FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Social Media Text Processing and Semantic Analysis for Smart Cities

João Filipe Figueiredo Pereira

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Rosaldo José Fernandes Rossetti

Co-supervisor: Pedro dos Santos Saleiro da Cruz

June 27, 2017

Social Media Text Processing and Semantic Analysis for Smart Cities

João Filipe Figueiredo Pereira

Mestrado Integrado em Engenharia Informática e Computação

June 27, 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 43 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 investigação 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 43 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

First of all, my deep gratitude to my friends for being on my side when I was a bit down.

To my companions at Lab I120, João Neto, José Pinto, João Pedro Dias and Luís Reis (ρ7 Boyz): thank you for the funny moments during the whole dissertation period, specifically, during the tough process of writing up the document.

To my colleagues, specially, Henrique Ferrolho: thank you for the friendship, patience and support in these five long years. Now, I am sure that more challenges are coming to us which may imply distance but besides that I truly believe that in the future we still would cross paths at the professional or even academic course.

To Professor Rosaldo Rossetti and Pedro Saleiro, thank you very much for all support, dedication, enthusiasm and knowledge passed to me. During each task you defined in the dissertation period, I was able to improve myself in both academic and social levels.

To the institution that host me, Faculty of Engineering of University of Porto (FEUP), as well as to all of its docents that guide me during this Master's program, I am thankful for everything I have learn until now.

Last and more important, I would like to express my deep gratitude to my mother, Ana Brito, and my father, Júlio Pereira, for all the sacrifice and effort made to assure my future and concede me this opportunity to fulfil a dream: be graduated. I hope this achievement of mine make you very proud and I wish all success for both mine and your's ambitions and goals in the future. Like always, you know that you can count on me for everything you need.

João Pereira

"Life is too short for long-term grudges."

Elon Musk

Contents

1 Introduction **1**
1.1 Scope and Motivation . . . 1
1.2 Problem Statement . . . 2
1.3 Aim and Goals . . . 3
1.4 Document Structure . . . 3

2 Background and Literature Review **5**
2.1 Smart Cities . . . 5
2.2 Intelligent Transportation Systems . . . 7
2.3 Social Media Analytics . . . 8
2.4 Text Mining . . . 9
2.4.1 Topic Modelling . . . 11
2.4.2 Text Classification . . . 13
2.5 Related Social Media Frameworks . . . 15
2.6 Summary . . . 17

3 Framework **19**
3.1 Requirements . . . 19
3.2 Architecture Overview . . . 20
3.3 Data Collection . . . 21
3.3.1 Data Filtering . . . 23
3.4 Text Pre-processing . . . 25
3.5 Text Analytics . . . 26
3.5.1 Topic Modelling . . . 26
3.5.2 Travel-related Classification . . . 28
3.6 Data Storage and Aggregation . . . 29
3.7 Visualization . . . 30
3.8 Summary . . . 30

4 Exploratory Data Analysis **31**
4.1 Geographic Distributions . . . 31
4.2 Temporal Frequencies . . . 36
4.3 Content Composition . . . 40
4.4 Summary . . . 42

5 Experiments **45**
5.1 Topic Modelling . . . 45
5.1.1 LDA Model Parametrization . . . 46

5.1.2 Results and Analysis . . . 46
5.1.3 Final Remarks . . . 49
5.2 Travel-related Classification . . . 50
5.2.1 Rio de Janeiro and São Paulo . . . 51
5.2.2 New York City . . . 56
5.2.3 Concluding Remarks . . . 61
5.3 Summary . . . 61

6 Conclusions and Future Work **63**
6.1 Final Remarks . . . 63
6.2 Contributions . . . 64
6.3 Future Work . . . 65
6.4 Publications . . . 66

Acronyms **69**

Glossary **71**

References **73**

List of Figures

2.1 *Smart City* conjecture of four forces. Source: [Ang15] 6

3.1 Framework Architecture Overview 22
3.2 Example of our filtering method for geo-located tweets with variable bounding-boxes 24
3.3 Plate notation of Latent Dirichlet Allocation (LDA) by Blei et al. [BNJ03] . . . 27

4.1 Search Bounding-boxes for the data collection 32
4.2 Exploratory analysis in Brazilian cities 34
4.3 Exploratory analysis in English-speaking cities 35
4.4 Daily volume of tweets 37
4.5 Days-of-the-week box-plots for the volume of tweets 38
4.6 Hour-of-the-day box-plots for the volume of tweets 39
4.7 Log-log plots of users distribution 42

5.1 Day-of-the-week Twitter activity 49
5.2 ROC Curve of SVM, LR and RF experiences 53
5.3 Positive Predicted Tweets per Day of Week 54
5.4 Rio de Janeiro heat map to the positive tweets 54
5.5 São Paulo heat map to the positive tweets 55
5.6 SVM model with BoE(200) for each travel mode 59
5.7 Spatial density of the predicted tweets 60

LIST OF FIGURES

List of Tables

2.1 Text Mining Issues by Stavrianou [SAN07] 10
2.2 Brief overview of the related work for topic modelling 13
2.3 Brief overview of the related work for text classification - Best Experiments . . . 14

4.1 Collecting Bounding-boxes Coordinates (South-West and North-East) 32
4.2 Twitter Default Bounding-boxes Coordinates (South-West and North-East) 33
4.3 Datasets composition according bounding-box analysis 33
4.4 Volume of tweets for each type of geo-location 33
4.5 Percentage of Metadata composing the datasets 41

5.1 Datasets composition 46
5.2 Example of the topics classification 47
5.3 Final results of the LDA topics aggregation 47
5.4 Portuguese datasets composition for the travel-related classification experiment . 51
5.5 Travel terms used to build the training set 52
5.6 Performance results with 100 sized vectors for BoE 53
5.7 English-speaking tweets datasets 57
5.8 New York City First Experiment Results 57
5.9 *Leave-one-group-out expirement datasets* 58
5.10 *Leave one group out* experiments results for SVM, LR and RF classifiers 58
5.11 Sample of tweet messages correctly classified 60

Chapter 1

Introduction

1.1 Scope and Motivation 1
1.2 Problem Statement 2
1.3 Aim and Goals 3
1.4 Document Structure 3

1.1 Scope 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 (UGC) [KH10] 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 [MSLdG15].

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 MicroBlog 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 Natural Language Processing (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 to assure that all information is fully related to the target scenario in 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 guarantee 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 User Interface (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 Aim and Goals

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 SMC 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 and technologies from 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 Document Structure

The effort applied to this dissertation generated a great diversity of points and due to that the remainder of this document is organized as follows. Chapter 2 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 Chapter 3, being each its composing modules depth described. Experiments performed to test each module of the framework are reported in Chapter 5. We end the document with Chapter 6 where conclusions, future work and a few final remarks are exposed.

Chapter 2

Background and Literature Review

2.1 Smart Cities 5
2.2 Intelligent Transportation Systems 7
2.3 Social Media Analytics 8
2.4 Text Mining 9
2.5 Related Social Media Frameworks 15
2.6 Summary 17

This section aims the analysis and reflection about some works that has as final goal, similarly to ours, the development of a framework with the purpose of exploring social media data to extract meaningful domain-specific information. Nonetheless, studying works from other authors may help or even find already proposed solutions in order to solve the aforementioned problems.

Hence, this section will contemplate a brief contextualization about how can an intelligent system contribute to the improvement of a *smart city* or transportation services. Moreover, technologies and methods that allow extraction of information from a text document or, in this particular case, from tweets will be described. Finally, an exploration through already existent frameworks regarding the information extraction from social media content as well as the identification of its application domain.

2.1 Smart Cities

Smart City is a concept appeared thanks to the continuous growth of a city's population which contributed to an aggressive level of urban and technological developments [URS16]. In the last few years, several definitions for its meaning have emerged but its main idealization is not yet fully known [Kom09]. Angelidou [Ang15] defined Smart City as

"Conceptual urban development model on the basis of the utilization of human, collective, and technological capital for the development of urban agglomerations",

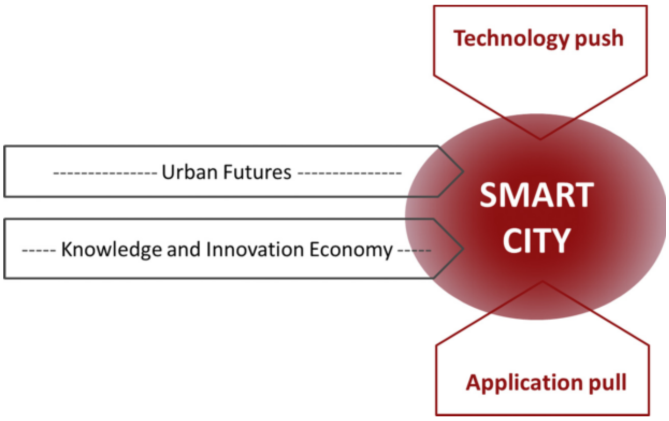


Figure 2.1: *Smart City* conjecture of four forces. Source: [Ang15]

enhancing *knowledge* and *innovation economy* as the primary factors that support the development of a city. Alongside with the previous factors, the author identifies other three distinct forces that shape the concept of a *smart city*:

1. *Technology Push*: The need of new products and solutions are introduced into the market due to a fast advance in science and technology.

2. *Demand Pull*: Current problems are solved originating new possibilities to respond society demands such as the continuous growth of the population.

3. *Urban Future*: Represents the final goal of a city constituting for that reason an important role in the whole transformation process.

4. *Knowledge and Innovation Economy*: The creation of new products using the most recent technologies is associated to solution for the efficiency and sustainability of a city.

The first two forces previous mentioned are directly dependent of the other ones as it is showed in Figure 2.1. However, the absence of desire to reach a better future having into consideration the city's economy and resources can result in the break of its dynamics and healthy, affecting services of a city due to the population discontentment.

The development environment of a city tagged as *smart* is another key factor to reach the success. Komninos [Kom09] associates collective sources of innovation to the improvement of life quality in cities. The globalization of innovation networks is responsible for the emergency of another types of environments and infrastructures, as so *"global innovation clusters and i-hubs, intelligent agglomerations, intelligent technology districts and intelligent clusters, living labs"* allowing the testing of products or services by the ordinary citizens in order to identify problems or even analyse their behaviour and reactions regarding what have experimented [Kom09]. Hence, it is possible to affirm that the development of a city has its starting point in the community but also depends on the quality of Information and Communications Technologys (ICTs) [Hol08], an essential requirement in the city's evolution process.

Last but not least, a *smart city* may focus its efforts in several sectors, such as the environment, culture and recreation, education, social and economic aspects, demography, and travel and transportation [CDBN11] in order to have equally advances in all of them.

2.2 Intelligent Transportation Systems

The transportation system is inherently connected to the progress of a city because people uses on a daily-basis transportation modes, i.e. bus, private cars, metropolitan, and others, in order to go to their jobs and make their own life and through that they contribute to the economic progress of it. Although this connection, such system is also influenced by the problem of population growth being relevant and necessary the finding of solutions to minimize or even erase 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 (ITS), it is necessary to introduce the meaning of Smart Mobility (SM). 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, their routes or even the knowledge of what citizens thought about the city's services [CL15]. The majority of ITS are expressed through smart applications where 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* and due to tha, they also serve as support to ITS, since the information provide by such infrastructures allows the piloting of activities such as traffic operations, as well as 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]. Szabo et al. [SFI+13] and Doran et al. [DSGD15] reported the highly economic cost to this kind of sensing, since it is require maintenance and replacement of this devices, as well as a tracking infrastructure to store and process the collected information. Hence, a new form of sensing has emerged - Crowdsensing or mobile crowdsensing - offering to the cities several ways to improve their services by exploring the participation of the citizens through social networks where there is a publicly sharing of opinions and thoughts regarding some problems around the city where they live are passing in [RMM+12]. This type of sensing consists in human-generated data provided by the population through the usage of mobile devices and social networks web-based platforms. Such data can be further used to extract some analytics regarding specific services in a city, namely the urban transportation system [RMM+12]. Having this considered, social media can be seen as a good source of data to extract valuable information aiming the direct use of it into the smartness evolution process of a city [SFI+13]. Recently, it is possible to verify that cities are increasingly opting for technological opportunities based on *crowd sensing*, once this type of exploration brings a considerable reduction of costs and support in the development of news valuable technologies.

In the last few years, several authors have published a widely range of social-media-based contributions focusing this specific domain. Kurkcu et al. [KOM16] use geo-located tweets to try and discover human mobility and activity patterns. The subject of transport modes was explored by Maghrebi et al. [MAW16] in the city of Melbourne, Australia. From a dataset of 300,000 geo-located tweets, authors tried to extract tweets related to several modes of transport using a keyword-based search method.

Additionally, there were also different efforts focused on the tracking of accidents using Twitter social media data. Mai and Hranac [MH13] tried to establish a correlation between the California Highway Patrol incident reports and the increased volume of tweets posted at the time they were reported. On the other hand, Rebelo et al. [RSR15] implemented a system capable of extract and analyse events related to road traffic, coined TwitterJam. In that study, authors also used geo-located tweets that were already confirmed as being related to events on the roads and compared their counts with official sources.

Performing robustness experiments over this domain is challeging since although the large number of recently publications, gold standards are yet not defined or even public being for this reason difficult to prove the methodology chosen or suppositions made. Maghrebi et al. [MAW16] enhances some terms related to the transportation domain, however they are limited and also very common ones. After a tough investigation work, it is worth noting a list produced by Gal-Tzur et al. [GTGMK+14] containing a large number of terms whose may serve as support for new scientific contributions using social media in studies of the transportation domain.

2.3 Social Media Analytics

In the last few years social networks have made impact on the business communications since users assumed the role of costumers through the publication of content on these networks, rising high levels of interaction between them, as well as with 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[1], Twitter[2] and other social networking websites 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 widely into this new communication technologies publicly sharing real-life events, their

[1] https://www.facebook.com/
[2] https://twitter.com/

opinions about certain topics and their on-time feelings in the network through a simple message [DDLM15].

Social Media Analytics (SMA) can be described as a type of digital analytics which focus is the study of interactions between, their opinions/thoughts, their own life, companies as so its products or services through the social media data. Such study provides important information to "analysts, brands, agencies or vendors" facilitating the generation of economic value to many organizations [Phi12]. In order to achieve the main goal of the SMA, companies focus their effort in the development automatic systems to make possible an easy collection, analysis, summarization and visualization of processed social media data establishing specific points about what is necessary to improved in their products [ZCLL10].

However the potential value that SMA can provide, Phillips [Phi12] enhance some important factors to be considered in the analytics process: (1) Users permissions; (2) Awareness/listening of real-time information; (3) Search mechanisms; (4) Text analysis methodologies and techniques; (5) Data access and integration; (6) System integration, customization and growth.

The previous mentioned factors will help during the identification and comprehension of possible necessary features in a social media analytics tool, as well as to establish potential parameters/metrics to test and evaluate such tool. Without careful conduction in the social media tool elaboration, for instance, use of a wrong technique of SMA could have a bad business impact for the company resulting possible bankruptcies and increase the unemployment tax of a city.

2.4 Text Mining

Text mining is a conjecture of fields such as information retrieval, data mining, machine learning, statistics and computational linguistics which aims the extraction of valuable information from unstructured textual data [HZL13]. The intensively usage of this analysis methodology is due to the massive amount of information stored in text documents being necessary automatic techniques to identify, extract, manage and integrate the knowledge acquired from these texts exploration in a efficiently and systematically way [ACK+05]. On the other hand, the emergency of social media applications have also contributed to the widely growth of text mining usage because of the "application's perspective and the associated unique technical and social science challenges and opportunities" [ZCLL10].

Text mining shares some of the issues presented by the Natural Language Processing (NLP) field. Texts are usually performed by humans and due to that, some problems in its construction can appear, such as spelling mistakes, wrong phrasal construction, slang among other. Before the mining process of a text, it's important to apply some preprocessing steps in order to eliminate or, at least reduce, undesired content (words) in the primary analysis process. A. Stavrianou et al. [SAN07] cite these issues very well alongside their work and some of them are observable in Table 2.1.

The removal of words from text may sometimes not be desirable because some sentences can lose its information or even leads to a different meaning compared with its original form. The

Table 2.1: Text Mining Issues by 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?
Part-of-speech 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?

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 speaking language of the text than its specific domain [SAN07]. The main goal of this technique is to reduce a word to its root form helping in the calculus of distances between texts, keywords or phrases, or even in the text representation.

The noisy data is derived from spelling mistakes, acronyms and abbreviations in texts and to solve this, a conversion of these terms should be done to maintain the integrity of data. Commonly solution approaches involve text edit distances (Levenshtein Distance[3]) and phonetic distances measures between known words and the misspelling ones to achieve good corrections [BDF+13]

Word Sense Disambiguation (WSD) focus on solving the ambiguity in the meaning of a word. 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[4] is a commonly used resource to extinguish this ambiguity [CSMA16]. There are two types of disambiguation: the unsupervised, where the task is support by a dictionary or a thesaurus [SAN07]; and, the supervised one, where different meanings of a word are unknown and normally learning algorithms with training examples are used to achieve good results regarding the performance of the disambiguation task [Yar95].

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, and others [HNP05]. Collocations are

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

groups and constitutes a very important step in some text mining approaches. Grouping two or more words to give its correct meaning is sometimes crucial to perform tasks such as sentiment analysis where negations (e.g. "don't like") needed to be composed by two or more words in order to assure the negative value of, for example, a verb. Collocations 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 splitting its content 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 term-frequency vector according the occurrence of a word 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].

Once been introduced some of the most preliminary important steps in text mining, the remainder subsection are focused in two different text analytics approaches: topic modelling and text classification. The majority of Social Media Analytics approaches focus its efforts in modelling and classification tasks in order to understand the large range of data collected and support commonly used techniques to extract information from it, such as sentiment analysis, trend analysis and topic modeling [FG13].

2.4.1 Topic Modelling

Topic modelling is a text mining unsupervised technique/method aiming the identification of similarities in unlabeled texts. Usually, this technique is applied over texts of large volume since to correctly identify the resulting patterns in its content requires the existence of lots of information.

One of the first studies made using Twitter data was proposed by Kwak et al. [KLPM10] and consisted in the collection of messages to classify the trends in its content. Results showed that almost 80% of the trends in Twitter are related to real-time news and the period in which each trend maintains itself in the top is limited. The authors proved that Twitter can be seen as a mirror of real-time occurring events/incidents in the world.

Several works were already proposed to identify social patterns in the population daily-basis life and mapping such patterns geographically by topic modelling techniques to discover latent topics in social media streams. Usually, studies about topic modelling, in particular LDA model, to text mining problems follow unsupervised approaches [LL16, OPST16] - where is not required the creation of a training dataset. Others improved the model and made it an supervised approach [RDL10], dependent of training data, and compare to the traditional one in order to prove better results.

Using entity-centric aggregations and topic modelling techniques, Oliveira et al. [OPST16] built a system focused in data visualization that allows an user to search for an entity during a specific period and shows which are the main topics identified in the Twitter messages. Ordinary weekday patterns were identified by Lansley et al. [LL16] in their study regarding the inner region of London. The authors used a LDA model to distribute 20 topics over 1.3M tweets. After crossing the results of the experiment with land-uses datasets it was possible to observe interesting patterns in specific zones and places of the British city. Nonetheless, Ramage et al. [RDL10] improved a LDA model by adding a supervised layer that automatic label each tweet used in their experiment.

Conditional Random Fields (CRF) are explored by Nikfarjam et. al [NSO+15] which have applied word embeddings in combination to other text features, such as adverse drug reactions lexicons, POS tagging and negation collocations in order to train a supervised model. Such model was able to demonstrate high performances on the extraction of concepts/topics from the social media user-generated content. To prove robustness and efficiency in the model, authors have compared the obtained results with DailyStrength corpus and were able to notice that due to the limited size of text in a tweet, the detection of different reactions about drugs is more complex, which could be simplified with access of greater amount of information provided in the training process of the model.

Differently from the majority of works involving topic modelling techniques, Tuarob and Tucker [TT15] take support of a LDA model to extract the most frequent words for groups of tweets previously collected. The overall work is focused in sentiment analysis approaches and aims the perception of what people fells about a specific product as well as its composing features. Authors used the LDA model to find what were the main 2 topics present in each product set of tweets and considered the most frequent 30 words. Moreover, POS tagging, disambiguation and stemming techniques were used in order to filter out and normalized words related to the product. Finally, an unsupervised method to calculate the sentiment polarity was applied to the data being final results coherent to the product feature/aspect extracted.

Topic modeling techniques consisting in supervised learning approaches were explored by Zhang et al. [ZNHG16], where authors have compared the results obtained from a SVM classification for accident-related tweets with a classification using a two-topic generative model SLDA (Supervised LDA). Contrarily to the unsupervised method, this one takes into consideration the label assigned to the training examples and can be trained as a genuine classification model. By comparing the final results between both models, it is possible to observe a significative increase of the precision and a decrease of only 0.04 points in the accuracy meaning that supervised topic modelling techniques to binary classification may compete well with conventional classification models, with respect to tweets.

Probabilistic topic models, such as LDA, are the most used techniques in topic detection tasks. Although high applicability, authors question themselves regarding the performance of this technique over social media data which present limitations, starting at the size of the message and ending in the bad phrasal construction and informality [MSBX13]. In this dissertation we will tackle this question and try to answer it by presenting results obtained in a real-world scenario

Table 2.2: Brief overview of the related work for topic modelling

Approach	Features	Methods	Goal	Potential Domain
H. Kwak et al. [LL16]	Twitter metadata	Aggregation of trending topics using external information	Quantitative study in order to reveal Twitter as both social media and news media platform	Smart City
J. Oliveira et al. [OPST16]	specific-entity words	Unsupervised Latent Dirichlet Allocation	Extract the most relevant entity-related topics	Smart City
G. Lansley and P. Longley [LL16]	Bag-of-words	Unsupervised Latent Dirichlet Allocation	Study social dynamics of London using Twitter topics	Smart City
S. Tuarob and C. Tucker [TT15]	Bag-of-words	Unsupervised Latent Dirichlet Allocation	Extraction of people's polarization sentiment about a specific feature of a product (aspect sentiment analysis)	Smart City - Economy
D. Ramage et al. [LL16]	Labeled bag-of-words	Supervised Latent Dirichlet Allocation	Proving the applicability of supervised approaches in conventional LDA model	Smart City
Z. Zhang et al. [ZNHG16]	Labeled bag-of-words	Supervised Latent Dirichlet Allocation [MB08]	Comparing performances with SVMs models to accident-related tweets	Smart City - Travel and Transportation
A. Nikfarjam et. al [NSO+15]	ADR Lexicons, POS Tagging Negation, Word Embeddings	CRF	Discrimination of adverse drug reactions in tweets content	Smart City - Health

experiments.

2.4.2 Text Classification

Text classification is a text mining task which main goal is the discrimination or characterization of a piece of text into a specific format value. Such value can vary from number (sentiment analysis tasks), labels (multi-labeling tasks), classes (binary or multi-class tasks). Classification in text analysis is a widely used methodology and had already been reported in several scientic contributions regarding the smart cities and transportation domains.

Suport Vector Machines (SVM), Ordinary Least Squares (OLS), Random Forests (RF), Multilayer Perceptron (MLP), Naïve Bayes (NB) and Decision Trees J48 (DT J48) are some of the supervised classification models used to analyse social media data over fields such as health [SSP11] and pharmacovigilance, political opinion [SGS16], transportation (travel classification [CSR10, KMN+17], traffic and incidents detection [ZNHG16]), financial sentiment analysis [SRSO17] and *online* reputation monitoring [SMRSO15].

Sifnorini et al. [SSP11] reported a study which main goal was the tracking of the disease Influenza A virus. Tweets collected by the authors using term-based search sum up more than 300 million examples. Their methodology consists in training SVM models with sets of frequency features composed by the most used weekly-terms over the whole dataset. Each model was specifically trained according a certain set of keywords and follow an iterative process, i.e. authors firstly have classified all illness-related tweets related and than used the resulting related subset of data to perform new classification regarding specific keywords, such as what was the disease source, countermeasures used and infected people characteristics. Final results allowed the verification of a decrease of Twitter activity while more new cases were appearing meaning less concerning about this epidemic through time.

Accident-related classification for Twitter data was proposed by Zhang et al. [ZNHG16]. Authors explored the Twitter Streaming API to collect geo-located tweets from Northern Virginia during a completed year, January to December of 2014, and recurring to auxiliary loop detectors that are, in intervals of 15 minutes, recording the traffic flow. In order to automatize the detection of accidents in that interval of time (were the sensors are not recording the scene), authors have built a binary classification model using Linear SVM with a balanced dataset composed by 400 training examples for each of the accident-related and non-related classes composed by a boolean-vectors according the final 3,000 tokens resulted from the token filtering and stemming process. Performance was improved by submitting the model to a 5-fold cross validation which was proved by values of accuracy and precision over than 70% of success.

Considering the task of discriminate travel-related tweets, Carvalho et al. [CSR10] have constructed a bag-of-words dependent classification model and achieved improvements at the model's performance with support of a bootstrapping approach implying a two phases train to the SVM model. By assuming the similarities, i.e. all four works were related to binary text classifications, we can induce an hypotheses that Linear SVM models have superior performances relatively to other models for this type of classification tasks.

Multi-class classification models were also applied to the transportation domain through text analysis of social media content. Kufliket et al. [KMN+17] build multiple classification models using methods such as NB and DT J48 to predict multiple modes of transport during three different sports events. Tweets sum up a total of 3.7M and were submitted to the models classification task in order to prove that an harvesting automatically information from SMC is possible and may help transportation entities in the planning and management of their services during social occasions as it is demonstrate in theirs use cases.

On the other hand, Saleiro et al. [SGS16] tried to predict the 2011 Portuguese bailout results analysing opinion within the tweets about all five political parties candidates. The opinion was measure using a OLS model trained with specific sentiment aggregate functions and proved to be capable of correctly predict who would be elected prime minister of Portugal only exploring sentiment analysis in social media data. In SemEval-2017 Task 5, Saleiro et al. [SRSO17] explored word embeddings techniques to extract the sentiment polarity and intensity in financial-related tweets. Authors have proved good performance of models trained with bag-of-words and bag-of-embeddings features together although the approach been applied to a specific domain. The usage of features representing syntactic and semantic similarities of texts, such as word embeddings, can be seen with great potential namely to the area of travel-related text classification.

Table 2.3: Brief overview of the related work for text classification - Best Experiments

Approach	Features	Classification Methods	Goal	Potential Domain
Sifnorini et al. [SSP11]	Bag-of-words	Linear SVM	Tracking the evolution of public sentiment and increasing of social media activity about the H1N1 pandemic	Smart City - Health
Zhang et al. [ZNHG16]	Boolean vectors matrix (3,000 different tokens)	Linear SVM	Improve transportation control by automatic discriminate accident-related tweets	Smart City - Travel and Transportation
Kuflik et al. [KMN+17]	Bag-of-words	Naïve Bayes, DT J48	Multi-class mode of transport classification and the purpose behind it	Smart City - Travel and Transportation
Carvalho et al. [CSR10]	Bag-of-words	Linear SVM with Bootstrapping	Discrimination of travel-related tweets	Smart City - Travel and Transportation
Saleiro et al. [SGS16]	Sentiment Aggregate Functions	OLS	Predicting Portuguese polls results through opinion mining	Smart Cities - Government
Saleiro et al. [SRSO17]	Word Embeddings, Bag-of-words, domain-specific lexicons	RF	Extraction of sentiment polarity and intensity from social media content and web news	Smart City - Economy

There is a wide diversity in text classification approaches. A worth noting fact in this review at the literature is that word embeddings have been supporting conventional techniques in order to improve performances in text classification tasks. Transportation domain lacks in studies having this particular feature in the training process of its classification models. Hence, it is of major importance perform experiments about this domain aiming conclusions and additional content to support the potential advantages brought by word embeddings.

2.4.2.1 Classification Evaluation Metrics

In order to measure the performance of a text classification model, there are several types of metrics that can help this process, depending of course the context of the task. Regarding binary classification tasks, the most common evaluation metrics used are precision, recall (sensitivity) and F1-score which is the harmonic mean or the weighted average of the previous two. Therefore, it is described each of these metrics as well the mathematical equation used in its calculation.

- **Precison:** Represents the fraction of correct predictions for the travel-related class (Equation 2.1).
- **Recall:** Represents the fraction of travel-related tweets correctly predicted (Equation 2.2).

$$Precision = \frac{tp}{tp + fp} \tag{2.1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2.2}$$

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.

$$F1_{score} = 2 * \frac{precision * recall}{precision + recall} \tag{2.3}$$

These first three metrics only showed us the performance of the classifier for a discrimination threshold of 0.5. The Receiver Operating Characteristic (ROC) curve gives us the True Positive Rate (TPR) and the False Positive Rate (FPR) for all possible variations of the discrimination threshold. Through the ROC curve, it is possible to compute the Area Under the Curve (AUC) to see what was the probability of the classifier to rank a random positive higher than a random negative one.

2.5 Related Social Media Frameworks

In the last few years, the number of proposals of frameworks to treat social media content and produce valuable information to the end-users has widely increased. For instance, each framework has it own domain of application and generalization is not the center focus. Event detection, *online* reputation monitoring, socio-semantic analysis to human reactions and traffic sensing are some of the application domains that research community present their contribute through framework proposals.

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 extracted 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. Author decided to implement a system that produce automatically analysis since they have find interesting results in the experiment conducted.

Regarding the field of event/incident detection, Abel et al [AHH+12] developed Twitcident, a real life accidents-aware web-based framework that is connected to a emergency broadcast system in order to detect incidents across the world. Then, an automatically system starts the collection and filtering of content from social media platforms and extracts information about entities using Named Entity Recognition and Disambiguation techniques. Data temporal distributions are also produced to analyse the time line of the events.

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.

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.

Traffic sensing over the city of Rio de Janeiro, Brazil, was studied by Rebelo et al. [RSR15] which have implemented a system capable of extract and analyse events related to road traffic, coined TwitterJam. In that study, authors used geo-located tweets that were already confirmed as being related to events on the roads and compared their counts with official sources. Finally but not least, authors present interesting geographic visualizations to the end-users in order to understand what is the current traffic-state of a certain road.

Social Media is used by Ludwig et al. [LSP15], in a framework that attempts the creation of voluntary and emergency activities, coined CrowdMonitor. The systems allows through the analyse of human mobality through tweets posted in the platform. Although absence of text analysis methodologies, such system intents to promote more cooperation between citizens and also promotes the applicability of crowd sensing, a crucial factor for the smartness evolution of a city.

Technological companies is the main target of the framework proposed by Lippizzi et al. [LIR15]. The system analyses social media content having in consideration specific products, such as mobile phones, tablets and others, and tries to extract information of what their customers think and talk about it. By measuring the sentiment of word clusters produced by the system, companies may take profit and additional insights about what in needed to be improved in their products.

CrowdPulse is a domain-agnostic framework proposed by Musto et al. [MSLdG15] which main objective is the presentation of text analytics to the end-users. Such framework is rich regarding implemented text methods, which range from entities disambiguation to sentiment analysis. Authors followed unsupervised approaches to implement all the framework composing methods, and applied the resulting system in two real-world scenarios, the earthquake of L'Aquila city and The Italian Hate Map. Further analysis of the results proved that simple techniques can provide faster insights about people sentiment regarding any type of domain.

A full-based text mining framework for *online* reputation monitoring is proposed by Saleiro et al. [SMRSO15] cabable of explore and extract multiple types of information from a wide range of Web sources. TextRep is divided in several modules in order to perform correctly the different text mining techniques, such as the collection of data, disambiguation and sentiment analysis. The system is adaptable to different domains as well and applications of it to political opinion mining and financial sentiment analysis are two of the use cases presented by the authors.

2.6 Summary

The literature review shows positives and negatives points that are necessary to be reported. First, the conceptualization of a meritorious system capable of bringing value to the smartness evolution of a city is a labourious and time-consuming process. Although iterative steps, it is necessary the stipulation of a detailed work-plan and what are/is indeed the final target/s and objectives of such system. Crowd sensing is a type of sensing that enables the study of what citizens think about a specific topic, and social media platforms can easily be explored in order to take its content to futher analysis and support the construction of a adaptable and profitable tool for the city's entities. Nowadays, text mining techniques allows the extraction of information from social media content, which can be represented, after accurate aggregations on the results, in visualization views facilitating analysis by the end-users of these systems. Last but not least, we could identify two unexplored approaches in this literature. Word embedding is a technique which has not been applied to transportation domain using social media content. Domain-agnostic frameworks using supervised learning methods are an hard task regarding its conception, however, due to the learning phase, models could learn new similarities from the text, and we see potential in this approach since it is not necessary construction of auxiliar dictionaries to perform the desired tasks.

Chapter 3

Framework

3.1 Requirements 19
3.2 Architecture Overview 20
3.3 Data Collection 21
3.4 Text Pre-processing 25
3.5 Text Analytics 26
3.6 Data Storage and Aggregation 29
3.7 Visualization 30
3.8 Summary 30

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.

3.1 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 [MSLdG15].

Social media data is mostly represented by text messages being necessary the application of 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 and characterization of the topic associated with a tweet. Each technique is represented as an independent module whose belongs to the boundary of text analytics. This framework needs also to 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.

The construction of this complex system requires careful planning since there are dependency between a task and the one that follows it, at least with respect to the filtering and preprocessing of data. Adaptability to different languages is considered and further addiction of new ones may be possible. For the same reason, but this time regarding new functionalities, the framework needs to follow a modular architecture allowing new text analysis layers as well as other type of data visualizations. The domain of *smart cities* is vast in terms of indicators and fields that constituting it and due to that, the final architecture may be designed in a way that allows configuration about the user's field of interest if he do not desire analytics visualization from all fields.

3.2 Architecture Overview

The framework proposed in this dissertation is divided into six different modules: (1) collection; (2) filtering; (3) preprocessing; (4) text analytics; (5) aggregation and (6) data visualization.

The current collection module is implemented to retrieve geo-location tweets from a specific bounding-box, however if the user demands, multiple locations can be explored at the same time. Other collection heuristics are also available, such as the keyword-search and users following. Depending on the target scenario and analytics to be explored, these two heuristics will need to be added in the module. This detail was considered during implementation period and flexibility was assured into the module composition.

Filtering tasks are directly related to `locations` heuristic of the collection module. Since this framework is designed to analyse cities or specific regions/zones of it, it is necessary guarantee if a tweet is actually inside of the searching bounding-box in order to do not induce information in the analysis from places far away of the target location. If other heuristics will be implemented, the filtering module can be configured to support other filtering-specific operations.

The preprocessing module is a module that has into consideration the future task in the framework. Having this considered, we implemented a segmented pipeline allowing the user a definition of the desired tasks he wants to analyse in the text messages since different text analysis may have different operation in the preprocessing routine. Methods implemented here are carefully described in Section 3.4.

Text analytics module is composed by two different sub-modules, both of them focusing in a specific text analysis method. Travel-related classification of tweets for two different speaking languages is available since one of the final goals regarding domain-agnostic framework is its adaptability into different scenarios and the language of texts constitutes one of them. Topic Modelling sub-module is available as a text analytics method provided by the framework. We trained a model over a sample of tweets and characterize each topic generated in order instantly characterize future tweets by only being necessary passing it over the transformation process to have their topic identified. In terms of generalization, the main module, text analytics module, was construct following adaptability and flexibility approaches to, in the future, new analysis be integrated.

By adding new functionalities, new aggregations are required in order to present the specific-task final results to the end-user. The aggregation module is structured into integrative methods facilitating future extensions or updates on it. Last but not least, aggregation results are communicated to the visualization module, where, similar to other modules, it is possible the inclusion of new data visualization charts, according to the new integrated functionalities.

3.3 Data Collection

In Section 3.1, 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 Application Programming Interfaces (APIs). The Search API is based on the Representational State Transfer (REST) protocol and only looks up for tweets published in the last 7 days, while the Streaming API creates basic endpoints (independent of the RESTful endpoints) and retrieves up to 1% of the Twitter Firehose. Regarding the proposed and developed framework, we chose the Streaming API due to its free-access for the community, smooth integration in the module implementation and due to the availability of real-time information. 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 the retrieval of tweets located inside a bounding-box [MKWP+16]. There are

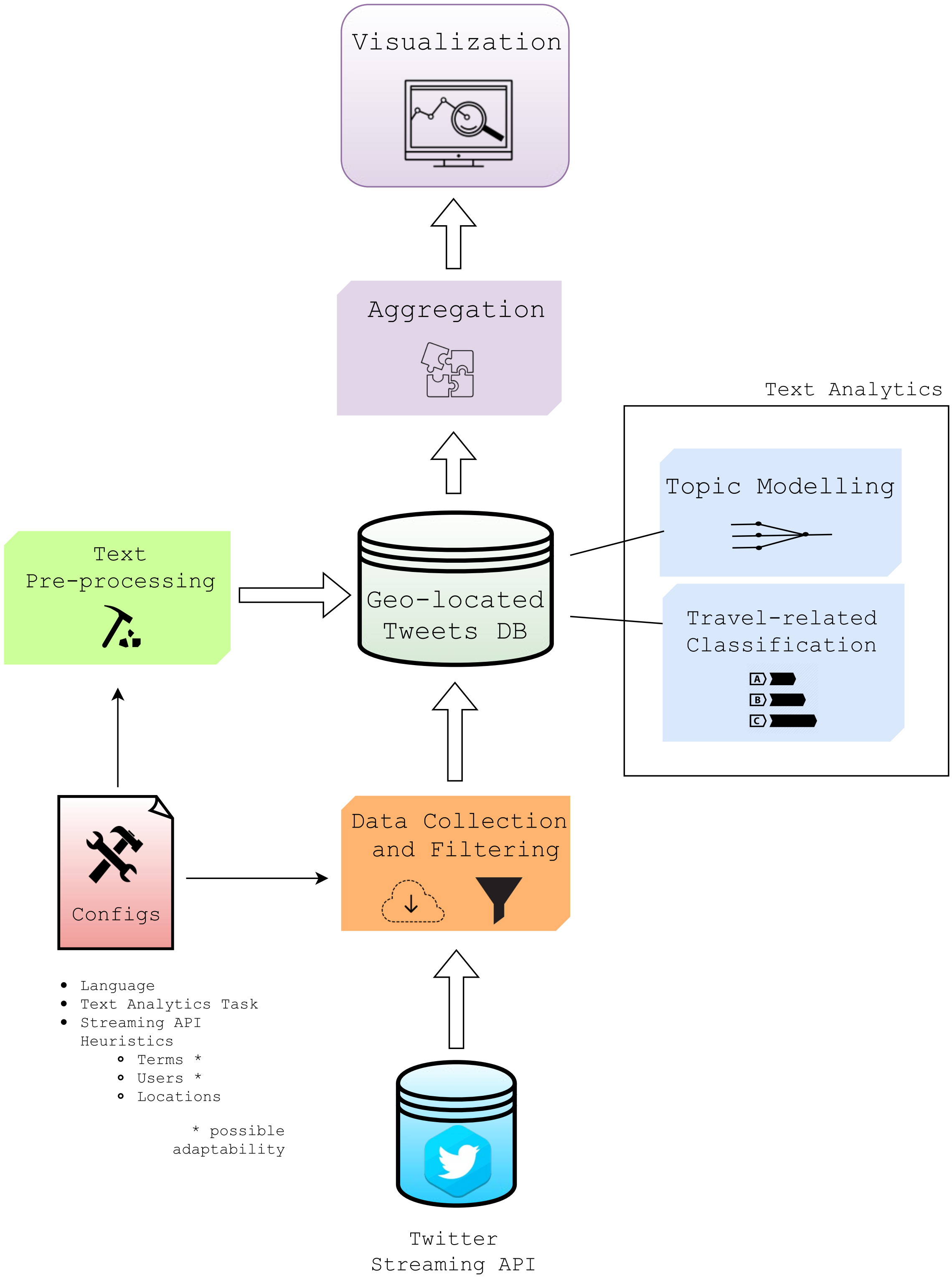


Figure 3.1: Architecture overview of the implemented framework

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[1]; 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 [2], which is integrated with the Google Maps API and allows an user to create a bounding-box for an existing place in that 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 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 so, we would need to be aware of 90% of the world, which is not the case.

3.3.1 Data Filtering

In the first attempts to study the data collected geographic distribution, we discover that not all tweets had a precise coordinate attached to it. Nonetheless, there were cases where tweets from other cities were collected by our crawler 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

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

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

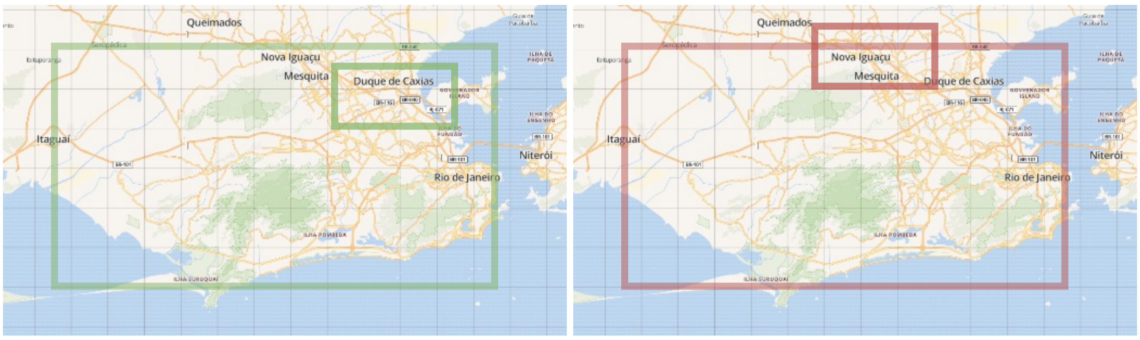
can activate the Global Positioning System (GPS) in the mobile application and associate to the tweet his precisely geo-location; (2) a user can choose a place from a predefined list provide by Twitter and associate the place to the tweet.

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 considered, it was necessary to understand how the Twitter Streaming API works and what kind of heuristics follows in order to retrieve such type of tweets. The documentation [3] 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 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 instance, if a tweet has a place such as Portugal and our filter bounding-box is defined for Porto, all tweets from place Portugal will be in our dataset, regardless the fact some tweets are posted elsewhere, such as in the city of Lisbon, very far away from Porto.



Positive Matching   Negative Matching

Figure 3.2: Example of our filtering method for geo-located tweets with variable bounding-boxes

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 in study appealing to the tweets *place* field. Such information was then used as the limited area in order to filter out tweets which *coordinates* field was not populated. The methodology behind the filtering process consists in the matching of the Twitter default bounding-box of the city against all places'

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

bounding-boxes in tweets. In Figure 3.2, we illustrate an example of our method in which the green color represents the matching of a tweet attached with place Duque de Caxias yielding a positive result, while the red color represents a tweet with place Nova Iguaçu yielding a negative match result with the Twitter default bounding-box for the city of Rio de Janeiro.

3.4 Text Pre-processing

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 (3.1), 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 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:

- **Lowercasing:** 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 casing in one letter, its representation/interpretation by text mining techniques may be disparate.

  Travel-related Classification and ***Topic Modelling*** modules explore this pre-processing operation.

- **Lemmatization:** Only plural words are transformed into singular ones (e.g. cars -> car).

  Topic Modelling module explores this pre-processing operation.

- **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] from Python's Natural Language Toolkit (NLTK), perform poorly and are not capable of dealing with *#hashtags*, *@mentions*, abbreviations, strings of punctuation (e.g. `...` or `%&$`), *emoticons* (e.g. `:)` or `:-)` or `=D`) 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 O'Connor et al. [OKA10], which is capable of dealing with these special characteristics of tweets.

  Topic Modelling module explores this pre-processing operation.

- **Transforming repeated characters:** Sequences of characters repeated more than three times were transformed, e.g. "loooool" was converted to "loool".

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

Travel-related Classification and ***Topic Modelling*** modules explore this pre-processing operation.

- **Punctuation removal:** Every punctuation symbols are removed from the text message, including the previous mentioned *emoticons*.

  Topic Modelling module explores this pre-processing operation.

- **Cleaning *Entities* and Numerical Symbols:** Removing *URLs*, user mentions, *hashtags* and digits from the text messages.

  Travel-related Classification and ***Topic Modelling*** modules explore this pre-processing operation.

- **Stop and short 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 for the stop words removal task. Other type of words, such as 'kkk' or 'aff' represent short words that do not bring any valuable information from the message analysis. For this reason, we conceive a short dictionary containing these words and removed it from the message.

  Travel-related Classification and ***Topic Modelling*** modules explore this pre-processing operation.

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` [5] 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.

3.5 Text Analytics

The extraction of information from texts can vary in several types depending on the task performed to achieve it. In this dissertation, we explored two different types of analysis to the tweets: topic modelling and travel-related classification.

3.5.1 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

[5] https://pypi.python.org/pypi/pytz

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 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.

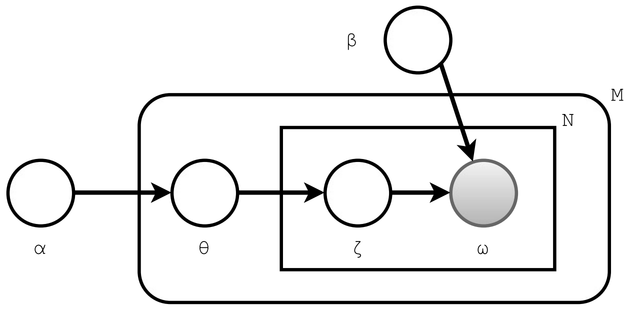


Figure 3.3: Plate Notation of the graphical model representation of Latent Dirichlet Allocation by Blei et al. Source: [BNJ03]

In Figure 3.3 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, and for this reason the task of topic modelling becomes very simple since only the frequency of words in 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 5.1 as well as potential improvements to the generated model.

3.5.1.1 Features

Topic modelling requires, like in other learning model, a group of features to be trained. In this case, we used the 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 Python's NLTK stop words list for the specific language in analysis. The minimal occurrence value for a word being considered was set to 10.

3.5.1.2 LDA Model Resulting Topics

The final model used in the implementation of our framework is defined to characterize a tweet into 50 different topics. Although that, in the experiment made to comprove the added-value brought by the model, we were obligated to cluster some of the topics due to the similarity presented in words constituting them. The final list of possible topics can be seen in Section 5.1.2, more specifically in Table 5.3.

3.5.2 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 it was created 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 produced 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 (BoW) features with Bag-of-embeddings (BoE) (word embeddings representation matrices), producing, for the best of our knowledge, the first travel-related classification model with both type of features.

3.5.2.1 Features

BoW 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. In this group of features, we only considered uni-grams as the basis of text representation form. The final dictionary of this form was produced with the 3,000 most frequent terms across the training set excluding the ones found in more than 60% of the documents (tweets).

The technique of word embeddings is used by 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 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:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log P(w_{t+j}|w_t) \tag{3.1}$$

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, Le and 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. Hence, we explored *paragraph2vec* to learn the vector representations of each tweet and tried several configurations in the model hyper-parameterization.

Using *paragraph2vec* [LM14], we created BoE representation matrices for the tweets in order to explore the learning distributed representations of words where each word is represented by a distribution of weights across a fixed number of dimensions. Mikolov et al. [MYZ13] proved that this kind of text representation is robust when encoding syntactic and semantic similarities in the embedding space. The training process of our classification models involved 10 iterations over the datasets using a context window of value 2 and feature vectors of 50, 100 and 200 dimensions. Then, the corresponding embedding matrix yielded the group of features fed into our classification routine.

Both previous described methods are available in the collection of Python scripts we used in this dissertation, coined `Gensim` [6], presented and lately improved by Řehůřek and Sojka [RS10].

The overall experiments regarding the travel-related classification of tweets are described and detailed in Sections 5.2.1 and 5.2.2. Concluded the experiments, we select the best classifiers for each case and used it in the implementation of the framework's travel-related modules allowing discrimination of potential new tweets related to the transportation domain.

3.6 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 data warehouse technology for our framework. MongoDB allows storage of JavaScript Object Notation (JSON) 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

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

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.

3.7 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 HyperText Markup Language (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.

3.8 Summary

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.

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

Chapter 4

Exploratory Data Analysis

4.1 Geographic Distributions . . . 31
4.2 Temporal Frequencies . . . 36
4.3 Content Composition . . . 40
4.4 Summary . . . 42

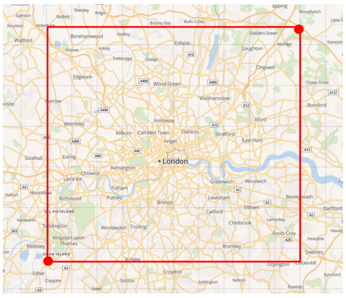
The main goal of this chapter is the devise of relevant analysis taking into consideration the five different collected datasets. Since this dissertation is supported in experiments using real-world data, such analysis is crucial in order to gain better knowledge of the intrinsic characteristics of it. 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.

4.1 Geographic Distributions

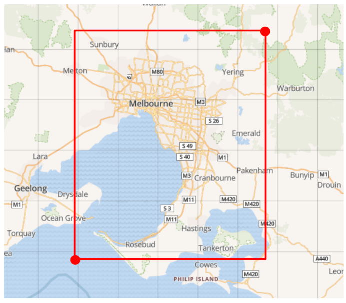
As previously mentioned, in Section 3.3, we exploit an auxiliary *online* tool to generate the coordinates for the bounding-boxes used in the collection process. The visual representation of the each city bounding-box is illustrated in Figure 4.1, as well as its the corresponding coordinates which are presented in Table 4.1.

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

(a) São Paulo



(b) London



(c) Melbourne

(d) Rio de Janeiro

(e) New York

Figure 4.1: Search Bounding-boxes for the data collection

To conduct the data filtering process, we extracted from data the Twitter default bounding-boxes for each city in study, being possible to observe their corresponding South-West and North-East coordinates in Table 4.2. The map visualization of these bounding-boxes is demonstrated in Figures 4.2 (subfigures 4.2b and 4.2a) and 4.3 (subfigures 4.3a, 4.3b and 4.3c), where the biggest rectangle represents the Twitter default bounding-boxes for each city.

The final volume of tweets located inside and outside the cities correspondent bounding-boxes are presented in Table 4.3. 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 (4.3) 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 and New York City, 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

Table 4.1: 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)

Table 4.2: 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 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 [1] has almost 12 millions habitants while Rio de Janeiro [2] has 6 million. Having only this amount of information it is impossible, at the moment, formulate a explanation to this phenomenon.

Table 4.3: 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%

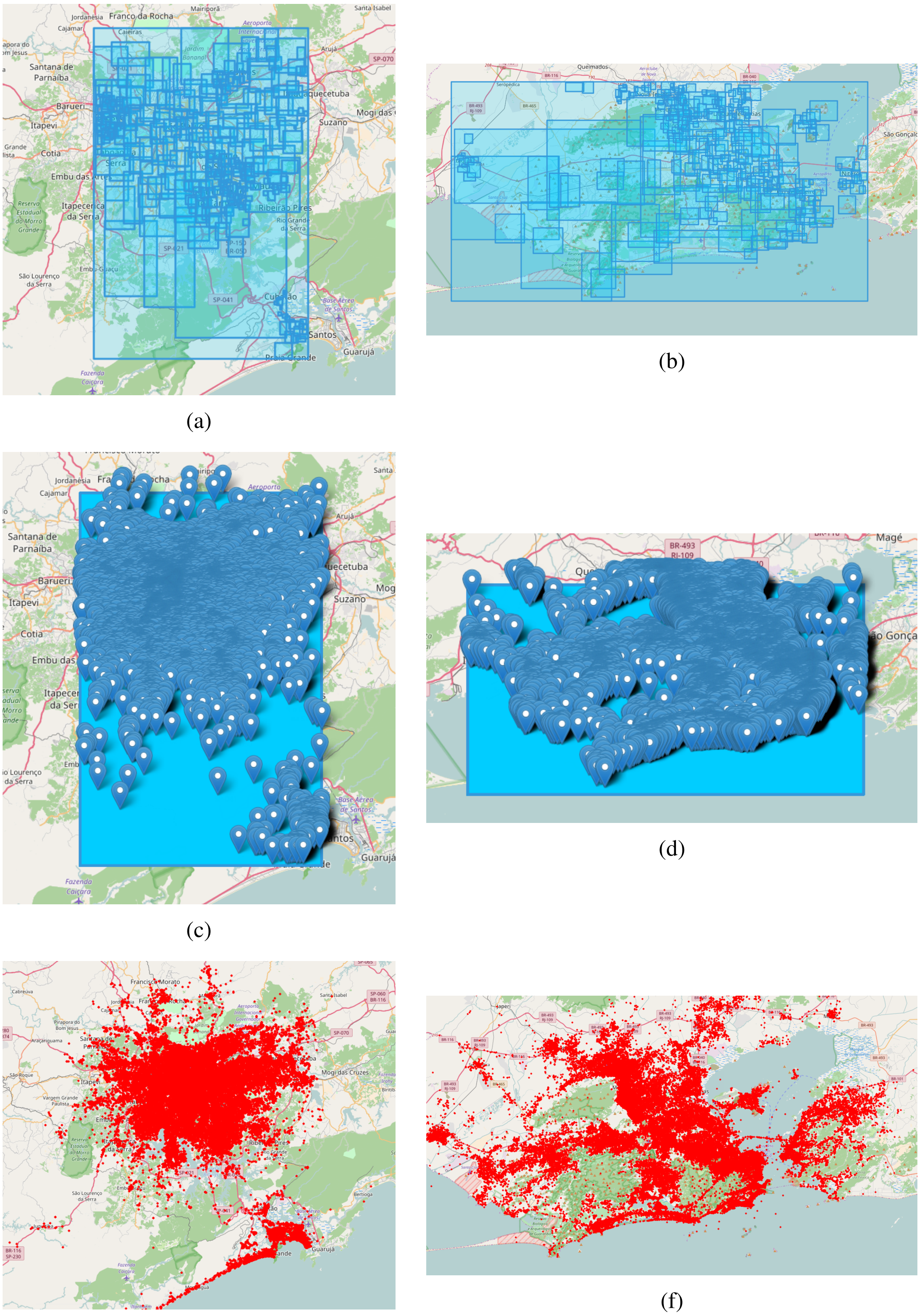
Later, after the filtering process, we tried to understand the volume, as well as the location of each tweet. Through this kind of analysis it was possible to find out that a tweet which *coordinates* field was empty and is, actually, represented with a bounding-box, can also be a specific place, i.e. 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 its location field being represented by a bounding-box format, 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 4.4: 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%

[1] https://cidades.ibge.gov.br/v4/brasil/sp/sao-paulo/panorama (Accessed on 17/06/2017)

[2] https://cidades.ibge.gov.br/v4/brasil/rj/rio-de-janeiro/panorama (Accessed on 17/06/2017)



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

Figure 4.2: 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

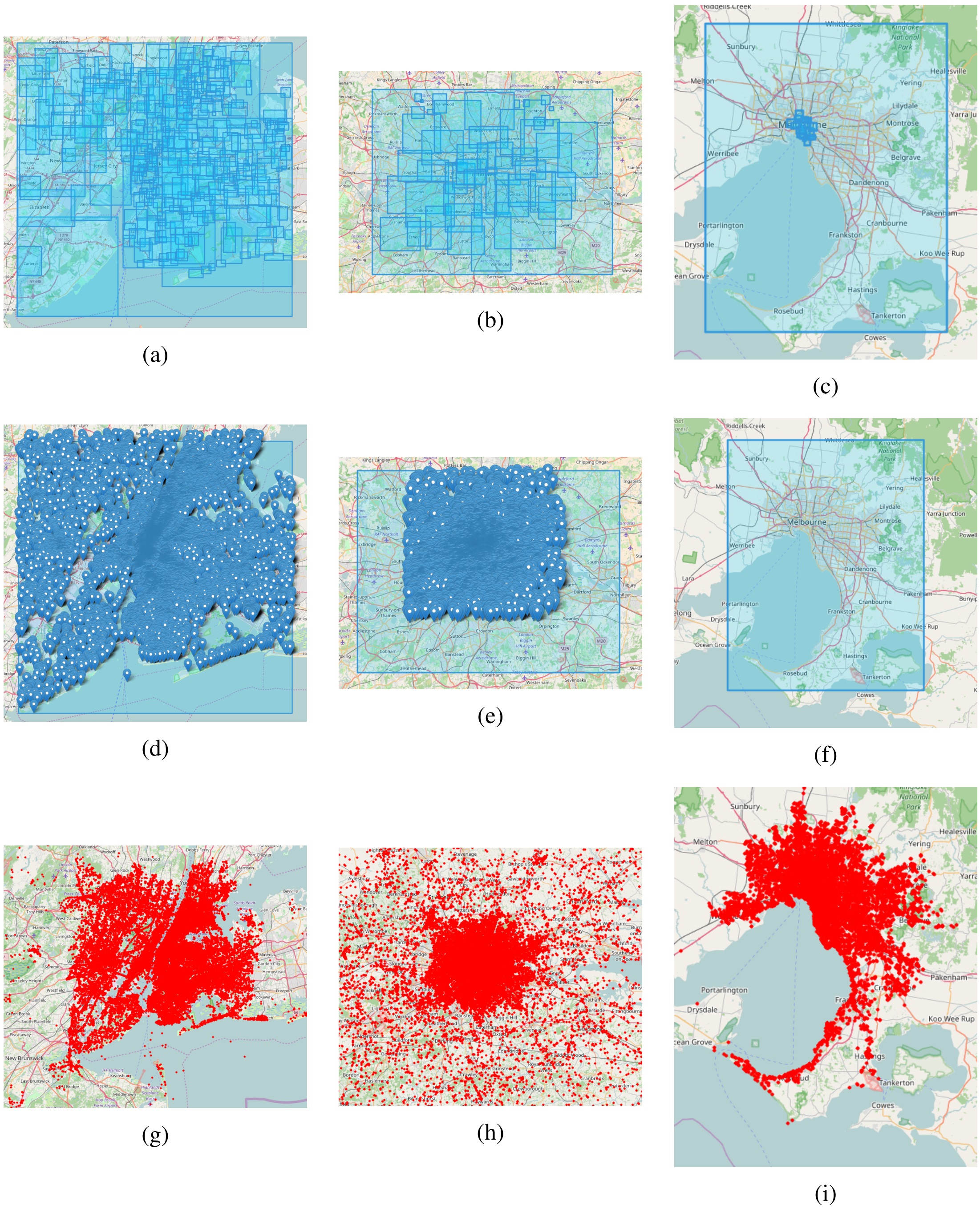


Figure 4.3: New York City (a, d, g), London (b, e, h) Geographical Distributions: (a, b) Bounding-boxes of places (c, d) Specific places (e, f) Geo-tagged tweets

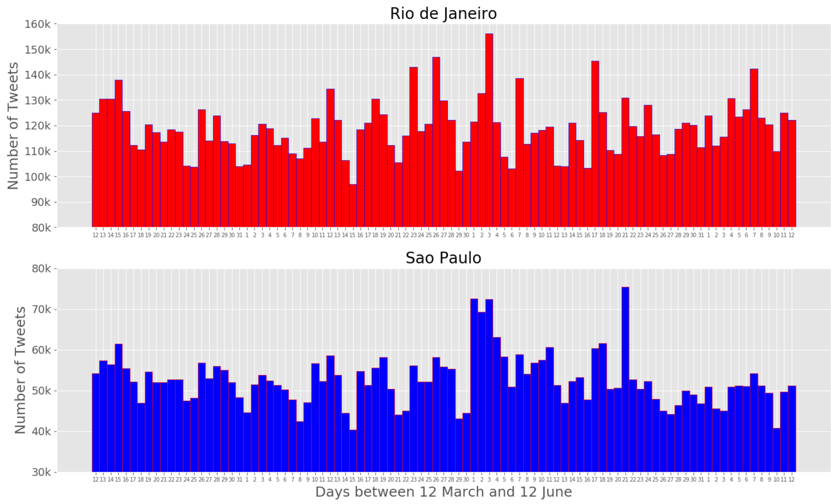
The final counts of the analysis for each identified type of geo-location are presented in Table 4.4. Looking at the numbers it is possible to conclude some facts applicable to all cities. Citizens tend to geo-locate themselves with a location which has variable bounding-box size since more than 70% 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 4.2 and 4.3. The variable bounding-boxes are showed in 4.2a, 4.2b, 4.3a, 4.3b and 4.3c proving that our filter method was able to correctly agglomerate places that were, indeed, inside of the Twitter default bounding-boxes. In 4.2c, 4.2d, 4.3d, 4.3e and 4.3f 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 concerning 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 4.2e, 4.2f, 4.3g, 4.3h and 4.3i 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.

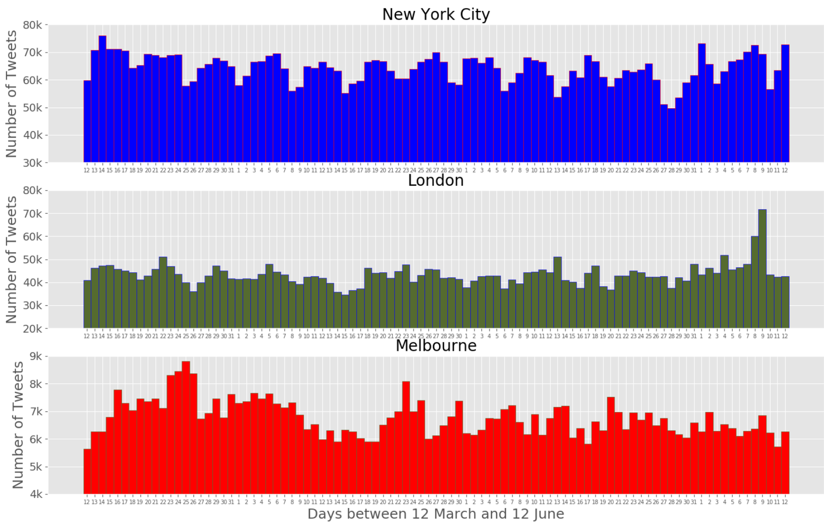
4.2 Temporal Frequencies

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 4.4 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



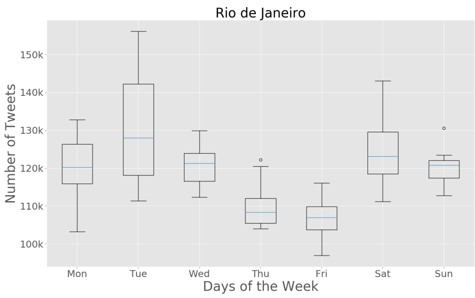
(a)



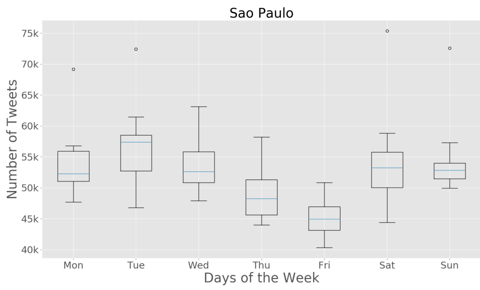
(b)

Figure 4.4: Daily volume of tweets (a) Rio de Janeiro and São Paulo - Portuguese Cities (b) New York City, London and Melbourne - English Cities

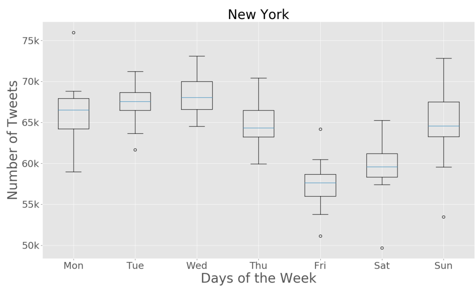
very similar, with exception of Melbourne whose activity is very low comparatively to the other cities (New York City and London). In the particular case of London, we can identify an abrupt increase of volume during days 8 and 9 of June. With the support of external sources such as news websites, we learnt about the United Kingdom General Elections 2017 [3] occurred on that period which suggests that an increase of the Twitter activity might be associated with that event.



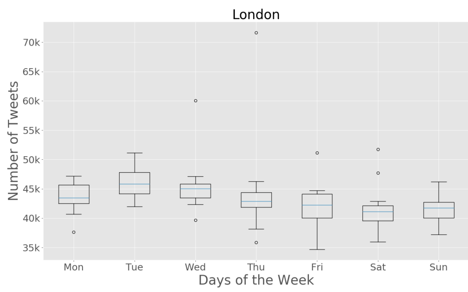
(a)



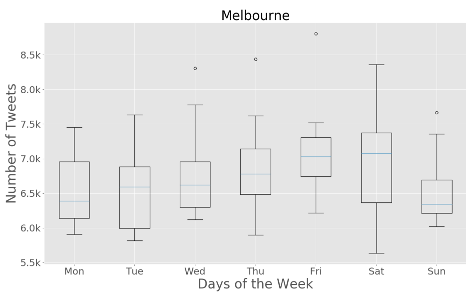
(b)



(c)



(d)

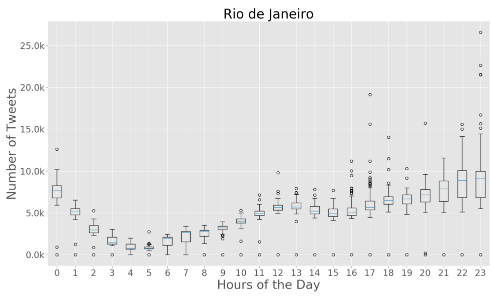


(e)

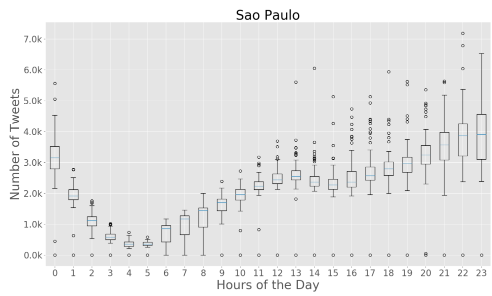
Figure 4.5: Days-of-the-week box-plots for the volume of tweets (a) Rio de Janeiro (b) São Paulo (c) New York City (d) London (e) Melbourne

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)

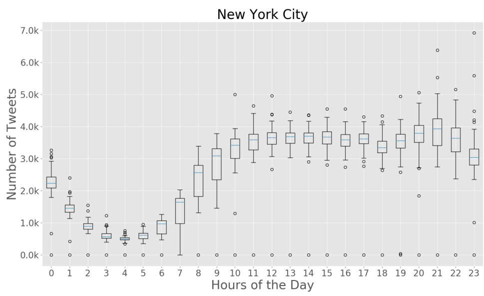
[3] https://www.theguardian.com/politics/general-election-2017 (Accessed on 17/06/2017)



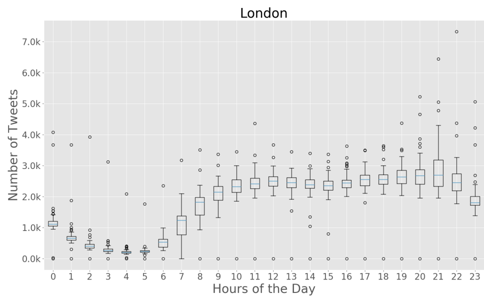
(a)



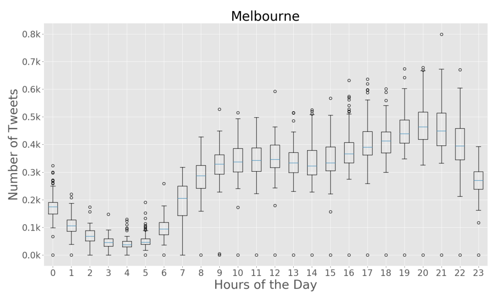
(b)



(c)



(d)



(e)

Figure 4.6: 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

third quartiles as well as (6) the Interquartile-range (IQR). Figures 4.5 and 4.6 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 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.

Looking at the hour-of-the-day box-plot (4.6), 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.

4.3 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 4.7. 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 (4.7a, 4.7b, 4.7c, 4.7d, 4.7e).

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.

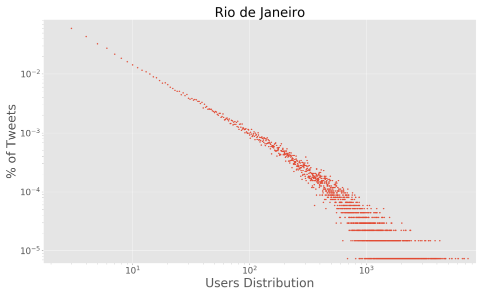
For example, the percentage of users that posted 20 tweets in a period of three months was almost 63% for the city of Rio de Janeiro, São Paulo registered 75%, New York City presented 84%, London showed 87% while Melbourne had 87% of his users with that number of tweets shared. Only taking this example 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.

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.

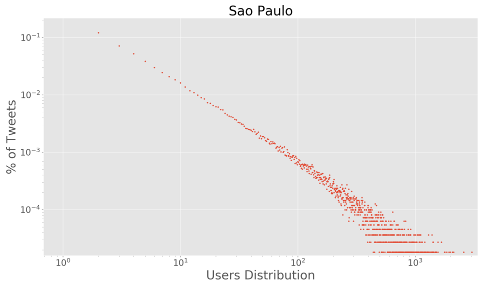
Having this considered, we counted the volume of each element constituting the previously mentioned *metadata* and calculate the percentage of tweets containing it. In Table 4.5 are listed the counts and the corresponding percentage of it relatively to the datasets. The resulting analysis and results 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 4.5: Percentage of Metadata composing the datasets

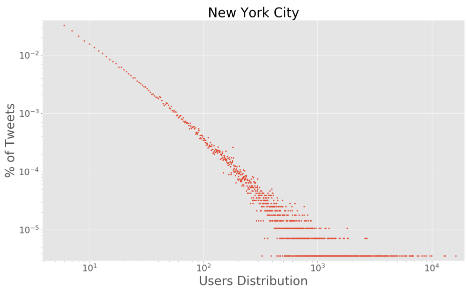
City	Total	Hashtags (#)		User Mentions (@)		URLs		Media	
		Total (tweets)	%	Total (tweets)	%	Total(tweets)	%	Total (tweets)	%
Rio de Janeiro	11,060,136	504,835	4,56%	1,336,329	12,08%	1,783,060	16,12%	409,500	3,70%
São Paulo	4,886,626	593,952	12,15%	1,030,341	21,08%	1,111,749	22,75%	325,385	6,66%
New 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%



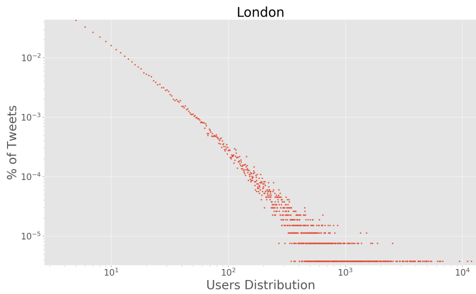
(a)



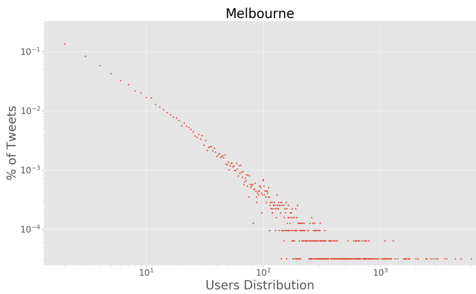
(b)



(c)



(d)



(e)

Figure 4.7: 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

4.4 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 5

Experiments

5.1 Topic Modelling 45
5.2 Travel-related Classification 50
5.3 Summary 61

Providing to the final users of our framework actionable and trustful information about the smart cities and transportation domains is a laborious and time consuming process. First, in order to extract information it is necessary the conduction of some experiments to test the text analytics modules described in Section 3.5. Second, the results obtained from each experiment need to be evaluated in order to prove robustness and efficiency in our implemented modules. Last but not least, it is necessary to enhance the value and importance of such analysis and what indeed social media content can provide to improve services and the whole cities that use our framework.

Hence, in this Chapter, we describe three different experiments that were perform during this dissertation period to implement the previous mentioned text analytics modules

5.1 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 topics 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.

Messages collected to conduct this experiment are 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 3.3, we filtered the data in order to only use the tweets that were actually inside the cities' areas. The final composition of the non-filtered datasets is presented in Table 5.1, and by observing its content it is possible to verify that almost 6M tweets were not located inside the cities' bounding-boxes.

Table 5.1: 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

After the filtering process, we select the subset of data composed by Portuguese tweets and located inside the cities' bounding-boxes, and used it 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.

Usually, to tackle topic modelling tasks in text documents several pre-processing operations are needed in order to make its content as clean and correct as possible since, as previous mentioned in Section 3.5.1, models responsible for this task handle with the most basic element in texts - words. The pre-processing of tweets increase the performance of the LDA model and 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 usually used in this type of analysis. Here, each tweet of both datasets was submitted to a specific and planned group of pre-processing operations, which are described in Section 3.4, and were used to train a LDA model and proceed with the experiment.

It is worth noting that, after the data preparation phase, 772,017 tweets have their message empty which means that its content is irrelevant for the final experiment phase.

5.1.1 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 desire was to reproduce the experiment made by G. Lansley et al. [LL16], which revealed interesting results regarding the characterization of land-uses in 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.

5.1.2 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 5.2 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 5.2: 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.

Table 5.3: 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**

It was found a group of 50 topics which had the largest number of distinct topics between them. We choose to use this model trained to characterize tweets into 50 topics because the final goal of our framework is to be able of compare or monitor cities, regions or even countries having into consideration high levels of diversification and generalization. However, by studying the 50 different topics, we conclude the existence of 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.

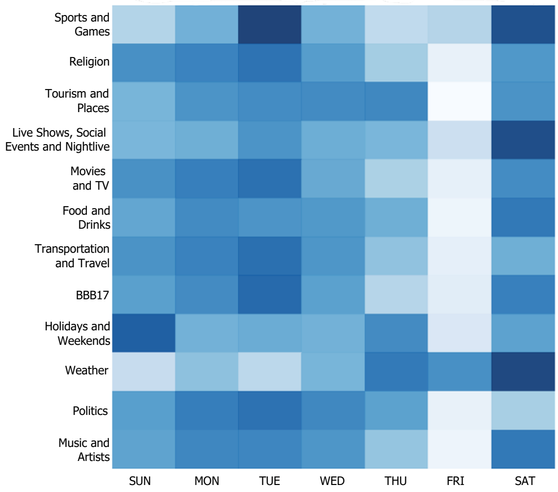
The final set of topics (50 topics) to be considered was selected accordantly to the most recurring subjects. The final classification and details associated with the whole dataset for each city is presented in Table 5.3. Almost every topics demonstrated a balanced distribution, with exception of *Relationships 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 in certain days than others.

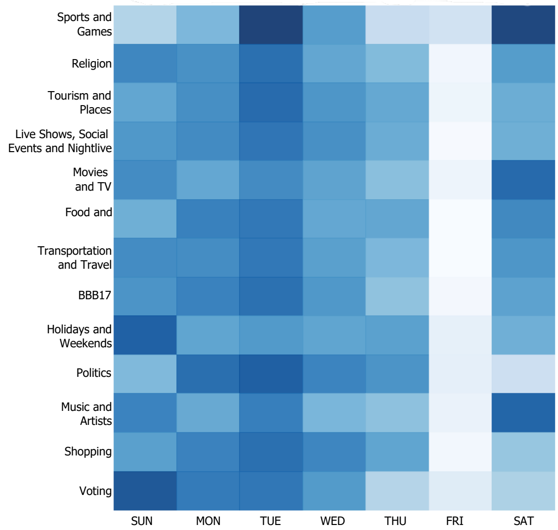
For making such observations some assumptions were made in relation with some *hot* topics. More specifically, we think that is valid to assume that people will talk more about *Religion* in the weekend, since they go to the church in those days. The same result is likely to happen for topics like *Holidays 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 5.1 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.



(a) Rio de Janeiro



(b) São Paulo

Figure 5.1: Day-of-the-week activity per each topic in both cities

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.

5.1.3 Final Remarks

This experiment demonstrates the capability of our framework to handle different topic modelling analysis under unregulated and non-conventional data such as the content found in most social media. The application of topic modeling technique to tweets from two different cities enables interesting comparisons between them since the whole analytics process accounts for what inhabitants talk about in their social networks. Through these analysis, cities' services are capable of monitoring human behaviour, activity patterns as well as of identifying regions where there may be some levels of intolerance on certain topics, making it possible to trigger preventive measures to solve problems in those specific areas.

LDA models usually require documents of large size, or at least ones with higher complexity than a single tweet so as to yield appropriate performance. A traditional approach was followed considering each tweet as a document instead of trying to aggregate tweets in more complex documents taking into consideration some criteria, e.g. grouping messages by date and hour. All topics resulting from our approach are similar in both cities but two, which are unique for each of the selected scenarios. The percentage difference between similar topics was within the interval 0.16-4.43% evidencing the fact that both cities are also similar besides different factors that characterize each other: population, culture, lifestyle and also the region where the city is located in.

In spite of the analysis carried out and reported in this paper, we can not assure that inside a topic other encapsulated topics might exist. The resulting amount of tweets for each topic was extremely high, turning a one-by-one verification into a very laborious and time consuming process. Therefore, our classification approach was limited to the verification of the top-50 words and the manual identification of such words in samples of 200 tweets per topic. Such a limitation, nonetheless, did not prevent us to draw important conclusions relatively to the results we have obtained after the application of our proposed solutions, as discussed in the Section **??**.

Future direction for this research will include application of spatio-temporal aggregation methods over both datasets in order to create more complex documents and verify whether results can be different taking into consideration temporal and spatial factors. To pursue this, it is required that a large dataset for both cities is available, which is expectable only in mid- to long-term. A possible future evaluation approach to be considered is the mapping of topics over specific areas of the cities, such as the identification of topics related to beaches alongside the coastal area in Rio de Janeiro, or the identification of transportation and travel topics over the metropolitan area in São Paulo. Finally, it is also necessary to explore other classification/evaluation approaches to enhance robustness, consistency, and efficiency of the topic modelling routine; one possible solution is the method aforementioned in Section **??**, which considers the addition of an extra layer to endow the framework with supervised labeling capabilities.

5.2 Travel-related Classification

The main goal of this section is to describe the experiments conducted to discriminate travel-related tweets in Rio de Janeiro, São Paulo and New York City. Considering the volume of the collected data for each scenario, it is 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 classification 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"/"ir para a escola"* 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 reminder subsections, we describe two different text classification experiments following distinct approaches across two speaking languages - Portuguese and English.

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

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 3.5.2.1, in the case of the RF classifier.

To evaluate the performance of 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.

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 (horizontally combination of bag-of-words and bag-of-embeddings matrices into a single one).

5.2.1 Rio de Janeiro and São Paulo

Messages were collected for a period of a 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 4.1, we filtered the data in order to use only tweets that were actually inside the cities' areas. The final composition of the datasets is presented in Table 5.4, and the subset of data considered in this experiment sum up a total of 7.7M tweets - 5.3M and 2.4M tweets for Rio de Janeiro and São Paulo, respectively.

Table 5.4: Portuguese datasets composition for the travel-related classification experiment

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	819,000	4,327,000	1,848,000	3,749,000
São Paulo	2,934,000	2,444,000	490,000	2,016,000	918,000	1,672,000

5.2.1.1 Training and Test Datasets

The construction of the training and test sets followed a semi-automatic labeling approach. We tried to built a balanced training set concerning the travel-related class and the non-related. The selection process of tweets have support on a strategy used in the study of Maghrebi et al. [MAW16], which consists in searching tweets from a collection using specific travel terms in conjecture with white-spaces in the start and end of a regular expression (e.g. *" carro "*, for the car mode of transport). Using the terms declared in Table 5.5 combined with the previous mentioned regular expression, 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. Although the randomly selection of tweets to produce such training set, we careful analyse the existence of all modes of transportation present in Table 5.5. After this manual annotation we selected 2,000 tweets and used them as positive samples in the training dataset.

Table 5.5: 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

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.

It is worth mention that we try to hide some terms from the training set in order to verify if the embeddings were able to discriminate tweets about *"Uber"/"Busão"*, which are terms related to the `car/taxi` and `bus` modes of transport, respectively. To assure such test, we incorporate specific-related tweets about these modes of transport in the 71 travel-related tweets of the test set. By obtaining good result regarding the performance of the classification model, we can induce several advantages regarding the use of word-embeddings-based features in social media classification tasks for a large diversity of domains, such as the smart cities and transportation domains.

5.2.1.2 Results and Analysis

Table 5.6 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 5.6, 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.

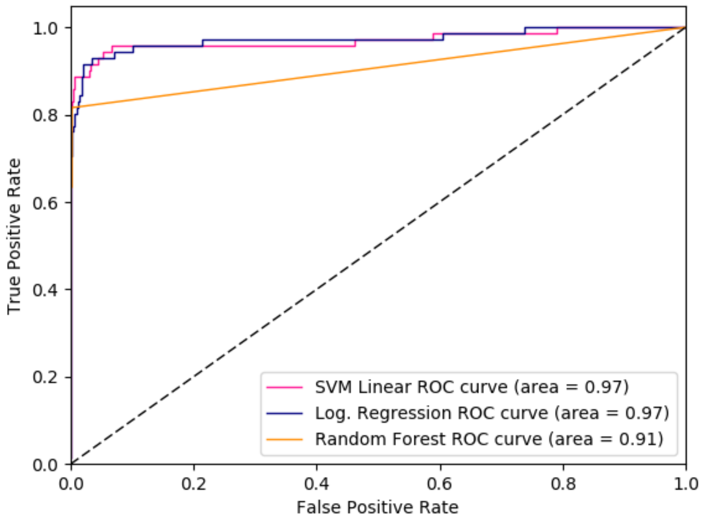


Figure 5.2: ROC Curve of SVM, LR and RF experiences

The performance of all three classifiers is illustrated using the ROC Curve in Figure 5.2. 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 5.6). From a total of 7.8M tweets, our classifier was able identified 37,300 travel-related entries.

Figure 5.3 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.

Table 5.6: Performance results with 100 sized vectors for BoE

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

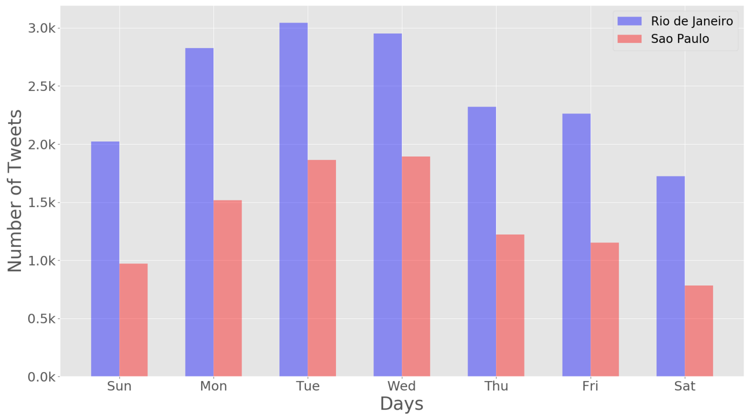


Figure 5.3: Positive Predicted Tweets per Day of Week

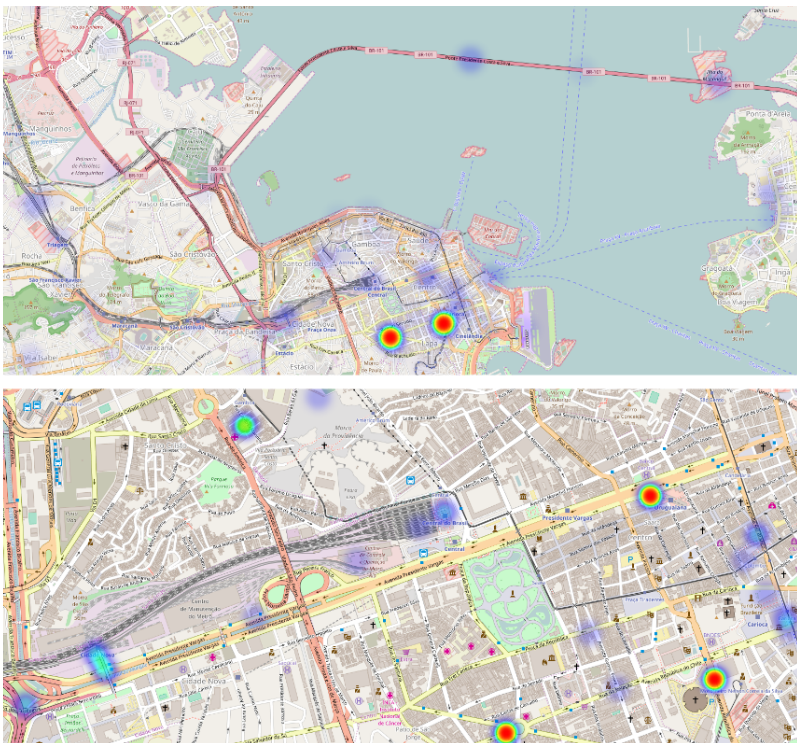


Figure 5.4: Rio de Janeiro heat map to the positive tweets

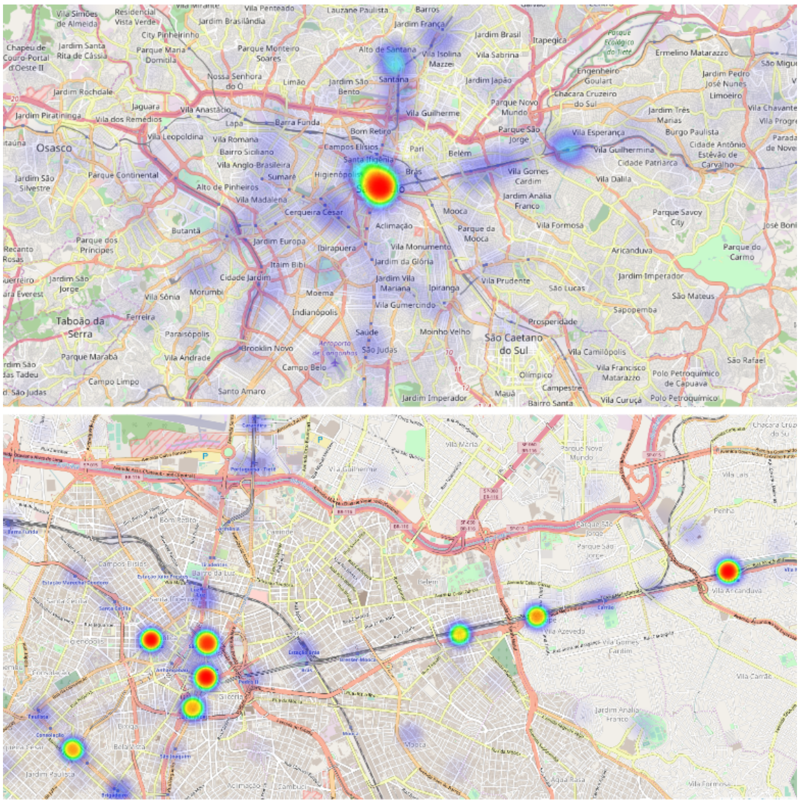


Figure 5.5: São Paulo heat map to the positive tweets

In order to understand the spatial distribution of travel-related tweets we generated a heatmap for both cities. From the heatmap of Rio de Janeiro, illustrated in Figure 5.4, it is possible to identify that 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 Figure 5.5, 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.

5.2.1.3 Final Remarks

The previous described experiment 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, 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. Modes of transport are always evolving and new services emerges making the identification of tweets related to it difficult. Overall, our experiment proved that word-embeddings features are actually an advantage regarding its applicability into instable real-world scenarios such as the transportation domain.

5.2.2 New York City

Similar to the experiment of Portuguese-speaking travel-related classification of tweets, we built a model to discriminate English-speaking travel-related tweets. However, the construction of the training and test sets in this experiment follows a different approach. While in the Rio de Janeiro and São Paulo experiment we explore an semi-automatic approach and tweets were almost instantaneous formed as a group, here we were obligate to follow a two-phase approach due to the polysemy level of English travel terms.

Differently from the Brazilian cities 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, as well as tweets located outside the bounding-box of New York City, the resulting dataset comprehends 4M tweets.

Regarding the preparation of data, we used the same preprocessing operations in both experiments, Brazilian and North-American. The operations were lowercasing, transformation of repeated characters and cleaning of *entities* (user mentions and URLs) from the message content.

5.2.2.1 Training and Test Datasets

In the Portuguese dictionary, travel-related terms do not have more than one meaning. For instance *"caminhar"* or even *"comboio"* possesses only one meaning. Regarding the English dictionary, travel-related tweets may have more than one meaning since some of them present high level of polysemy. Terms such as *"walk"* may be used to describe the action of walk or, for example, the action of *walk into*. On the other hand, the term *"train"* can be used to describe the mode of transport train or a type of behaviour through practice and instruction.

The polysemy level of such terms was took into consideration while the construction process of our training set of tweets for the English-language travel-related classification model. In the first stage of the construction process, we used the same strategy of the Portuguese training set. By take support on a semi-automatic labeling technique using a regular expression, we find out almost 16,000 tweets. The next step in the construction process was a manually verification followed by a manually annotation. Overall, 1,686 tweets were selected for each of both binary classes, 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.

Table 5.7: Composition of the training and test datasets for the English travel-related tweets classification

	Training Set	
Mode of Tranport	**Travel-related**	**Non-related**
Bike	300	1686
Bus	311	
Car	317	
Taxi	314	
Train	317	
Walk	217	
Total	3372	

Nonetheless, we include into the training set tweets which polysemy level may induce doubts regarding the context of the message in order to make possible higher levels of discrimination in our model. This inclusion may help the learning process of our model making it capable of correctly identify which are the tweets that are actually related to the travel and transportation domain. The final composition of the training datasets is presented in Table 5.7.

5.2.2.2 Preliminary Results

Due to the laborious and time-consuming effort made in the construction of the training set, we opt to apply a different approach in the training phase of our model classification model. In order to enhance the differences between tweets whose terms present high levels of polysemy, the model was trained using a k-fold cross-validation technique with 10 iterations for all groups of features: bag-of-words and bag-of-embeddings and both combined. Results showed good performance for all models regarding the selected evaluation metrics. The best model in this experiment was the Logistic Regression classifier trained with bag-of-words and bag-of-embeddings features, presenting a F1-score of 0,98324.

The fact that all models performed incredibly well, in particular models using the features group of BoW and BoW+BoE raise to us some questions and doubts about the robustness of the

Table 5.8: Preliminary results (it is only demonstrated the best result for the bag-of-embeddings group)

Classifier	Features	Precision	Recall	F1-score
Linear SVM	BoE(200)	0,90883	0,83634	0,87089
	BoW	0,96298	0,97652	0,96962
	BoE(200) + BoW	**0,97251**	**0,99114**	**0,98170**
Logistic Regression	BoE(100)	0,90172	0,84948	0,87447
	BoW	0,96431	0,98042	0,97222
	BoE(200) + BoW	**0,97391**	**0,99285**	**0,98324**
Random Forests	BoE(100)	0,81283	0,83600	0,82394
	BoW	0,96569	0,98997	0,97764
	BoE(50) + BoW	**0,93688**	**0,99939**	**0,96701**

features used in the training process. First, in the Brazilian cities experiment, by following the same approach over the training set construction process we did not obtain results of this kind. Second, the selected tweets are very specific and our model may be overfitted due to training data. In order to pursue and have answers to our questions, we designed another experiment using the same dataset.

5.2.2.3 *Leave-one-group-out*

It is worth noting that in our first experiment all travel-mode classes were known by the model before the classification of the test set (the remaining sub-dataset in the 10-fold cross-validation). Comparing with real-world scenarios, this may not be true since new modes of transport and companies, such as Uber, Lyft and Cabify, arise from unpredictable moments. This 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. Hence, the behaviour of the learned model when facing a completely unknown travel-mode class can be evaluated. A model for each hidden transport-mode class was built and evaluated using the same training conditions and metrics. The datasets composition of each experiment led in this strategy can be observed in Table 5.9.

Table 5.9: Datasets composition used in the *leave-one-group-out* strategy

<table>
<tr><th rowspan="2">Travel-Mode Class</th><th colspan="2">Training Set</th><th colspan="2">Test Set</th></tr>
<tr><th>Pos.</th><th>Neg.</th><th>Pos.</th><th>Neg.</th></tr>
<tr><td>Taxi</td><td>1,372</td><td rowspan="6">1,686</td><td>314</td><td rowspan="6">300</td></tr>
<tr><td>Train</td><td>1,369</td><td>317</td></tr>
<tr><td>Car</td><td>1,369</td><td>317</td></tr>
<tr><td>Bike</td><td>1,386</td><td>300</td></tr>
<tr><td>Walk</td><td>1,469</td><td>217</td></tr>
<tr><td>Bus</td><td>1,375</td><td>311</td></tr>
</table>

For each experiment of the learning models, we maintain a 10-fold cross-validation approach, however it was built a test set with a hidden travel-mode class and 300 non-related tweets (negative class). Here, only bag-of-words and bag-of-embeddings features were fed into our models classification routine since the main goal of this experiment is to check the features robustness. Table 5.10 presents the best results for each model, as so the group of features feeding it. To achieve the final results of this experiment, we calculated the mean between all models' results to each of the hidden transport-mode classes.

Table 5.10: *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**

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 version. The generalization power is an important and crucial characteristic for our desired solution since 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. Having this considered, the bag-of-words features group presents lack of robustness as we doubt in our first experiment (Section 5.2.2.2).

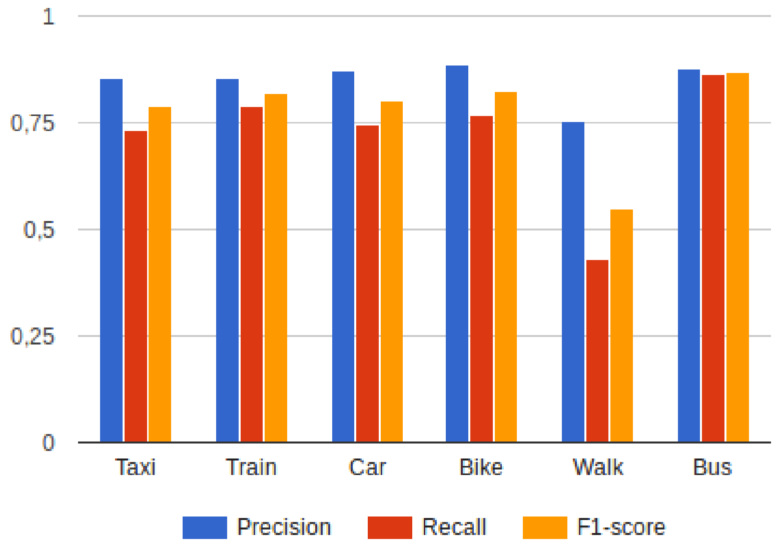


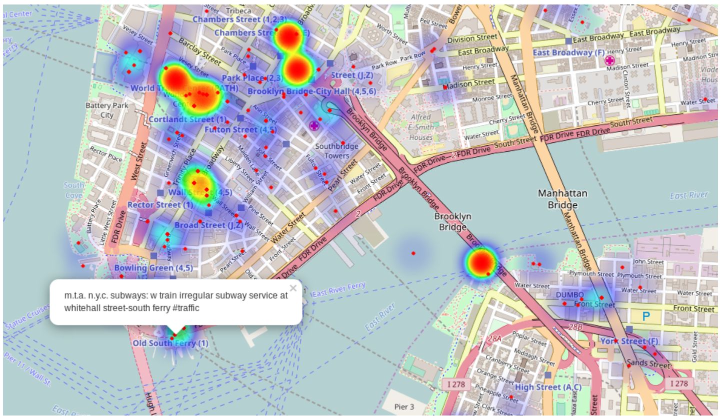
Figure 5.6: SVM model with BoE(200) for each travel mode

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 5.6 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.

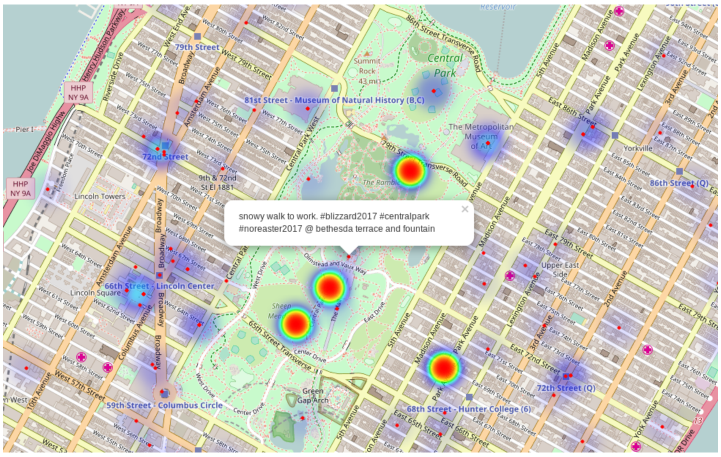
Having the experiments concluded, we used the best model, in this case, Linear SVM for the dimensionality of 200, to predict the 4M tweets that composed the NYC dataset. Almost 300,000 tweets were classified as travel-related. After the classification step, a sample of 10,000 tweets was taken from all 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 5.7a, 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.

Table 5.11: 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 5.7b**
m.t.a. n.y.c subways: w train irregular subway service at whitehall street-south ferry #traffic - **Figure 5.7a**



(a)



(b)

Figure 5.7: Spatial density of the travel-related predicted tweets in New York City: (a) South of Manhattan and over the Brooklyn Bridge, (b) Central Park

5.2.3 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 which is not desired when facing real-world scenarios and lots of changes happen in short periods of time. On the other hand, by looking at the results, the almost perfect performance lead us to doubt about the existence of overfitting, and so, a *leave-one-group-out* strategy was applied to validate the robustness of features. There, 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 and robustness 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 travel-related 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.

5.3 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, 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.

Moreover, 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. Although huge restrictions regardind the labelling of each topic, results show promising contributions and informations to the *smart cities* entities, allowing until this
point possible identifications of what are the most *hot* topics in each region. 2

Chapter 6

Conclusions and Future Work

6.1 Final Remarks . . . 63
6.2 Contributions . . . 64
6.3 Future Work . . . 65
6.4 Publications . . . 66

6.1 Final Remarks

The literature review, studied in Chapter 2, shows the main challenges across the evolution process of a city in order to be titled as *smart*. Moreover, the biggest restrictions in the development of intelligent systems using social media data are enunciated as well as possible methodologies and techniques to solve it. By combining the challenges of a *smart city* and the restrictions present in the analysis of text messages, in particular, social media messages, the problem around this dissertation was divided into five distinct ones. The solutions presented to each of the sub-problems took into consideration lacks observed in the literature review. Domain generalization in the conception of an automatic system capable of collect, filter, processing, aggregate and demonstrate, through graphical representations, valuable information to final entities/users is one of the identified lacks, in terms of being constructed with the support of supervised methods. The transportation domain also present lacks regarding discrimination of travel-related tweets using methods that take advantages from semantic and syntactic similarities in texts. The majority of works present conventional techniques such as bag-of-words, which although good performances represents high risks when implementing supervised learning models due to the possibility of overfit the model in the training routine.

Assuming the challenges identified for this dissertation as well as the previous mentioned lacks in the literature review, we propose and develop a domain-agnostic framework and test it using five different cities over the world as use cases for three distinct analysis: simple statistics, travel-related classification for English and Portuguese languages and Twitter topics identification over two Brazilian megacities, Rio de Janeiro and São Paulo.

Travel-related classification of tweets using a combined approach of bag-of-words and bag-of-embeddings proved, as P. Saleiro [SMRSO15] also reported for sentiment polarity of financial tweets, that each representation completes the other since results showed consistency and robustness over the classification performed to Portuguese texts. On the contrary, English speaking tweets do not present similar results, however, as it was previous mentioned, there was a suspicion of overffiting in the model training process. Further experiments, proved such theory since the model was able to maintain its performance only using bag-of-embeddings as training features.

Characterization of topics is a very common type of analysis in terms of information extraction from tweets. In this dissertation, we explore this analysis and implement a specific module for this task, integrating it in the final framework. Our experiment reveal promising results, however auxiliary methods would probably improved the information obtained turning it more concise and accurate. Literature review shows approaches potential which can help in this future improvements of the model responsible for identification of topics in tweets.

It is worth nothing that every experiment was performed having only into consideration geo-located tweets. This choice, as previously mentioned, was due to the additional information contained in this type of tweets. By analyse the location and combine it with results of the classification tasks, characterizations and studies over specific areas in cities are possible to be reported as weel as identifying existent and notable patterns regarding travel-related problems and urban dynamics.

Having all considered, it is necessary to divulge that the final framework is still far of its full potential, and supported on this, we consider it as a scalable, flexible and adaptive prototype that must improved over time since implementing supervised learning systems is a laborious and time-consuming process

6.2 Contributions

At the end of this dissertation, efforts applied are summarized in three different types of contributions.

- **Scientific Contributions**

  In order to test each module composing the framework, several experiments using conventional and recently text mining methods were followed. Our desire to share the advantages obtained on such experiments take us to perform three attempts of scientific contributions. The first one is about automatic classification of travel-related Portuguese speaking tweets, for the cities of Rio de Janeiro and São Paulo, and is currently under the press phase in the EPIA 2017. Attempts are performed taking into consideration different types of features in the training phase of the model, being the most accurate the one combining bag-of-words and bag-of-embedding.

Further experiment reports the previous mentioned method over English speaking tweets from New York City. The final results reveal differences comparing to the Portuguese experiment. Having this considered, another approach was chosen to prove signs of overfitting in the training process, *leave-one-group-out strategy*. Final remarks demonstrate the consistency of word embeddings model for hidden modes of transport classes, while bag-of-words model prove to be dependent of the examples used in the training phase. The overall experiment was submitted to the CIKM-2017 and is currently in review.

Finally, the experiment regarding topic modelling is reported to the IEEE S3C 2017. There is described the use of LDA model to characterize the topic present in a tweet. Promising results were obtained after a difficult topic classification phase. The final model was then used to implement the topic modelling sub-module of the developed framework in this dissertation. It is worth noting that this contribution, similar to the previous one, is under review phase.

- **Technical Contributions**

  At the end of this work, we report that every implementation performed during the dissertation period will be open-sourced to help future candidates in the integration of new functionalities to the framework. Besides that, the implementation of the travel-related classification models require the conception of labeled datasets regarding the transportation domain. These datasets, containing Portuguese and English speaking tweets will be uploaded in order to fulfill the absence of public datasets, with hope of being considered a gold standard in future developments of this kind.

- **Applicational Contributions**

  The most important contributions of this dissertation are the analysis provided by the developed automatic analysis-based system. The information provide by such system can serve to support monitoring tasks in cities as well as help in future decision-making policies by the responsible entities' services. Although the final framework being presented as a prototype, with integration of new features to the system, there are infinite possibilities for its use as well as its potential for the smart cities domain.

6.3 Future Work

The dissertation purpose had as it main focus the conception of an automatic system capable of analyse real-time data streams from social media platforms in order to produce valuable information for users of services or even its responsible entities. For achieve the proposed goals, we tried to explore already consistent state-of-the-art methodologies as well as unexplored ones regarding specific domains. Since this framework can be seen as a prototype of a future complex system, several improvements can be invested here. Although already existent modules and text analysis devised, it worth noting the conjecture of a additional sentiment analysis module in order to infer the sentiment polarity value regarding specific zones where the travel-related tweets were located in, as so the overall sentiment in an identified topic.

Another important work to pursue in the future is to correlate the results of this study with official sources of transportation agencies relatively to traffic congestions and other events on the transportation network, including all modes of transports and their integration interfaces and modules. This kind of association will be useful both to validate the proposed approach as well as to improve the inference process and knowledge extraction. The automatic classifier herein presented will then be integrated into data fusion routines to enhance transportation supply and demand prediction processes alongside other sensors and sources of information.

A possible future direction to improve the topic modelling approach is the application of spatio-temporal aggregation methods under a sample of data to create more complex documents, retrain the model and verify if the results can be different taking into consideration some of the factors that distinguish both cities: demographics, culture and location. An attempt to pursue good performances using supervised LDA models also needs to be enhanced here.

Lastly, there is a need of creation of other specific models to other fields of a *smart city* in order to assure equally performances for any of its fields.

6.4 Publications

During the period of this dissertation, we published three different scientific papers in order to share our experiments' methodologies and results.

- 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, Rosaldo J. F. Rossetti and Javier Sanchez-Medina. Classifying Travel-related Tweets Using Word Embeddings. In *IEEE 20th International Conference on Intelligent Transportation Systems* (IEEE ITSC), 2017. Under review.

- João Pereira, Arian Pasquali, Pedro Saleiro, Rosaldo J. F. Rossetti and Nélio Cacho. Characterizing Geo-located Tweets in Brazilian Megacities. In *The Third International Smart Cities Conference* (ISC2), 2017. Under review.

Acronyms

API Application Programming Interface. 21, 23, 24, 29, 67

AUC Area Under the Curve. 15, 67

BoE Bag-of-embeddings. 28, 29, 57, 67

BoW Bag-of-words. 28, 52, 57, 67

DT J48 Decision Trees J48. 13, 14, 67

FPR False Positive Rate. 15, 67

GPS Global Positioning System. 24, 67

HTML HyperText Markup Language. 30, 67

ICT Information and Communications Technology. 6, 7, 67

IQR Interquartile-range. 40, 67

ITS Intelligent Transportation Systems. 7, 67

JSON JavaScript Object Notation. 29, 67

LDA Latent Dirichlet Allocation. xi, 11–13, 27, 67

MLP Multilayer Perceptron. 13, 67

NB Naïve Bayes. 13, 14, 67

NLP Natural Language Processing. 2, 19, 20, 67

NLTK Natural Language Toolkit. 25, 26, 28, 67

OLS Ordinary Least Squares. 13, 14, 67

REST Representational State Transfer. 21, 67

RF Random Forests. 13, 67

ROC Receiver Operating Characteristic. 15, 67

SM Smart Mobility. 7, 67

SMC Social Media Content. 1, 3, 14, 67

SVM Suport Vector Machines. 13, 14, 67

TPR True Positive Rate. 15, 67

UGC User Generated Content. 1, 67

UI User Interface. 3, 67

UTC Coordinated Universal Timezone. 26, 67

Glossary

Crowdsensing or mobile crowdsensing Technique used to collectively share and extract information from large groups of infividuals in order to analyse, infer or even measure processes of common interest.. 7, 67

Influenza A Influenza A is a type of virus capable of infecting animals, although it is more common for people to suffer the ailments associated with this type of flu.. 13, 67

k-fold cross-validation It is a technique where the original dataset is randomly partitioned into k equal sized sub-datasets. Of the k sub-datasets, only one is retained as the validation data for testing the model, and the remaining $k\ \ 1$ sub-datasets are used as training data.. 57, 67

MicroBlog It is a tool that allows quick abd short status updates, and if possible, through multiple different platforms.. 1, 2, 67

Twitter Firehose It is a paid Twitter service that guarantees the delivery of 100% of the tweets matched with certain criteria.. 21, 67

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.

[ACK$^+$05] Sophia Ananiadou, Julia Chruszcz, John Keane, John McNaught, and Paul Watry. The national centre for text mining: Aims and objectives, January 2005. [Accessed on 25/06/2017].

[AHH$^+$12] Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. Twitcident. *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, page 305, 2012.

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

[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.

[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.

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

[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.

[CDBN11] Andrea Caragliu, Chiara Del Bo, and Peter Nijkamp. Smart cities in europe. *Journal of urban technology*, 18(2):65–82, 2011.

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

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

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

[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.

[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.

[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.

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

[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.

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

[Hol08] Robert G Hollands. Will the real smart city please stand up? intelligent, progressive or entrepreneurial? *City*, 12(3):303–320, 2008.

[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.

[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.

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[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.

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

[KOM16] Abdullah Kurkcu, Kaan Ozbay, and Ender Faruk Morgul. Evaluating the usability of geo-located twitter as a tool for human activity and mobility patterns: A case study for new york city. In *Transportation Research Board 95th Annual Meeting*, number 16-3901, 2016.

[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.

[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.

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

[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.

[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.

[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.

[MB08] Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.

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

[MH13] Eric Mai and Rob Hranac. Twitter interactions as a data source for transportation incidents. In *Proc. Transportation Research Board 92nd Ann. Meeting*, number 13-1636, 2013.

[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.

[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.

[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.

[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.

[MSLdG15] 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.

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

[NSO+15] Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681, 2015.

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

[OPST16] João Oliveira, Mike Pinto, Pedro Saleiro, and Jorge Teixeira. Sentibubbles: Topic modeling and sentiment visualization of entity-centric tweets. In *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering*, pages 123–124. ACM, 2016.

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

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

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

[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.

[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.

[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.

[RSR15] Francisco Rebelo, Carlos Soares, and Rosaldo JF Rossetti. Twitterjam: Identification of mobility patterns in urban centers based on tweets. In *Smart Cities Conference (ISC2), 2015 IEEE First International*, pages 1–6. IEEE, 2015.

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

[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.

[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.

[SGS16] Pedro Saleiro, Luís Gomes, and Carlos Soares. Sentiment aggregate functions for political opinion polling using microblog streams. In *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering*, pages 44–50. ACM, 2016.

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

[SMRSO15] Pedro Saleiro, Eduarda Mendes Rodrigues, Carlos Soares, and Eugenio Oliveira. Texrep: A text mining framework for online reputation monitoring. *New Generation Computing*, 2015.

[SRSO17] Pedro Saleiro, Eduarda Mendes Rodrigues, Carlos Soares, and Eugénio Oliveira. Feup at semeval-2017 task 5: Predicting sentiment polarity and intensity with financial word embeddings. *arXiv preprint arXiv:1704.05091*, 2017.

[SSP11] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467, 2011.

[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.

[TT15] Suppawong Tuarob and Conrad S Tucker. Quantifying product favorability and extracting notable product features using large scale social media data. *Journal of Computing and Information Science in Engineering*, 15(3):031003, 2015.

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

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

[ZCLL10] Daniel Zeng, Hsinchun Chen, Robert Lusch, and Shu-Hsing Li. Social media analytics and intelligence. *IEEE Intelligent Systems*, 25(6):13–16, 2010.

[ZNHG16] Zhenhua Zhang, Ming Ni, Qing He, and Jing Gao. Mining transportation information from social media for planned and unplanned events. 2016.