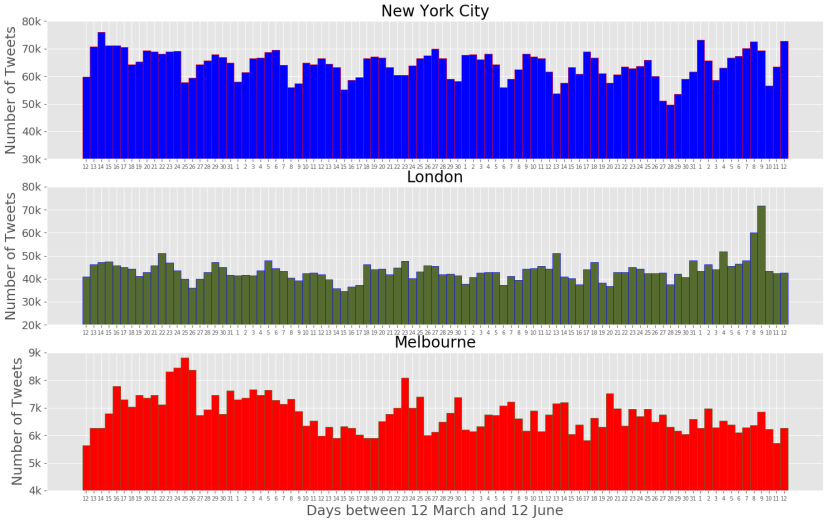
(d) (e) (f)

(g) (h) (i)

Figure 5.3: New York City (a, d, g), London (b, e, h) Geographical Distributions: (a, b) Bounding-boxes of places (c, d) Specific places (e, f) Geo-tagged tweets



(a)



(b)

Figure 5.4: Daily volume of tweets (a) Rio de Janeiro and São Paulo - Portuguese Cities (b) New York City, London and Melbourne - English Cities

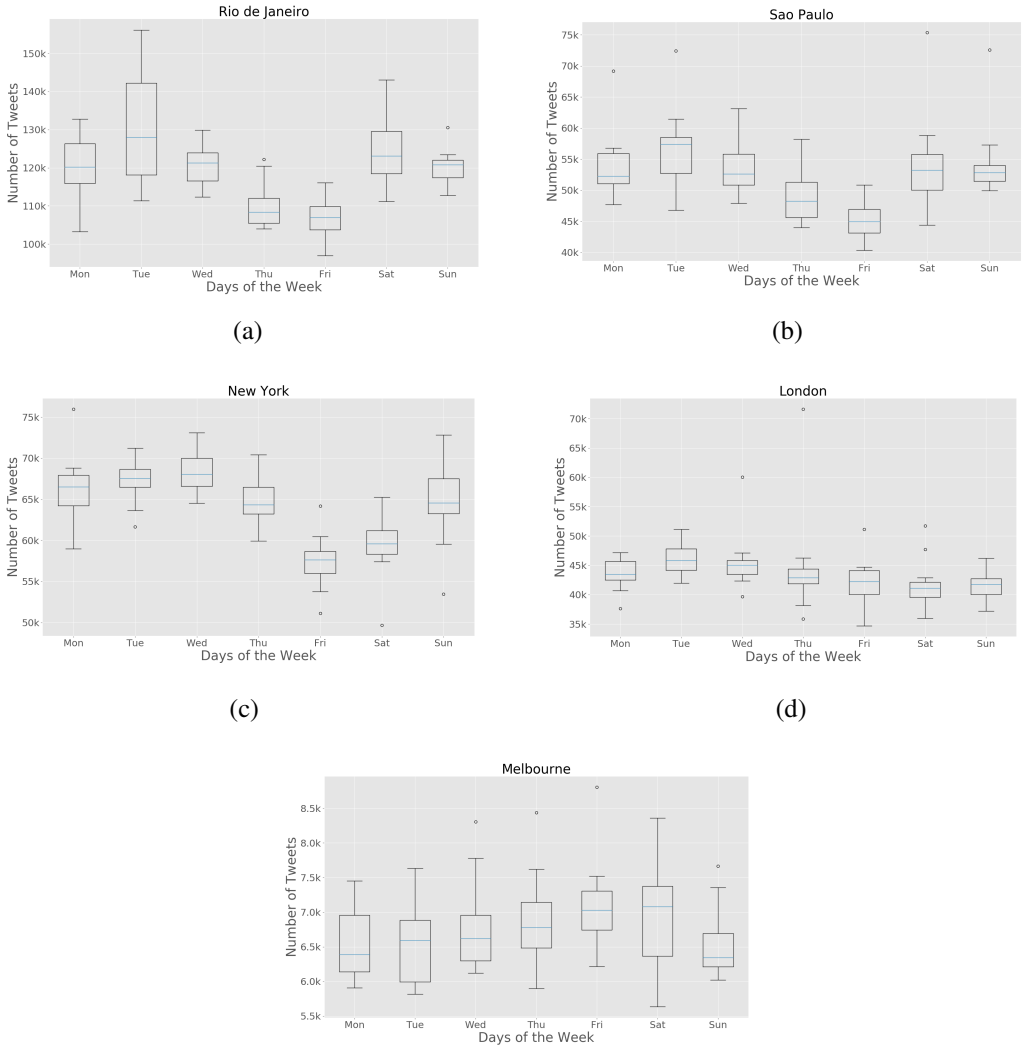
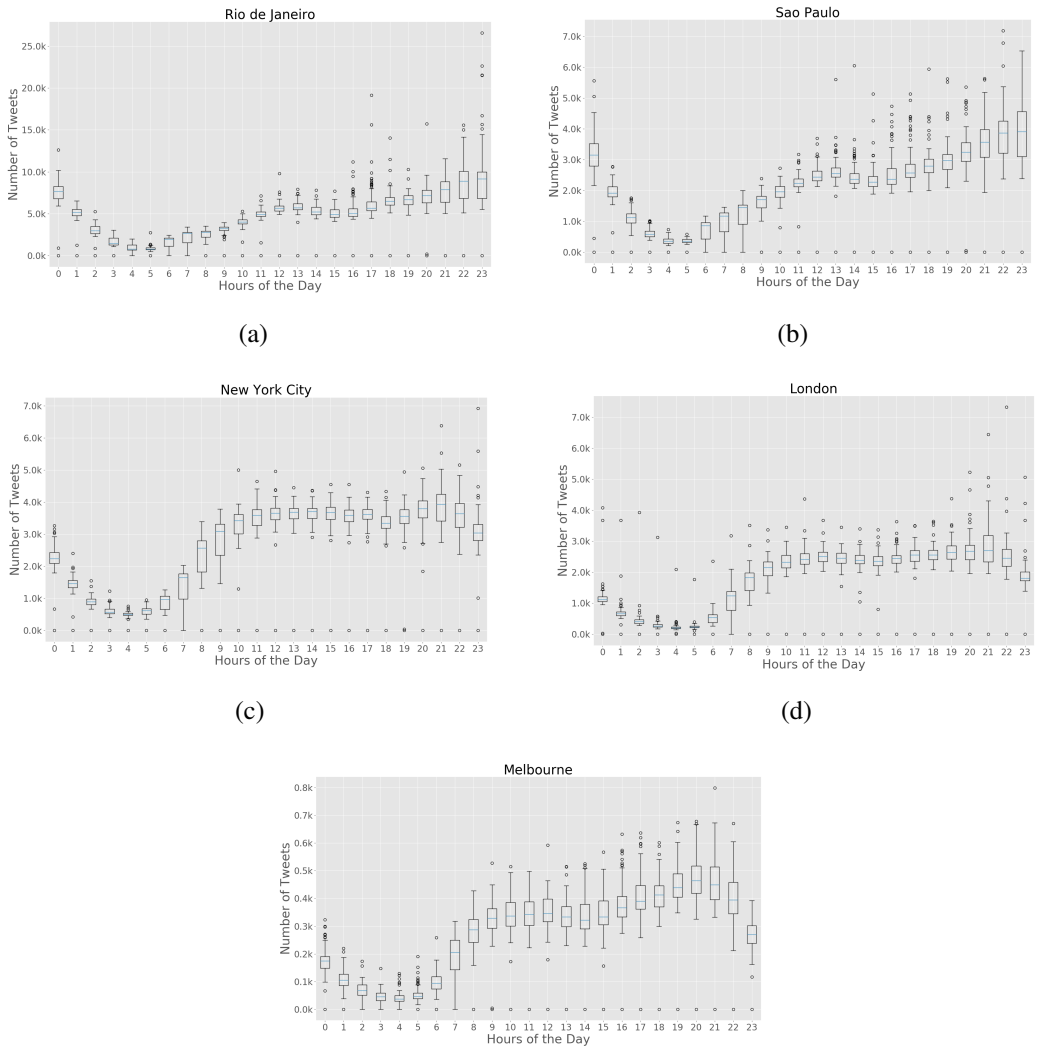


Figure 5.5: Days-of-the-week box-plots for the volume of tweets (a) Rio de Janeiro (b) São Paulo (c) New York City (d) London (e) Melbourne

levels of activity is spread between week and weekend days. The interquartile range in the plots can tell us the amount of days whose activity was above and behold the median value, and through that we identify Rio de Janeiro and Melbourne as the cities where this phenomenon happen more times. São Paulo, New York City and London present an almost regular IQR which means that the days of weeks are similarly regarding the activity on Twitter.



(a) (b) (c) (d) (e)

Figure 5.6: Hour-of-the-day box-plots for the volume of tweets (a) Rio de Janeiro (b) São Paulo (c) New York City (d) London (e) Melbourne

Looking at the hour-of-the-day box-plot (5.6), it is possible to verify an decrease in terms of activity on Twitter during the night period to all cities. More specifically, there were cases in which the volume of tweets was inexistent and based on this fact, two possible reason are suggested: (1) the absence of tweets during this period is explained through the zero activity of users in the city, regarding geo-located tweets; (2) the service on Twitter was in maintenance and due to that, any tweet was retrieved by the API. Although the observable increase of activity during day-time, the

peak of it is similiar to all cities and it is established between the 19 and 23 hours.

5.3 Content Composition

Tweets although its classification as text messages, also contain other kind of *metadata* which exploration of it can sometimes be transformed in added-value information. The *metadata* present in a tweet is represented by the *hashtags*, *user mentions*, *URLs* and *media* attached to it. Other point to explore is the number of distinct users that contributed to the datasets composition. Users which number of posts are unnatural may sometimes be *bots*. If there is a time pattern associated to the post of tweets by a user, for example, the user posts a tweet in a period of 5 minutes over the whole day, then this user is a potential *bot*. The existence of *bots* is not considered in this dissertation because the information provide by such automatic system can also be valuable. In this subsection, we demonstrated the distribution of users over the number of posts made by themselves, as well as the counts of the different type of *metadata* contained in the data.

Social media platforms present similar characteristics between themselves. One of the most studied ones is the behviour of the its users activity in its services (social media services). The visualization of users activity usually is similar to the power-law distribution long tail [MPP+13]. Here, we tried to reproduce such visualization in order to establish this kind of correlation as so to prove this behaviour over social media services. The results are present in Figure 5.7. Each city proved to have a high number of users with few posts and that is observable in the long-tail showed in the cities corresponding sub-figures (5.7a, 5.7b, 5.7c, 5.7d, 5.7e).

The counts and percentages of users that have posted a certain number of tweets was calculated in order to assure the trustiness of the aforementioned distribution. Rio de Janeiro although the highest number of tweets in the datasets only was composed by 135,449 distinct users followed by São Paulo with a lower number 110,352 individuals. The English speaking cities revealed to be very different comparatively to the Portuguese speaking cities in this factor. New York City dataset was composed by 279,554 distinct users, London presented 266,128 users and Melbourne only was composed by 31,733 individuals. Looking at these numbers, we may conclude that Rio de Janeiro has a high percentage of users with more than a certain number of tweets and following this assumption, the log-log distribution made to correlate the behaviour of a power-law distribution must be different from the other cities, at least the English speaking ones.

For example, the percentage of users that posted 20 tweets in a period of three months was almost 63% for the city of Rio de Janeiro, São Paulo registered 75%, New York City presented 84%, London showed 87% while Melbourne had 87% of his users with that number of tweets shared. Only taking this example in consideration we proved the assumption mentioned before. The distributions also presented differences if the x-axis is considered. The scale at such axis is one magnitude higher for the English speaking cities, and this means that the number of users with lower number of tweets posted in a three months period is much higher than the users with the same number for the city of Rio de Janeiro.

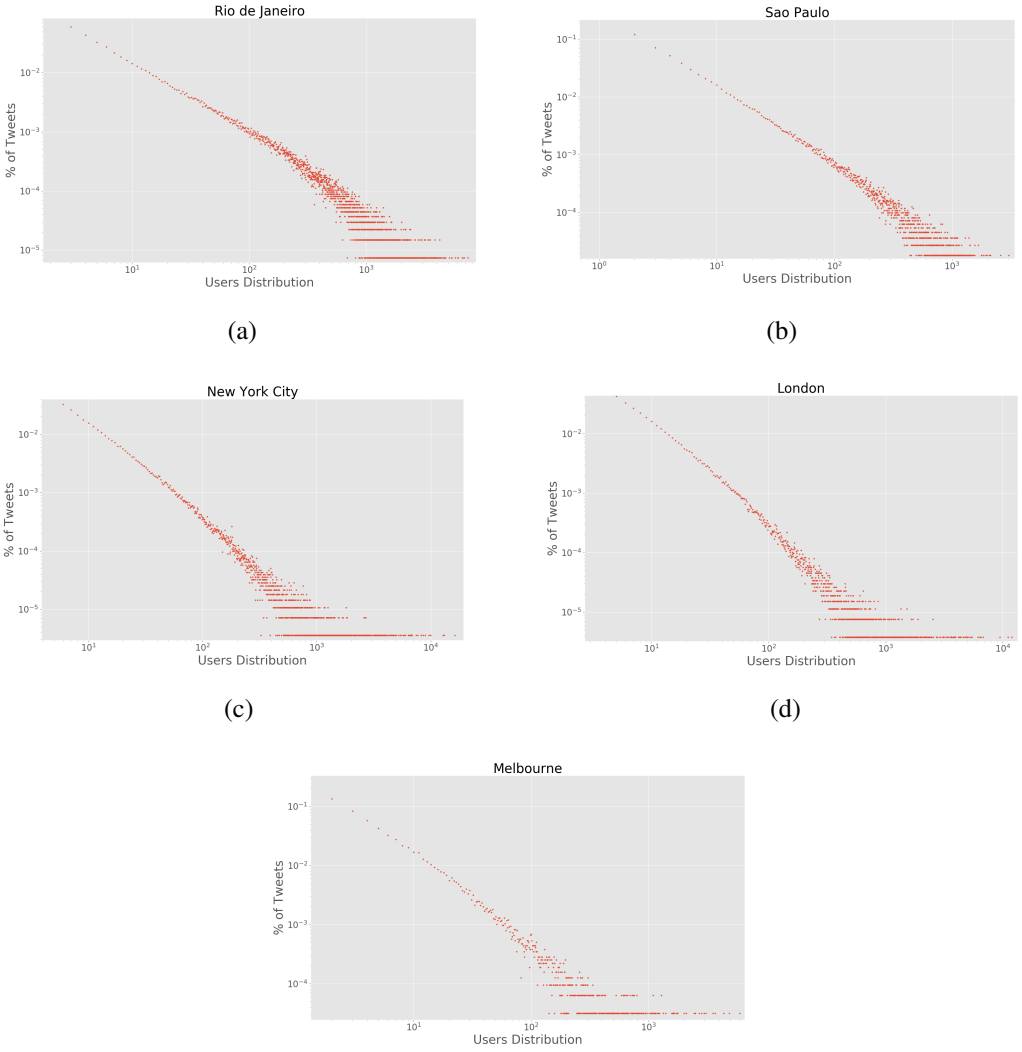


Figure 5.7: Log-log plots for the users distribution over the number of tweets posted (a) Rio de Janeiro (b) São Paulo (c) New York City (d) London (e) Melbourne

The last analysis presented in this subsection is related to the *metadata* contained in the tweets. Here, we want to characterize the different cities with respect to the amount of extra content used by the users in the posts and what kind of information such results suggests for each city.

Having this considered, we counted the volume of each element constituting the previously mentioned *metadata* and calculate the percentage of tweets containing it. In Table 5.5 are listed the counts and the corresponding percentage of it relatively to the datasets. The resulting analysis and results were performed over the tweets with the city's native language and located inside the bounding-box area used in the filtering process. The most observable evidence in the results is the greater use of this elements in the English speaking cities. User mentions, as well as *URLs* are the most used *metadata*. This elements may suggest that citizens tend to tag other people in their messages when posting and also share information about certain topic through urls. Regarding the

Brazilian cities, the *metadata* usage is not so noticeable. This fact may me related to the number of users composing each dataset because, as it was previously mentioned, the English speaking cities possesses almost two times more users than the Brazilian cities and this characteristic contributes to the increase of this type of *metadata* usage since when someone tag another one in a message, usually a re-post is sent tagging the person responsible by the starting of the conversation. To prove this so, an intensive study about social media tracking and mapping of the flow of each Twitter conversation is needed.

Table 5.5: Percentage of Metadata composing the datasets

City	Total	Hashtags (#)		User Mentions (@)		URLs		Media	
		Total (tweets)	%	Total (tweets)	%	Total(tweets)	%	Total (tweets)	%
Rio de Janeiro	11,060,136	504,835	4,56%	1,336,329	12,08%	1,783,060	16,12%	409,500	3,70%
São Paulo	4,886,626	593,952	12,15%	1,030,341	21,08%	1,111,749	22,75%	325,385	6,66%
New York City	5,956,355	1,697,416	28,50%	1,752,839	29,43%	2,839,794	47,68%	535,945	9,00%
London	4,040,092	1,163,981	28,81%	1,744,051	43,17%	1,812,152	44,85%	465,610	11,52%
Melbourne	629,424	195,967	31,13%	271,970	43,21%	258,278	41,03%	65,941	10,48%

5.4 Summary

In this chapter we tried to identify interesting patterns and valuable information recurring only to the simple characteristics provided by a tweet: location, date of creation and *metadata* content. First, it was possible to find out existing problems regarding the collection of geo-located tweets. More than one problem is mentioned and possible solutions were designed to surpass them. Our datasets represent only three months of data, however supporting in the analysis made, we conclude that the majority of tweets are tagged with variable sized bounding-boxes instead of precisely geo-coordinates. Furthermore, we tried to instigate temporal patterns using the, already, filtered tweets and proved that it is possible to learnt about remarkable events only seeing abrupt activity on Twitter for some days. By studying the Twitter users distribution it was possible correlate the behaviour of it with the famous power-law distribution. Last but not least, a brief analysis of the *metadata* was performed in order to see the amount of possible topics identified on it (hashtags), the volume of tweets mentioning another user and how many information can be share through the use of urls in this microblog, named Twitter.

Chapter 6

Experiments

The developed framework presented and described in Chapter 4 obligate us to the validation of each module in order to assure consistency and robustness in the results that such system produces. Having this considered, we stipulate several experiments, being each of them related to a specific task.

6.1 Topic Modelling

This section is related to the experiment of automatically characterize tweets in two different Brazilian cities, Rio de Janeiro and São Paulo. We used an unsupervised learning approach to tackle the task of topic modelling in order to compare both cities and see if there are differences between subjects people talked about. Automatic characterization of text messages is a laborious and time consuming task since it is necessary to assure the right level of abstraction in the learning model; very much similarly to human minds, which essentially present a bounded rationality nature, our learning model needs to be trained in order to assimilate the necessary knowledge and perform the appropriate analogies so as to discover different topics within the tweets' contents. The premises to implement such a mechanism are presented and discussed in the following subsections.

6.1.1 Data Selection

The data selected to conduct this experiment is correspondent to a period of two months, between days March 12 and May 12, 2017.

The resulting datasets sum up a total of 12.5M and 6.3M tweets for Rio de Janeiro and for São Paulo, respectively. Due to the problem detected in Section 4.3, we filtered the data in order to only use the tweets that were actually inside the cities' areas. The final composition of the datasets is presented in Table 6.1, and the results of the filtering process shown that almost 6M tweets were not located inside the bounding-boxes of the cities.

Table 6.1: Datasets composition

City	All	PT	Non-PT	In Bounding-Box	Out Bounding-Box	PT and In Bounding-Box
Rio de Janeiro	12,531,000	10,570,000	1,961,000	8,644,000	3,886,000	7,353,000
São Paulo	6,352,000	4,886,000	1,466,000	4,247,000	2,105,000	3,313,000

The subset of data composed by Portuguese tweets and located inside the cities' bounding-boxes was used to conduct the experiment described in this section. Such subset can be sum up to a total of 7.3M and 3.3M for Rio de Janeiro and São Paulo, respectively.

6.1.2 Data Preparation

Usually, to tackle topic modelling tasks in text documents it is required several pre-processing steps. Such pre-processing to the data helps the operations made by the LDA model, which is the technique used here. Removing unnecessary words, transforming words into their root form as so deleting all the punctuation are some of the common text mining pre-processing steps. Here, each tweet of both datasets was submitted to a required group of pre-processing operations in order to train a LDA model and proceed with the experiments. The pre-processing steps were the ones detailed below.

- **Lowercasing:** Every message presented in a tweet was converted into lower case;

- **Cleaning Entities and Numbers:** Removing *URLs*, user mentions, *hashtags* and digits from the text message;
- **Lemmatization:** Only plural words were transformed into singular ones;
- **Transforming repeated characters:** Sequences of characters repeated more than three times were transformed, e.g. "looooool" was converted to "loool";
- **Punctuation Removal:** Every punctuation was removed as well as smiles (e.g. `:)`, `:-)`, `=D`) or even *emoticons*;
- **Stop Words Removal:** The removing of this kind of words was made using the Portuguese NLTK dictionary;
- **Short Tokens Removal:** Words such as 'kkk', 'aaa', 'aff' and other of the same style were removed.

After the data preparation phase, 772,017 tweets have their message empty which conclude that its content was irrelevant for the final experiment phase.

6.1.3 Features Selection

Topic modelling requires, like in other learning model, a group of features to be trained. In this case, we used the Bag-of-Words representation matrix - which is a representation where each document is converted to a frequency vector according to the number of occurrences of each word in the message. The set of features was limit to a dictionary containing 10,000 words and it only took into account uni-grams in the message content. The dictionary was also limited to words that occur in a maximum percentage of 40% in the whole dataset, avoiding common words that were not removed because they were not included in the NLTK Stop Words list. The minimal occurrence value for a word being considered was set to 10.

6.1.4 LDA Model Parametrization

In order to understand and see the LDA model performance, we set five different numbers for the topics results parameter of the training process: 5, 10, 20, 25 and 50 topics, being this the one with better results. The number of iterations to train the model was set to 20, since our desired was to reproduce the experiment made by G. Lansley et al. [LL16] to the city of London. Finally but not the least, each tweet in the datasets was treated as a single document comprehending that, in total, 6,580,983 different documents were used in the model training process. The complete pipeline according to all the steps taken to conduct this experiment is observable in Figure **??**.

6.1.5 Results and Analysis

To evaluate the experimental results obtained for each model (where the difference underlies on the variation of the number of topics), a list with the most frequent 50 words for each topic was

extracted. In Table 6.2 we can observe a sample (20 top words) selected out of the 50 studied. Nonetheless, the final evaluation took into consideration all the 50 outputted words.

Table 6.2: Example of the topics classification

Words (only 20 words)	**Topic Classification**
paulo, vai, hoje, dia, jogo, ser, melhor, time, vamo, brazil, todo, santo, brasil, gol, cara, aqui, agora, corinthiam, ano, palmeiro, vem, ...	Sports and Games
vou, dia, dormir, queria, hoje, ficar, casa, semano, quero, ter, ainda, hora, agora, sono, aula, acordar, acordei, cedo, fazer, prova, ...	Wake-up Messages
top, social, artist, vote, the, award, army, bom, voting, doi, bogo, oitenta, sipda, today, vinte, prepara, cypher, oito, quatro, man, ...	Voting and Numbers
marco, nada, falar, emilly, gente, quer, nao, pessoa, nunca, fala, vai, falando, sobre, chama, agora, manda, vem, mensagem, vivian, bbb, ...	Big Brother Brazil 2017
paulo, brazil, sao, santo, vila, just, parque, posted, photo, shopping, paulista, centro, bernardo, jardim, cidade, avenido, praia, santa, campo, academia	Tourism and Places

We also selected and manually analyse a random sample (with the size of 200) of tweets for each topic. This sampling was done in order to get better consistency and trustiness about the classification and characterization of the tweets.

It was found a group of 50 topics which had the largest number of distinct topics between them. However, there were topics which theme was the same (e.g. Love and Romance Problems or Brazilian Football *versus* European Football). Within this, such groups were aggregate into the same topic, *Relationships* and *Sports and Games*, respectively. After this grouping process, a total of 29 different topics was achieved.

Some tweets that have added complexity to our classification objective, such as, for example, "*queria namorar um mano parecido com o josh*" (Relationship) and "*como eu queria meus amigos aqui agora cmg*" (Friendship), raised some doubts about which topic this tweets may belong: Relationship, Friendship or even Actions or Intentions. In a perspective of context, the first tweet belongs to the theme `flirt`, which is directly related to Relationship. The theme on the second tweet is missing the company of friends, i.e. conviviality, which is related to Friendship. The decision of join the two topics was due to the proximity between them which have as content both types of tweets, talking about love/relationship and friendship, and with this in consideration both topics should be aggregated in order to assure the desired consistency in the classification.

The final set of topics (50 topics) to be considered was selected accordantly to the most recurring subjects. The final classification and details associated with the whole dataset for each city is presented in Table 6.3. Almost every topics demonstrated a balanced distribution, with exception of *Relationships and Friendship* and *Personal Feelings* for Rio de Janeiro and São Paulo, respectively. The difference that appear in this topics is a consequence of the final grouping process, since there was a considerable number of words been shared among this topics. This issue complicated our classification task, compelling to an high amount of undesired aggregations.

Additionally to the manual verification of a sample of tweets for each topic, we also produced a temporal week day distribution, with the objective to observe if some topics had more mentions

Table 6.3: Final results of the LDA topics aggregation

Topic Group	Rio de Janeiro		São Paulo		Diff (%)
	No. Tweets	Percentage (%)	No. Tweets	Percentage (%)	
Academic Activities	101,590	1,54%	90,616	3,30%	-1,76%
Actions or Intentions	600,030	9,12%	128,710	4,69%	**+4,43%**
Antecipation and Socialising	132,606	2,01%	0	0,00%	**+2,01%**
BBB17	122,054	1,85%	68,385	2,49%	-0,64%
Body, Appearances and Clothes	160,342	2,44%	71,447	2,60%	-0,17%
Food and Drink	167,204	2,54%	58,407	2,13%	+0,41%
Health	119,013	1,81%	0	0,00%	**+1,81%**
Holidays and Weekends	104,695	1,59%	79,610	2,90%	-1,31%
Informal Conversations	272,502	4,14%	138,848	5,06%	-0,92%
Live Shows, Social Events and Nightlife	359,342	5,46%	140,240	5,11%	+0,35%
Mood	139,287	2,12%	138,399	5,04%	**-2,92%**
Movies and TV	285,198	4,33%	39,778	1,45%	**+2,89%**
Music and Artists	84,407	1,28%	78,142	2,85%	1,56%
Negativism, Pessimism and Anger	229,104	3,48%	183,050	6,67%	**-3,18%**
Numbers, Quantities and Classification	86,897	1,32%	78,160	2,85%	-1,53%
Optimism and Positivism	106,714	1,62%	39,725	1,45%	+0,18%
Personal Fellings	375,735	5,71%	532,331	19,38%	**-13,67%**
Politics	81,254	1,23%	46,758	1,70%	0,47%
Relationships and Friendship	1,524,804	23,17%	187,541	6,83%	**+16,34%**
Religion	183,174	2,78%	66,788	2,43%	+0,35%
Routine Activities	334,216	5,08%	82,421	3,00%	+2,08%
Slang and Profinities	241,676	3,67%	44,620	1,62%	+2,05%
Social Media Applications	105,809	1,61%	44,073	1,60%	+0,01%
Sport and Games	382,479	5,81%	133,047	4,84%	+0,97%
Tourism and Places	59,288	0,90%	86,519	3,15%	-2,25%
Transportation and Travel	130,261	1,98%	63,923	2,33%	-0,35%
Weather	91,302	1,39%	42,588	1,55%	-0,16%
Shopping	0	0,00%	44,470	1,62%	**-1,62%**
Voting	0	0,00%	37,687	1,37%	**-1,37%**

in certain days than others.

For making such observations some assumptions were made in relation with some *hot* topics. More specifically, we think that is valid to assume that people will talk more about *Religion* in the weekend, since they go to the church in those days. The same result is likely to happen for topics like *Holidays and Weekends* or *Sports and Games*, since events related to this thematics occur during specific time-frames.

Only 12 topics of the finals 29 were selected for this part of the study, predicting them and comparing the final results, such as, but not limited to, *Sports and Games*, *Religion*, *Holidays and Weekends*, *Movies and TV*, *Live Shows, Social Events and Nightlife*. The temporal distribution is showed in Figure 6.1 as a heat map, where each row is independent from the others.

The necessity of applying such restrictions is due to the need of seeing in which days each topic is more talked about. For both cities the topic *Sports and Games* is more mentioned in Tuesdays and Saturdays. Indeed, this observation correlates with the days that topic-related events happens. Namely, Tuesdays and Wednesday correspond to the days when the *UEFA Champions League* competition happens and Saturdays and Sundays to the days of *Brazilian Football League* games. *Holidays and Weekends* was a topic with interesting results regarding the temporal distribution, presenting Sundays as the day where more people talk about it.

Furthermore, it is worth mentioning that our model had successfully discover a topic related to

Big Brother Brazil 2017 (BBB17), a well-known reality show. The amount of geo-located tweets concerning this topic was considerable (1.85% and 2.49%, in RJ and SP, respectively), rising the question about what led people to geo-located them in such topic.

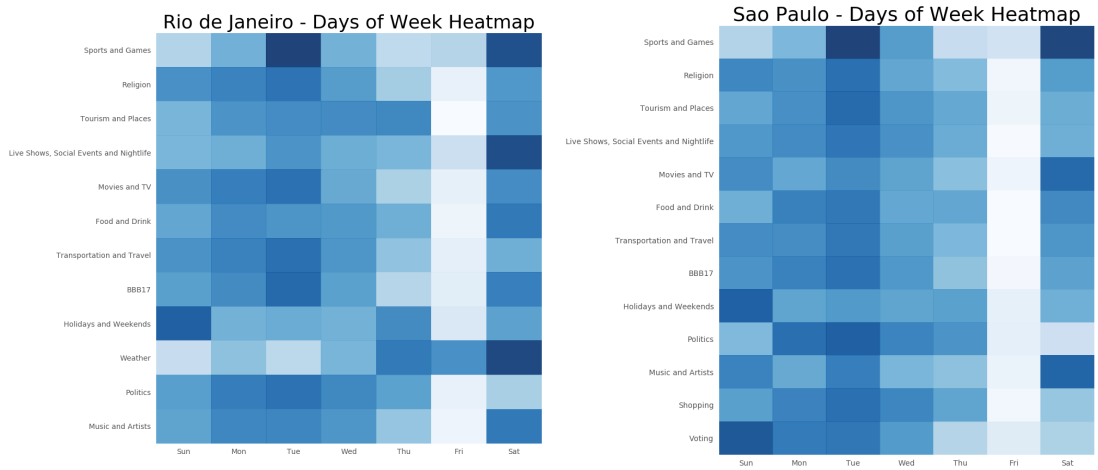


Figure 6.1: Day-of-the-week activity per each topic in both cities, Rio de Janeiro and São Paulo

6.1.6 Final Remarks

The methodology reported across this experiment is concerned with topic modelling over two datasets from two Brazilian cities in order to characterize the topics that people talked about and compare the results in both scenarios. LDA models usually requires documents of large size, or at least more complex than a single tweet, in order to get good performance. A traditional approach was followed considering each tweet as a document instead of trying aggregate tweets in more complex documents taking into consideration some criteria, e.g. grouping by date and hour. The final results showed that topics in both cities are very similar and only two of them are unique. With exception of topics - *Relationships and Friendship* and *Personal Feelings*, the percentage difference between similar topics was comprehended in the interval 0.16-4.43% evidencing the fact that both cities are similar besides the different factors that characterize each one: population, culture, lifestyle and also the region where the city is located in. Although all this analysis, we can not assure that inside a topic we do not have more topics hidden. Our classification was limited to the verification of the 50 top words and the manually verification of a sample of 200 tweets since the resulting amount of tweets for each topic is impossible to verify one by one. Due to this, another classification approach need to be explored and a promising one was proposed by D. Ramage et al. [RDL10]. The classification will be automatic by adding a supervised extra layer to the pipeline. However, to assure trustiness in the results the data may be manually labelled for the training phase of the model classification or, at least, have reliable sources, for example, exploring the topics provide by the Wikipedia articles[1].

[1] https://dumps.wikimedia.org/ptwiki/20170601/

6.2 Portuguese Travel-related Classification

The main goal of this section is to detail the experiment that supports the characterization of travel-related tweets in Rio de Janeiro and São Paulo. Considering the volume of the collected data, it was then necessary to automatically identify tweets whose content somehow suggests to be related to the transportation domain. Conventional approaches would require us to specify travel-related keywords to classify such tweets. On the contrary, our approach consisted in training a classifier model to automatically discriminate travel-related tweets from non-related ones.

One big challenge always present in text analysis is the sparse nature of data, which is especially the case in Twitter messages. Conventional techniques such as Bag-of-Words tend to produce sparse representations, which become even worse when data is composed by informal and noisy content.

Word embeddings, on the other hand, is a text representation technique that tries to capture syntactic and semantic relations from words. The result is a more cohesive representation where similar words are represented by similar vectors. For instance, *"taxi"/"uber"*, *"bus/busão/ônibus"*, *"go to work"/"go to school"* would yield similar vectors respectively. We are particularly interested in exploring the characteristics of word embeddings techniques to understand which extent it is possible to improve the performance of our classifier to capture such travel-related expressions. In the following subsections, we describe the necessary steps to build our classification model.

6.2.1 Data Selection

Messages were collected for a period of one whole month, between days March 12 and April 12, 2017, and the resulting datasets sum up a total of 6.1M and 2.9M tweets for Rio de Janeiro and São Paulo, respectively. Due to the problem detected in Section **??**, we filtered the data in order to only use the tweets that were actually inside the cities' areas. The final composition of the datasets is presented in Table 6.4, and according the previous mentioned criteria, a sum up of 7.7M tweets (5.3M and 2.4M tweets for Rio de Janeiro and São Paulo, respectively) was considered in this experiment.

Table 6.4: Rio de Janeiro and São Paulo datasets composition for the travel-related classification

City	All	PT	Non-PT	Inside Bounding-Box	Outside Bounding-Box	PT and Inside Bounding-Box
Rio de Janeiro	6,175,000	5,355,000	0,819,000	4,327,000	1,848,000	3,749,000
São Paulo	2,934,000	2,444,000	0,490,000	2,016,000	0,918,000	1,672,000

6.2.2 Data Preparation

Each tweet of our training and test sets was submitted to a small and basic group of pre-processing operations, as detailed below. Regarding the *bag-of-words* group, we limited each tweet representation to the 3,000 most frequent terms excluding also words present in more than 60% of the tweets. For *bag-of-embeddings*

- **Lowercasing:** Every message presented in a tweet was converted into lower case;
- **Transforming repeated characters:** Sequences of characters repeated more than three times were transformed, e.g. "loooool" was converted to "loool";
- **Cleaning:** URLs and user mentions were removed from the text.

6.2.3 Features Selection

We established the use of different groups of features to train our classification model, namely bag-of-words, bag-of-embeddings - word embeddings dependent technique - and both combined. Such groups are detailed below.

- **Bag-of-words (BoW):** This group of features was obtained using unigrams with standard bag-of-words techniques. We considered the 3,000 most frequent terms across the training set excluding the ones found in more than 60% of the documents (tweets);
- **Bag-of-embeddings (BoE):** We applied bag-of-embeddings to each tweet using a *doc2vec* model [2] combining Deep Learning and *paragraph2vec*. The model was trained with 10 iterations over the whole Portuguese dataset using a context window of value 2 and feature vectors of 50, 100 and 200 dimensions. We then took the corresponding embedding matrix to yield the group of features fed into our classification routine.
- **Bag-of-words plus Bag-of-embeddings:** We horizontally combined both the above matrices into a single one and used it as a single group of features.

6.2.4 Training and Test Datasets

The construction of the training and test sets followed a traditional approach. We thus tried to select balanced training sets, to which it was necessary to identify tweets that could possibly be travel-related. We were inspired by a strategy used in the study by Maghrebi et al. [MAW16], which consists in searching tweets from a collection using specific travel terms and regular expressions.

Using the terms declared in Table 6.5 combined with the regular expression $space + term + space$, we found about 30,000 tweets. From this subset, we randomly selected a small sample of 3,000 tweets to manually confirm if they were indeed related to travel topics. After this manual annotation we selected 2,000 tweets and used them as positive samples in the training dataset.

In order to select negative samples for the training dataset we randomly selected 2,000 tweets and also manually verified their content to assure that they were not travel-related. Finally, our training set was composed by 4,000 tweets, from which 2,000 were travel-related and 2,000 were not. We selected 1,000 tweets randomly that were not present in the training set to build the test set, and then manually classified them as travel-related or non-travel-related. In the end, 71 tweets were found to be travel-related and whereas 929 were not.

[2] https://radimrehurek.com/gensim/models/doc2vec.html (Accessed on 09/06/2017)

Table 6.5: Travel terms used to build the training set

Mode of Transport	Terms	
	Portuguese Language	**English Language**
Bike	bicicleta, moto	bicycle, bike
Bus	onibus, ônibus	bus
Car	carro	car
Taxi	taxi, táxi	taxi, cab
Train	metro, metrô, trem	metro, train, subway
Walk	caminhar	walk

6.2.5 Estimators and Evaluation Metrics

Support Vector Machines (SVM), Logistic Regression (LR) and Random Forests (RF) were the classifiers used in our experiments. The SVM classifier was tested under three different kernels, namely *rbf*, *sigmoid* and *linear*; the latter proved to obtain the best results.

The LR classifier was used with the standard parameters, whereas the RF classifier used 100 trees in the forest. The gini criterion and the maximum number of features were limited to those as aforementioned in Section **??**, in the case of the RF classifier.

To evaluate the performance of the classifiers in our experiences we used five different metrics. Firstly we compute a group of three per-class metrics, namely precision, recall and the F1-score. Bearing in mind this study considers a binary classification, metrics were associated with the travel-related class only, i.e. the positive class. Therefore, the interpretation for each metric is provided below:

- **Precison:** Represents the fraction of correct predictions for the travel-related class (Equation 6.1).
- **Recall:** Represents the fraction of travel-related tweets correctly predicted (Equation 6.2).

$$Precision = \frac{tp}{tp + fp} \qquad (6.1) \qquad\qquad Recall = \frac{tp}{tp + fn} \qquad (6.2)$$

  where tp is related to the true positives classified tweets, fp represents the false positives and fn are the false negatives.

- **F1-score:** Represents the harmonic mean of precision and recall.

$$F1_{score} = 2 * \frac{precision * recall}{precision + recall} \qquad (6.3)$$

Once these first three metrics only showed us the performance of the classifier for a discrimination threshold of 0.5, we decided to calculate another metric. The ROC (Receiver operating characteristic) curve gives us the TPR (True positive rate) and the FPR (False positive rate) for all possible variations of the discrimination threshold. Through the ROC curve, we compute the

area under the curve (AUC) to see what was the probability of the classifier to rank a random travel-related tweet higher than a random non-related one.

6.2.6 Results and Analysis

Table 6.6 presents the results obtained using the different features combination for our test set composed by 1,000 tweets manually annotated. According to the evaluation metrics we conclude that the bag-of-word and bag-of-embeddings combined produced better classification models. The model produced by the Linear SVM performed slightly better than the LR and the RF. Interesting to note is that BoW features have influence on the precision scores obtained from our results, producing more conservative classifiers. Regarding the recall results, we can see that the Logistic Regression using only bag-of-embeddings features was the model with best results; perhaps if the precision is taken into consideration, the same conclusions will not be possible. Analysing the scores provided in Table 6.6, the best model under the F1-score was the Linear SVM, with a score of 0.85. It is worth noting that combining Bag-of-words and Bag-of-embedding with size 100 was the group of features with best performance taking into consideration the evaluation metrics used in this experiment.

Table 6.6: Performance results with 100 sized vectors for BoE

Classifier	Features	Precision	Recall	F1-score
Linear SVM	BoW	1.0	0.6761	0.8067
	BoE	0.4338	0.8309	0.5700
	BoW + BoE	**1.0**	**0.7465**	**0.8548**
Logistic Regression	BoW	1.0	0.6338	0.7759
	BoE	0.4444	0.8451	0.5825
	BoW + BoE	1.0	0.6761	0.8067
Random Forest	BoW	1.0	0.6338	0.7759
	BoE	0.2298	0.8028	0.3574
	BoW + BoE	1.0	0.6338	0.7759

The performance of all three classifiers is illustrated using the ROC Curve in Fig. 6.2. The area under the curve of the Receiver Operating Characteristic (AUROC) was very similar for both the Logistic Regression and the Linear SVM models. The results obtained from the Random Forest model were not so promising as expected.

After the selection of our classification model, we decided to classify all the Portuguese dataset and draw some statistics from the results. The trained Linear SVM classifier was used to predict whether tweets were travel-related or not, since it was the model presenting the best score under the F1-score metric (as shown in Table 6.6). From a total of 7.8M tweets, our classifier was able identified 37,300 travel-related entries.

Fig. 6.3 depicts the distribution of travel-related tweets over the days of the week. We can see that the first three business days (Monday, Tuesday and Wednesday) are the ones on which the Twitter activity is higher for both cities in our study.

In order to understand the spatial distribution of travel-related tweets we generated a heatmap for both cities. From the heatmap of RJ, illustrated in Fig. 6.4, it is possible to identify that

Figure 6.2: ROC Curve of SVM, LR and RF experiences

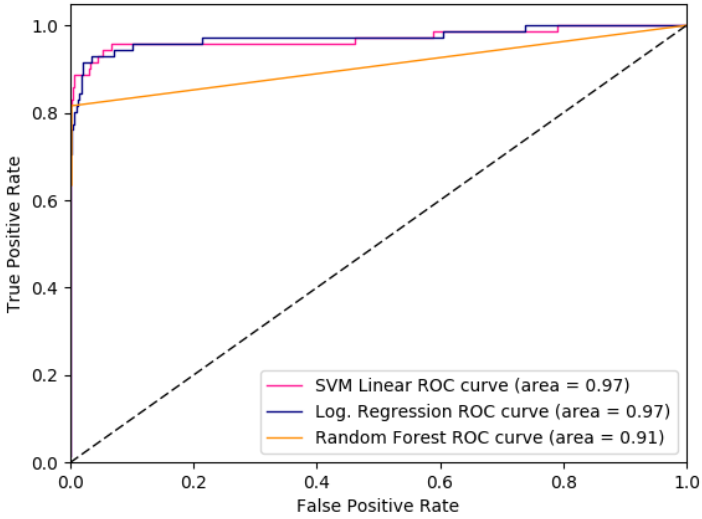


Figure 6.3: Positive Predicted Tweets per Day of Week

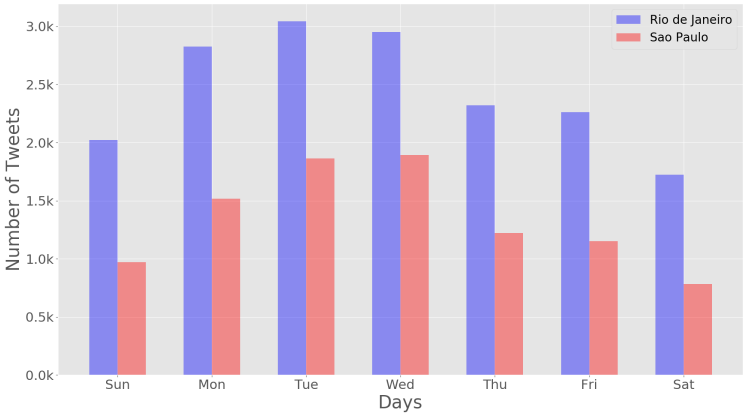
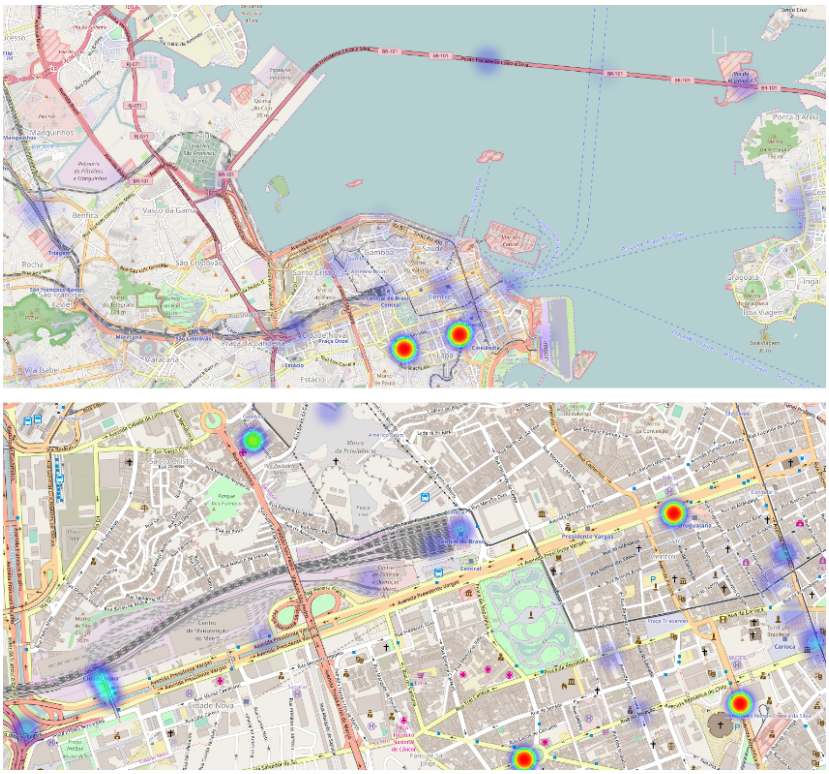


Figure 6.4: Rio de Janeiro Heatmap to the positive tweets



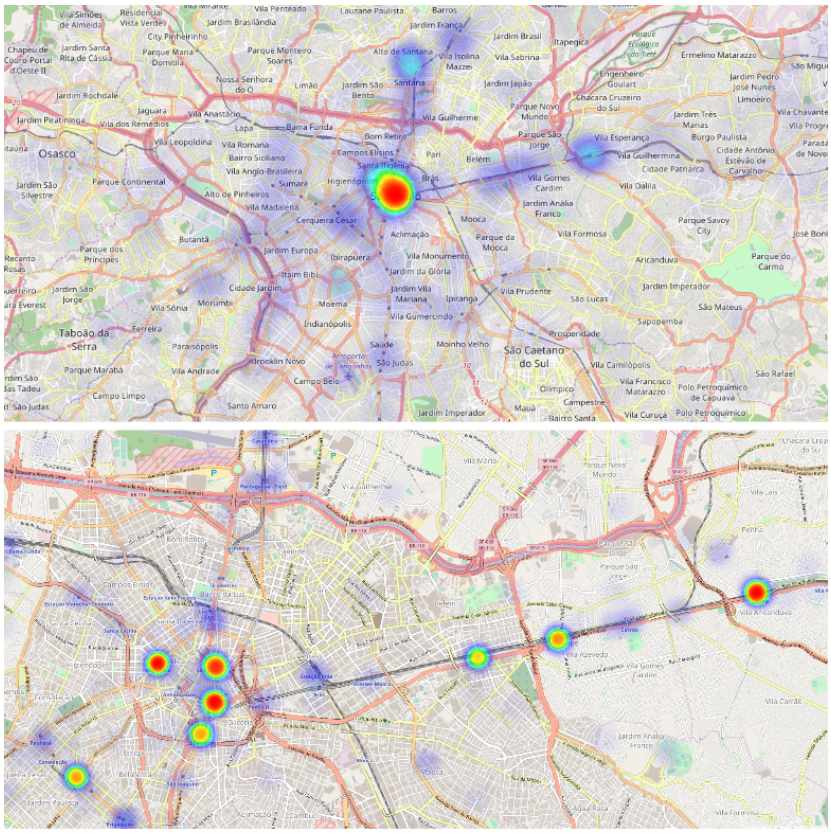
some agglomerations of tweets are located at Central do Brasil, Cidade Nova and Triagem train stations, as well as at Uruguaiana, Maracanã and Carioca metro stations. The Rio-Niterói bridge, connecting Rio de Janeiro to Niterói, as well as the piers on both sides also presented considerable clouds of tweets classified as travel-related.

The heatmap for the city of SP, illustrated in Fig. 6.5, was also an interesting case to observe. Almost every agglomeration matched some metro or train station. Estação Brás, Tatuapé, Belém, Estação Paulista, Sé, Liberdade were some of the stations highlighted in the heatmap. We could also identify a little agglomeration of travel-related tweets at Congonhas airport, even though no tweets seemed to mention the word *plane* explicitly in the training of our classification model.

6.2.7 Final Remarks

The experiment previous described explores an approach of supervised learning using as training examples a set of manually annotated tweets extracted from the whole datasets with the support of a term-based regular expression. The overall methodology is concerned with the problem of construct a fine-grained Twitter training set for the travel domain and also the automatic identification of travel-related tweets from a large scale corpus. We combined different word representations to verify whether our classification model could learn relations between words at both syntactic and semantic levels. After using standard techniques such as bag-of-words and bag-of-embeddings,

Figure 6.5: São Paulo Heatmap to the positive tweets



we have used them combined yielding results that showed that these different groups of features can complement each other, with respect to Portuguese-speaking tweets.

6.3 English Travel-related Classification

Similar to the experiment of Portuguese travel-related classification, we built a model to discriminate english-speaking travel-related tweets. However, by following the same approach, final results were not improved with the combination of two different groups of features, bag-of-words and bag-of-embeddings.

The overall experiment steps as well as the final results are showed in the following subsections.

6.3.1 Data Collection and Preparation

Differently from the Portuguese experiment, tweets were collected from New York City during a period of two months, between days March 12 and May 12, 2017. Ignoring all non-English tweets the resulting dataset comprehends 4M tweets.

Regarding the preparation of data, we used the same preprocessing operations for each tweet present in our dataset:

- **Lowercasing:** The message was converted to lowercase;
- **Transforming repeated characters:** Sequences of characters repeated more than three times were transformed, e.g. "sooooo" was converted to "sooo";
- **Cleaning:** Removing URLs and user mentions.

6.3.2 Features Selection

The features groups used in this experiments were the same presented in Section 6.2.3.

6.3.3 Training and Test Datasets

The construction of the training and test sets were supported by the same term-based approach used in Section 6.2.4 in order to filter tweets from the whole collection, i.e. we used the regular expression $space + term + space$ with each term presented in Table 6.5. Firstly, 1,686 tweets were selected for each of both cases, travel-related and non-related. The travel-related set was strictly balanced in order to have almost the same amount of examples for each of the travel-modes involved in this study. The non-related training set is composed of several subjects that are not related to travel, e.g. football, leisure, politician, personal tweets, among others.

6.3.4 Classification

We choose a supervised learning approach in order to provide a robust solution for the classification task. Three learning algorithms were selected to conduct our experiments, namely Support Vector Machines (SVM), Logistic Regression (LR) and Random Forests (RF). The SVM classifier was tested under the *linear* kernel function. To the LR classifier, standard parameters were applied, whereas the RF classifier was defined with 100 trees in the forest. The *gini criterion* and the maximum number of features were limited to those previous mentioned in Section 6.2.3, in the case of the RF classifier. The performance of the resulting models will be compared in terms of *precision*, *recall* and the *F1-score*.

6.3.5 Preliminary Results

In our first attempt, 10-fold cross-validation was applied for each model using, independently, bag-of-words and bag-of-embeddings as features. Results showed us that all the models obtained good performance regarding the selected evaluation metrics. The best model in this experiment was the Random Forests classifier trained with bag-of-words features, performing an F1-score of 0,977. Indeed, all the models that used bag-of-words features, in particular, revealed high scores as can be observed in Table 6.7. This may be explained by the similar vocabulary present in both training and test sets. One important note is that all travel-mode classes are known by the model before the classification of the test set. This may not be true in real-world scenarios. Although the results presented in Table 6.7, we tried to combine both features and conclude that, contrarily to the

Portuguese travel-related experiment, the performance was decreased when comparing it with the one obtained from the usage BoW features in the experiment. To further investigate the robustness of the best features group we designed another experiment that is explained in Section 6.3.6.

Table 6.7: Preliminary Results

Classifier	Features	Precision	Recall	F1-score
Linear SVM	BoE (200)	0,90883	0,83634	0,87089
	BoW	**0,96298**	**0,97652**	**0,96962**
Logistic Regression	BoE (100)	0,90172	0,84948	0,87447
	BoW	**0,96431**	**0,98042**	**0,97222**
Random Forests	BoE (100)	0,81283	0,83600	0,82394
	BoW	**0,96569**	**0,98997**	**0,97764**

6.3.6 *Leave-one-group-out*

The second experiment follows a *leave-one-group-out* strategy. Meaning that one travel-mode class if left out of the training set and moved into the test set. This way, the behaviour of the learned model when facing a completely unknown travel-mode class can be evaluated. A model for each hidden mode of transport class was built, and evaluation is carried as the previous experiment. The datasets composition of each experiment led in this strategy can be observed in Table 6.8.

Table 6.8: Datasets Composition

Travel-Mode Class	Training Set		Test Set	
	Pos.	Neg.	Pos.	Neg.
Taxi	1,372	1,686	314	300
Train	1,369		317	
Car	1,369		317	
Bike	1,386		300	
Walk	1,469		217	
Bus	1,375		311	

Each learning model experiment was made varying the hidden travel-mode class, which is unknown for our classifier in the training process. This method was performed in order to evaluate the sensitivity and robustness of the models built in our first experiment, described in Section 6.3.5. Table 6.9 presents the best results for each model, as so its features and tuning parameters. The results from the models using bag-of-embeddings features revealed a consistent performance, i.e. they do not change even with the variation of the size of the feature vectors.

According to results, all classification models have performed reasonably well under the bag-of-embeddings features group, although the dimensionality used being different for the Linear SVM classifier.

After testing each model with a hidden travel-mode class, the models trained with bag-of-words features demonstrated poor performance when facing unknown travel-modes, revealing higher sensitivity and lower generalization capabilities in comparison to the bag-of-embeddings

Table 6.9: *Leave one group out* experiments results for SVM, LR and RF classifiers

Classifier	Features	Precision	Recall	F1-score
Random Forests	BoW	0,40774	0,07474	0,12629
	BoE (50)	**0,80278**	**0,76194**	**0,78447**
Logistic Regression	BoW	0,40774	0,07474	0,12629
	BoE (50)	**0,84882**	**0,75702**	**0,80219**
Linear SVM	BoW	0,41527	0,07153	0,12203
	BoE (200)	**0,86374**	**0,75715**	**0,81289**

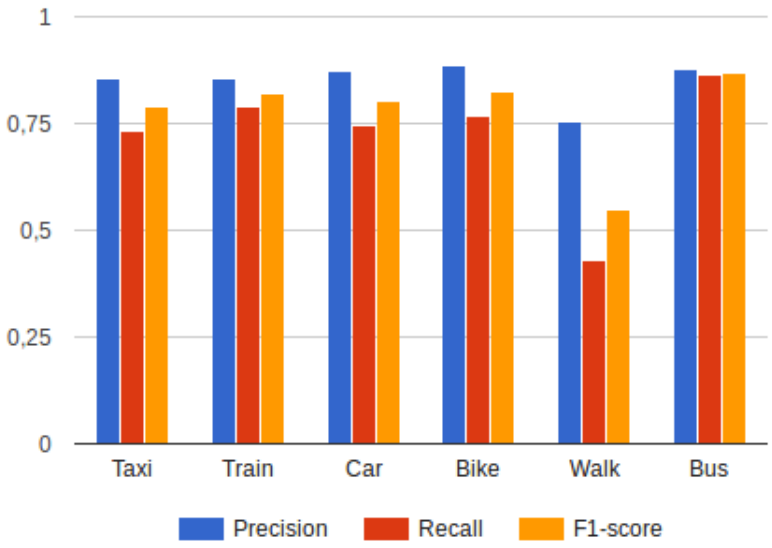


Figure 6.6: SVM model with BoE(200) for each travel mode

version. The generalization power is an important and crucial characteristic for our desired solution. In a real world scenario is very likely that we will face a higher variety of categories that were not taken into consideration in the training phase of our model.

Table 6.10: Sample of tweet messages correctly classified

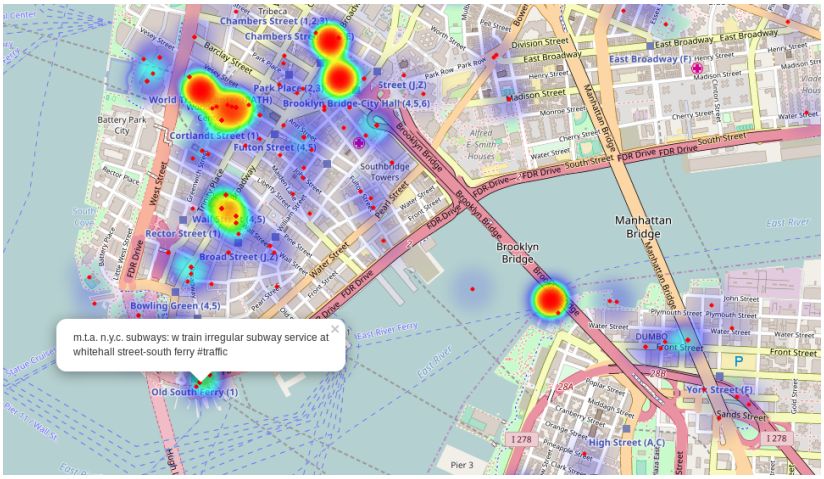
when you get into your uber and he has a pipe in the back
a ground stop for #ewr is no longer in effect #flightdelay
snowy walk to work. #blizzard2017 #centralpark #noreaster2017 bethesda terrace fountain - **Figure 6.7b**
m.t.a. n.y.c subways: w train irregular subway service at whitehall street-south ferry #traffic - **Figure 6.7a**

The best result of the *leave-one-group-out* was the Linear SVM model, with the dimensionality of 200 in the size of the feature vectors. Figure 6.6 presents the results of each experiment led for the different hidden travel-mode classes. An interesting point to observe is the low performance obtained to the experiment with the travel-mode class "Walk" hidden. This is due to the different semantic and syntactic contexts that the word *walk* is used. Although all other classes can be used in the same context, for example, *car*, *train*, or *bus*, usually the word *walk* is not applied in the same way.

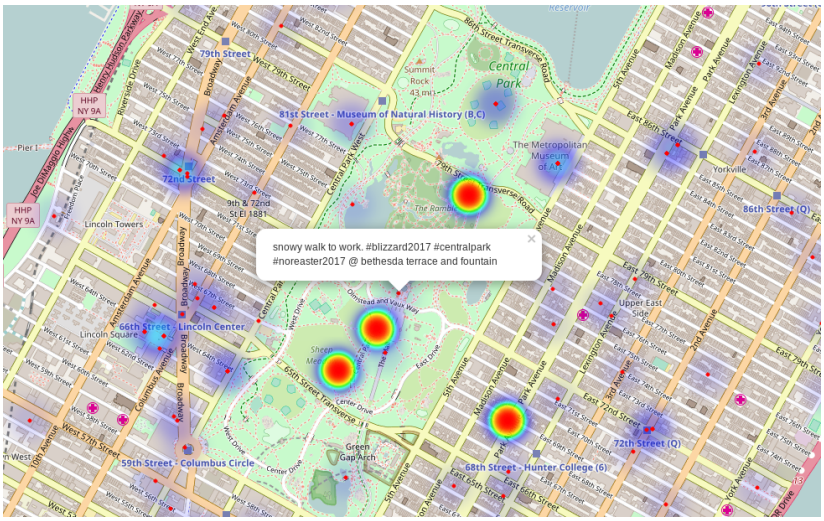
Having the experiments concluded, we used the best model, in this case, Linear SVM for the dimensionality of 200, to predict the 4M tweets that composed the NYC dataset. Almost 300,000 tweets were classified as travel-related. After the classification step, a sample of 10,000 tweets was taken from all the travel-related classified tweets and it was produced a heat-map distribution in order to verify which are the most concentrated zones. Such distribution enables the identification of associations with metro, train, bus stations. In Figure 6.7a, that shows the south of the Manhattan island and also the Brooklyn bridge, it is possible no note some agglomerations over the bridge and also in the port and closed to the Wall Street(4.5) where there are some metro stations. The Central Park is one place that also took our attention since presented several agglomerations of tweets. In this particular place, tweets related to the walk class were correctly identified.

6.3.7 Concluding Remarks

The main objective of this experiment was to devise a travel-related tweet classifier using word embeddings trained with geo-located English-speaking tweets. Similar to the Portuguese travel-related classification, we tried to build our model using a combined approach relying on bag-of-words and bag-of-embeddings features; however, results presented signs of dependency in the bag-of-words features and the performance have also decreased. By looking in the results of the best group, bag-of-words, we doubt about the existence of overfitting, and so, a *leave-one-group-out* strategy was applied to attempt reproduce and validate the results obtained from classification models in preliminary experiment. Such an strategy shows that our training and test sets were very similar to each other. In this second experiment, we excluded one of the travel-modes classes, which resulted in the fact that models using bag-of-words features could not maintain the performance previously demonstrated. Comparatively to the approach based on bag-of-words, the models using bag-of-embeddings features revealed consistency, robustness, and effectiveness



(a)



(b)

Figure 6.7: Spatial density of the travel-related predicted tweets in New York City: (a) South of Manhattan and over the Brooklyn Bridge, (b) Central Park

in the classification task. The Linear SVM model proved to be the best option with respect to the performance metrics considered in this work. We thus used that model trained with bag-of-embeddings to predict all the English tweets from our NYC dataset, whose results showed significant improvement over a standard bag-of-words baseline. Finally, we applied the resulting classifier to a stream of geo-located tweets in New York City, which was able to depict important spatio-temporal patterns.

6.4 Summary

This chapter has the purpose of report the experiments conduct over this dissertation period in order to help the implementation of the different modules designed in our framework architecture. Firstly, topic modelling techniques were applied under Portuguese-speaking tweets for two different *megacities*, Rio de Janeiro and São Paulo, in order to extract information that may enabling interesting characterizations in different regions/zones of the cities regarding temporal and geographical distributions. Moreover, two different classification models for travel-related tweets were developed taking into consideration two possible languages in texts, Portuguese and English. Under the implementation of the Portuguese classification, we were able to prove that the combination of conventional techniques (bag-of-words) and recent ones (word embeddings) performed very well. However, for the English classification, the high performance values obtained using only bag-of-words led us to suspect of the existence of overfitting in the examples used as training. An *leave-one-group-out* strategy was taken to proved such phenomenon and conclude our suspicions of similar words being shared in training and test datasets. When a transport-class was omited, the model with bag-of-words performed worst than the one using only bag-of-embeddings. For this reason we were obligated to the application of two different classification models in the development of the frameworks' travel-related classification module. This allows consistency and robustness in the classification of tweets for two distinct speaking languages.

Chapter 7

Conclusions and Future Work

This planning report had two distinct objectives. The first one is the search of related works in order to see what is already developed to the problem context of this dissertation. The second objective is the initial planning of the dissertation work, as well as the approach and methodology chosen to tackle the problem in hands. From the all work made so far, it is possible to make some conclusions.

This dissertation proposes to tackle the problem of extraction of aspect-based sentiment from the citizens opinions about the services of a city, in social media streams, through a framework that may be capable of processing the messages and build some appealing visual indicators.

Hence, the problem was decomposed in some sub-problems. The literature review served to find interesting solutions for each sub-problem. There, a great diversity of approaches was found, not only about sentiment analysis that is the most important task in this dissertation but also for another problems like the content filtering and disambiguation.

The proposed framework can be seen as a potential tool to the users of the city's services and for the responsible entities, allowing that only good decisions are made to improve the quality of the cities and, in this particular case, the urban transportation systems.

To summarize the conclusions of all the work made so far, a SWOT analysis was conceived and the points that composed it are present below.

1. **Strengths**
    - Added value proposal by combining multiple State-of-the-Art approaches to tackle chained sub-tasks;
    - Well defined sub-tasks/modules will make it easier to track errors.
2. **Weaknesses**

- It might be difficult to collect enough relevant data for specific scenarios (e.g. the quality of the urban transportation in Porto);
- Twitter data might not be so reliable if there are few relevant messages.

Opportunities

- New scenario application for aspect-based sentiment analysis: transportation systems and Smart Cities;
- Extending State-of-the-Art approaches in each sub-task/module if the target scenario presents specific constraints.

Threats

- Absence of ground-truths for the target scenarios may lead to underperformed modules;
- Limited time for implementation is a risk of some unforeseen difficulties arise.

7.1 Expected Contributions

The work to be developed in this dissertation should present contributions both at the technological and scientific level. Some of the most important contributions are listed below:

- A brief review of related literature to help contextualize readers in the subject of information extraction, in particular the sentiment analysis, from social media streams and how difficult is this task;
- Development of a tool that could bring a potential value to the cities in order to improve the quality of its services;
- The studies of use cases about Smart Cities and Transportation Systems using aspect-based sentiment analysis may be considered something innovative since there are very few works related with both scenarios.

7.2 Task Planning and Scheduling

The tasks to be undertaken are mostly based in the modules described in Section 4.2 for the proposed framework architecture. The first task is to choose what are the specific scenarios that will serve to test the developed framework. A priori, two different scenarios will be enough to prove the good functionality and usability of the tool. Hence, the crawler module will be used to collect social media streams from the middle of February until, approximately, the ending of May. Meanwhile, the setup of the framework environment needs to be done. After this first step, the development of the modules will occur. The first module to tackle is the aspect-based sentiment

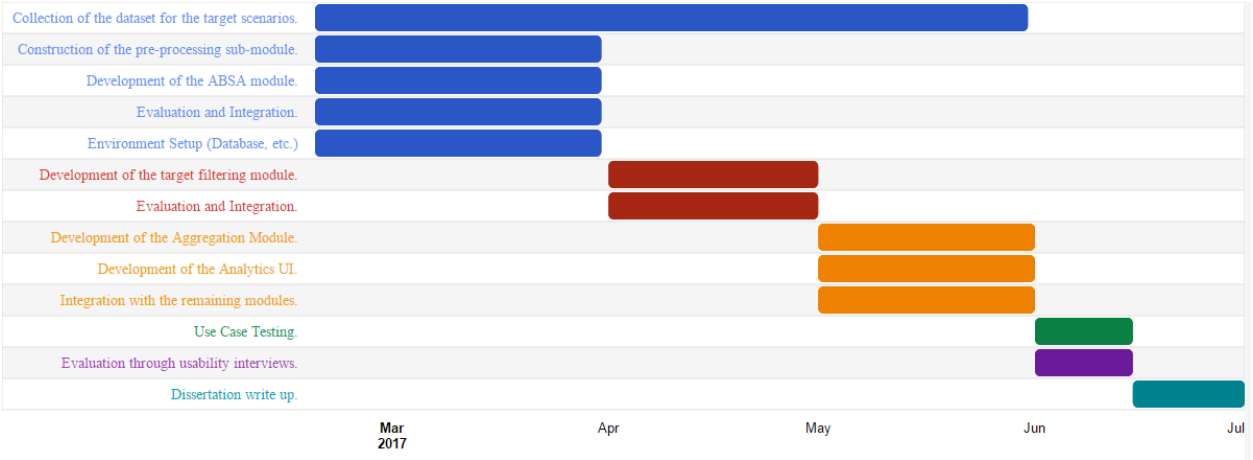


Figure 7.1: Dissertation working plan.

analysis and the sub-module of preprocessing. With the estimation of a possible margin of error, these tasks are ready to employment in the begin or middle of April. The target filtering module will be developed, if everything is going as planned, between the beginning/middle of April until the middle of May. The remaining month of May will serve to work on the aggregation module and the analytics UI. The month of June will be to test the final framework into the collected dataset about the two different scenarios. In order to evaluate the usability of the framework, it's planned the existence of a bunch of interviews to see if it's really possible that users of this tool are capable of immediately identify some conclusion from the analysis presented. This evaluation step will occur in the first two weeks of June, being the remaining two to the final dissertation report write up.

In the Figure 7.1 it's possible to visualize a Gantt chart scheduling according the mentioned tasks and the ideal scenario in case there are no delays.

2 , being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, $IQR = Q3Q1$

References

[AAB+13] G. Anastasi, M. Antonelli, A. Bechini, S. Brienza, E. D'Andrea, D. De Guglielmo, P. Ducange, B. Lazzerini, F. Marcelloni, and A. Segatori. Urban and social sensing for sustainable mobility in smart cities. pages 1–4, Oct 2013. Cited on page 9.

[AMB+13] Leonardo Allisio, Valeria Mussa, Cristina Bosco, Viviana Patti, and Giancarlo Ruffo. Felicitt??: Visualizing and estimating happiness in Italian cities from geotagged Tweets. *CEUR Workshop Proceedings*, 1096:95–106, 2013. Cited on pages 20 and 24.

[Ang15] Margarita Angelidou. Smart cities: A conjuncture of four forces. *Cities*, 47:95–106, 2015. Cited on page 8.

[AZ12] Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012. Cited on pages 14, 15, and 21.

[BAG+12] Michael Batty, Kay W Axhausen, Fosca Giannotti, Alexei Pozdnoukhov, Armando Bazzani, Monica Wachowicz, Georgios Ouzounis, and Yuval Portugali. Smart cities of the future. *The European Physical Journal Special Topics*, 214(1):481–518, 2012. Cited on page 1.

[BCJ+12] Nilanjan Banerjee, Dipanjan Chakraborty, Anupam Joshi, Sumit Mittal, Angshu Rai, and B Ravindran. Towards Analyzing Micro-Blogs for Detection and Classification of Real-Time Intentions Towards Analyzing Micro-Blogs for. (January), 2012. Cited on page 10.

[BDF+13] Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A Greenwood, Diana Maynard, and Niraj Aswani. Twitie: An open-source information extraction pipeline for microblog text. pages 83–90, 2013. Cited on page 13.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. Cited on pages xi and 35.

[Bor99] Andrew Borthwick. *A maximum entropy approach to named entity recognition*. PhD thesis, Citeseer, 1999. Cited on page 15.

[BSW99] Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. An algorithm that learns what‘s in a name. *Mach. Learn.*, 34(1-3):211–231, February 1999. Cited on page 15.

[CBB+13] Jean-Valère Cossu, Benjamin Bigot, Ludovic Bonnefoy, Mohamed Morchid, Xavier Bost, Grégory Senay, Richard Dufour, Vincent Bouvier, Juan-Manuel Torres-Moreno, and Marc El-Bèze. Lia@replab 2013. 2013. Cited on page 18.

REFERENCES

[CBRB12] José M. Chenlo, Jordi Atserias Batalla, Carlos Rodriguez, and Roi Blanco. Fbm-yahoo! at replab 2012. 2012. Cited on page 18.

[CD15] Andrea Caragliu and Chiara F. Del Bo. Do Smart Cities Invest in Smarter Policies? Learning From the Past, Planning for the Future. *Social Science Computer Review*, 34(6):1–16, 2015. Cited on page 8.

[Chi08] Ed H Chi. The social web: Research and opportunities. *IEEE Computer*, 41(9):88–91, 2008. Cited on page 1.

[CL15] Byung-tae Chun and Seong-hoon Lee. Review on ITS in Smart City. *Advanced Science and Technology Letters*, 98:52–54, 2015. Cited on page 8.

[CSMA16] Angel X Chang, Valentin I Spitkovsky, Christopher D Manning, and Eneko Agirre. A comparison of named-entity disambiguation and word sense disambiguation. 2016. Cited on page 13.

[CSR10] Sara Carvalho, Luís Sarmento, and Rosaldo J. F. Rossetti. Real-time sensing of traffic information in twitter messages. In *4th Workshop on Artificial Transportation Systems and Simulation (ATSS), 2010 13th International IEEE Conference on Intelligent Transportation Systems (ITSC 2010), Funchal, Portugal, 19-22 Sept. 2010*, pages 1–4, 2010. Cited on page 2.

[CVSO10] Miguel Ángel García Cumbreras, Manuel García Vega, Fernando Martínez Santiago, and José Manuel Perea Ortega. Sinai at weps-3: Online reputation management. 2010. Cited on page 18.

[DDLM15] Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, and Francesco Marcelloni. Real-Time Detection of Traffic from Twitter Stream Analysis. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):2269–2283, 2015. Cited on page 10.

[DSGD15] Derek Doran, Karl Severin, Swapna Gokhale, and Aldo Dagnino. Social media enabled human sensing for smart cities. *AI Communications*, 29(1):57–75, 2015. Cited on page 9.

[dSHJ14] Nádia F.F. da Silva, Eduardo R. Hruschka, and Estevam R. Hruschka Jr. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170 – 179, 2014. Cited on pages xi, 22, and 23.

[Fel13] Ronen Feldman. Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4):82–89, April 2013. Cited on page 20.

[FG13] Weiguo Fan and Michael D Gordon. Unveiling the Power of Social Media Analytics. *Communications of the ACM*, 12(JUNE 2014):1–26, 2013. Cited on page 10.

[FS10] Paolo Ferragina and Ugo Scaiella. Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). pages 1625–1628, 2010. Cited on page 17.

[GBH09] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009. Cited on page 22.

REFERENCES

[GC16] Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Comput. Surv.*, 49(2):28:1–28:41, June 2016. Cited on pages 22, 23, 24, and 25.

[GSZ13] M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, 40(16):6266 – 6282, 2013. Cited on page 25.

[GSZS14] Oshini Goonetilleke, Timos Sellis, Xiuzhen Zhang, and Saket Sathe. Twitter analytics. *ACM SIGKDD Explorations Newsletter*, 16(1):11–20, 2014. Cited on page 10.

[GTGMK+14] Ayelet Gal-Tzur, Susan M Grant-Muller, Tsvi Kuflik, Einat Minkov, Silvio Nocera, and Itay Shoor. The potential of social media in delivering transport policy goals. *Transport Policy*, 32:115–123, 2014. Cited on page 1.

[GWS13] Stefan Gindl, Albert Weichselbraun, and Arno Scharl. Rule-based opinion target and aspect extraction to acquire affective knowledge. pages 557–564, 2013. Cited on page 25.

[HAZ13] A. Hassan, A. Abbasi, and D. Zeng. Twitter sentiment analysis: A bootstrap ensemble framework. pages 357–364, Sept 2013. Cited on page 23.

[HBB13] Hussam Hamdan, Frederic Béchet, and Patrice Bellot. Experiments with dbpedia, wordnet and sentiwordnet as resources for sentiment analysis in micro-blogging. 2:455–459, 2013. Cited on page 22.

[HD10] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. pages 80–88, 2010. Cited on page 19.

[HF13] Viktor Hangya and Richárd Farkas. Filtering and polarity detection for reputation management on tweets. 2013. Cited on page 18.

[HL04] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. pages 168–177, 2004. Cited on page 25.

[HNP05] Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß. A brief survey of text mining. 20(1):19–62, 2005. Cited on pages 12 and 14.

[HZL13] Wu He, Shenghua Zha, and Ling Li. Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3):464–472, 2013. Cited on page 12.

[IK02] Hideki Isozaki and Hideto Kazawa. Efficient support vector classifiers for named entity recognition. pages 1–7, 2002. Cited on page 15.

[IK10] Utku Irmak and Reiner Kraft. A scalable machine-learning approach for semi-structured named entity recognition. pages 461–470, 2010. Cited on page 15.

[Iso01] Hideki Isozaki. Japanese named entity recognition based on a simple rule generator and decision tree learning. pages 314–321, 2001. Cited on page 15.

[JG10] Niklas Jakob and Iryna Gurevych. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. pages 1035–1045, 2010. Cited on pages 25 and 26.

REFERENCES

[JHS09] Wei Jin, Hung Hay Ho, and Rohini K. Srihari. Opinionminer: A novel machine learning system for web opinion mining and extraction. pages 1195–1204, 2009. Cited on pages 25 and 26.

[Kap12] Rianne Kaptein. Learning to analyze relevancy and polarity of tweets. 2012. Cited on page 18.

[KC13] Patrick Gage Kelley and Justin Cranshaw. Conducting research on twitter: A call for guidelines and metrics. 2013. Cited on page 12.

[KH10] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010. Cited on page 1.

[KMN+17] Tsvi Kuflik, Einat Minkov, Silvio Nocera, Susan Grant-Muller, Ayelet Gal-Tzur, and Itay Shoor. Automating a framework to extract and analyse transport related social media content: The potential and the challenges. *Transportation Research Part C: Emerging Technologies*, 77:275–291, 2017. Cited on page 1.

[Kom09] Nicos Komninos. Intelligent cities: towards interactive and global innovation environments. *International Journal of Innovation and Regional Development*, 1(4):337–355, 2009. Cited on page 8.

[KS12] Akshi Kumar and Teeja Mary Sebastian. Sentiment analysis on twitter. *IJCSI International Journal of Computer Science Issues*, 9(3):372–378, 2012. Cited on page 25.

[KWM11] E Kouloumpis, T Wilson, and J Moore. Twitter sentiment analysis: The good the bad and the omg! *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 11)*, pages 538–541, 2011. Cited on page 20.

[LAR12] Wendy Liu, Faiyaz Al Zamal, and Derek Ruths. Using Social Media to Infer Gender Composition of Commuter Populations. *Sixth International AAAI Conference on Weblogs and Social Media*, pages 26–29, 2012. Cited on page 9.

[LIR15] Carlo Lipizzi, Luca Iandoli, and Jos?? Emmanuel Ramirez Marquez. Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers' reactions to the launch of new products using Twitter streams. *International Journal of Information Management*, 35(4):490–503, 2015. Cited on page 10.

[LL16] Guy Lansley and Paul A Longley. The geography of twitter topics in london. *Computers, Environment and Urban Systems*, 58:85–96, 2016. Cited on page 55.

[LM14] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, 2014. Cited on page 34.

[LSP15] Thomas Ludwig, Tim Siebigteroth, and Volkmar Pipek. Crowdmonitor: Monitoring physical and digital activities of citizens during emergencies. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8852:421–428, 2015. Cited on page 9.

REFERENCES

[LWH+13] Yang Li, Chi Wang, Fangqiu Han, Jiawei Han, Dan Roth, and Xifeng Yan. Mining evidences for named entity disambiguation. pages 1070–1078, 2013. Cited on page 15.

[MAW16] Mojtaba Maghrebi, Alireza Abbasi, and S Travis Waller. Transportation application of social media: Travel mode extraction. In *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, pages 1648–1653. IEEE, 2016. Cited on page 60.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. Cited on page 34.

[MHK14] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014. Cited on page 21.

[MKWP+16] Sunghwan Mac Kim, Stephen Wan, Cécile Paris, Brian Jin, and Bella Robinson. The effects of data collection methods in twitter. *NLP+ CSS 2016*, page 86, 2016. Cited on page 32.

[MPLC13] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv:1306.5204*, 2013. Cited on page 32.

[MPP+13] Lev Muchnik, Sen Pei, Lucas C Parra, Saulo DS Reis, José S Andrade Jr, Shlomo Havlin, and Hernán A Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *arXiv preprint arXiv:1304.4523*, 2013. Cited on page 49.

[MSBX13] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. pages 889–892, 2013. Cited on page 19.

[MSLG15] Cataldo Musto, Giovanni Semeraro, Pasquale Lops, and Marco De Gemmis. CrowdPulse: A framework for real-time semantic analysis of social streams. *Information Systems*, 54:127–146, 2015. Cited on pages 1, 11, 20, 24, and 29.

[MWdR12] Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. Adding semantics to microblog posts. pages 563–572, 2012. Cited on page 17.

[MYZ13] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Hlt-naacl*, volume 13, 2013. Cited on page 34.

[OKA10] Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*, pages 384–385, 2010. Cited on page 33.

[PdRS12] Maria-Hendrike Peetz, Maarten de Rijke, and Anne Schuth. From sentiment to reputation. 2012. Cited on page 18.

[Phi12] Judah Phillips. *Social Media Analytics*, pages 247–269. John Wiley & Sons, Inc., 2012. Cited on page 10.

REFERENCES

[PP15] Maithilee L. Patawar and M. A. Potey. Approaches to named entity recognition: A survey. *International Journal of Innovative Research in Computer and Communication Engineering*, 3, December 2015. Cited on page 15.

[QYOP15] M. Atif Qureshi, Arjumand Younus, Colm O'Riordan, and Gabriella Pasi. *Company Name Disambiguation in Tweets: A Two-Step Filtering Approach*. Springer International Publishing, Cham, 2015. Cited on page 18.

[RDL10] Daniel Ramage, Susan T Dumais, and Daniel J Liebling. Characterizing microblogs with topic models. *ICWSM*, 10:1–1, 2010. Cited on page 58.

[Ril95] Ellen Riloff. Little words can make a big difference for text classification. pages 130–136, 1995. Cited on page 13.

[RL14] Jo Royle and Audrey Laing. The digital marketing skills gap : Developing a Digital Marketer Model for the communication industries. *International Journal of Information Management*, 34(2):65–73, 2014. Cited on page 10.

[RM99] L. A. Ramshaw and M. P. Marcus. *Text Chunking Using Transformation-Based Learning*. Springer Netherlands, Dordrecht, 1999. Cited on page 15.

[RMM+12] Haggai Roitman, Jonathan Mamou, Sameep Mehta, Aharon Satt, and L.V. Subramaniam. Harnessing the crowds for smart city sensing. *Proceedings of the 1st international workshop on Multimodal crowd sensing - CrowdSens '12*, (November):17, 2012. Cited on page 9.

[RR09] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. pages 147–155, 2009. Cited on page 15.

[RS10] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010. Cited on page 35.

[SAN07] Anna Stavrianou, Periklis Andritsos, and Nicolas Nicoloyannis. Overview and semantic issues of text mining. *ACM Sigmod Record*, 36(3):23–34, 2007. Cited on pages xiii, 12, 13, and 14.

[SDX13] Stefan Stieglitz and Linh Dang-Xuan. Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining*, 3(4):1277–1291, 2013. Cited on page 11.

[SFD+10] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short Text Classification in Twitter to Improve Information Filtering. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval SE - SIGIR '10*, (January 2010):841–842, 2010. Cited on pages 10, 11, and 14.

[SFI+13] Róbert Szabó, Károly Farkas, Márton Ispány, András A Benczur, Norbert Bátfai, Péter Jeszenszky, Sándor Laki, Anikó Vágner, Lajos Kollár, Cs Sidló, et al. Framework for smart city applications based on participatory sensing. pages 295–300, 2013. Cited on page 9.

[SGA13] Damiano Spina, Julio Gonzalo, and Enrique Amigó. Discovering filter keywords for company name disambiguation in twitter. *Expert Systems with Applications*, 40(12):4986–5003, 2013. Cited on page 16.

[SGS16] Pedro Saleiro, Luís Gomes, and Carlos Soares. Sentiment Aggregate Functions for Political Opinion Polling using Microblog Streams. *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering - C3S2E '16*, pages 44–50, 2016. Cited on pages 20 and 22.

[SIN13] SINTEF. Big data, for better or worse: 90last two years. Available at https://www.sciencedaily.com/releases/2013/05/130522085217.htm, May 2013. Cited on page 10.

[Spi14] Damiano Spina. *Entity-based filtering and topic detection For online reputation monitoring in Twitter*. PhD thesis, Universidad Nacional de Educación a Distancia, 2014. Cited on page 17.

[SRP+13] Pedro Saleiro, Lus Rei, Arian Pasquali, Carlos Soares, Jorge Teixeira, Fabio Pinto, Mohammad Nozari, Catarina Felix, and Pedro Strecht. POPSTAR at RepLab 2013: Name ambiguity resolution on Twitter. *CEUR Workshop Proceedings*, 1179, 2013. Cited on pages 14 and 18.

[SST+09] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. TwitterStand. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09*, (January 2009):42, 2009. Cited on pages xi, 2, and 11.

[TBP+10] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, December 2010. Cited on page 24.

[TWQ+14] Duyu Tang, Furu Wei, Bing Qin, Ting Liu, and Ming Zhou. Coooolll: A deep learning system for twitter sentiment classification. pages 208–212, 2014. Cited on page 23.

[TWY+14] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. pages 1555–1565, 2014. Cited on page 23.

[URS16] Daniela Ulloa, Rosaldo J. F. Rossetti, and Pedro Saleiro. A Framework for Open Innovation through Automatic Analysis of Social Media Data. 2016. Cited on pages 8, 10, and 14.

[VRLSM+12] Julio Villena-Román, Sara Lana-Serrano, Cristina Moreno, Janine García-Morera, and José Carlos González Cristóbal. Daedalus at replab 2012: Polarity classification and filtering on twitter data. 60, 2012. Cited on page 18.

[Yar95] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. pages 189–196, 1995. Cited on page 13.

[YMA10] Surender Reddy Yerva, Zoltán Miklós, and Karl Aberer. It was easy, when apples and blackberries were only fruits. 2010. Cited on pages 17 and 18.

[YMO+10] Minoru Yoshida, Shin Matsushima, Shingo Ono, Issei Sato, and Hiroshi Nakagawa. Itc-ut: Tweet categorization by query categorization for on-line reputation management. *CLEF (Notebook Papers/LABs/Workshops)*, 2010. Cited on page 17.

[ZCLL10] Daniel Zeng, Hsinchun Chen, Robert Lusch, and Shu-Hsing Li. Social Media Analytics and Intelligence. *IEEE Intelligent Systems*, 25(6):13–16, 2010. Cited on page 10.

[ZJW+11] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. pages 338–349, 2011. Cited on pages 19 and 20.