FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Social Media Text Processing and Semantic Analysis for Smart Cities

João Filipe Figueiredo Pereira

DISSERTATION PLANNING

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Rosaldo José Fernandes Rossetti

Supervisor: Pedro dos Santos Saleiro da Cruz

February 20, 2017

Social Media Text Processing and Semantic Analysis for Smart Cities

João Filipe Figueiredo Pereira

Mestrado Integrado em Engenharia Informática e Computação

February 20, 2017

Abstract

The *Smart Cities* are a future that most cities want to achieve by appealing to the citizens' participation to improve its services. The participation process is comprised of simple contributions, opinions or even validations of technological projects by the citizens. In this way, the social networks are a rich source of data where ordinary citizens publicly share their comments, and for this reason, it's resorted by many text mining projects since this information provides interesting analytics based on *human-generated* data.

The extraction of information from social media texts, in this case the *Twitter*, regarding the services of a *smart city*, is the major topic that will be discussed here. The texts that are obtained from a social network can sometimes be difficult to analyze, since the writing of a person is not always correct and/or coherent, such as, wrongly typing words and/or incorrect phrasal structuring, respectively. With all of this points, it's possible to slice the main problem into four sub-problems. The first one is to collect data from the social network; the second one is focused on the semantic analysis and elimination of ambiguous lexicons of the text; the third one will be the extraction of aspects from the messages and the classification of the sentiment polarity; finally, it's necessary to aggregate the results obtained into a set of indicators to categorize the founded problems.

We propose to tackle the aforementioned problems through the creation of a framework capable of collecting and extracting value from social media streams. The framework will be composed by a total of four modules, in order to solve each sub-problem mentioned above. The first one should handle social-data-collection tasks according to some user-defined heuristics. The second module should semantically analyze the text and remove ambiguities from the lexicon, using named entity linking techniques in order to filter the unrelated messages from the dataset. After this filtering, the texts will be subjected to the module of sentiment analysis to find out if its content has positive or negative value regarding some specific topics. Finally, the analytics module will aim the aggregation of the results obtained from the previously module in a set of statistical visual indicators.

The expected results in this dissertation rely heavily in the UI presented to the end-user. The knowledge derived from the text processing tasks may help ordinary users of cities services or even the responsible entities in the identification and interpretation of some problems, and, thus, the decision-making process can be easier.

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Abbreviations

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

Chapter 1

Introduction

1.1 Context and Motivation

The rise of social media services, in the last few years, has led to an excessive amount of information being placed online by the population. The need to explore this type of information has a steadily grown in order to realize what kind of value could bring to the areas of marketing, business or even politician [?]. Micro-blogging platforms, such as Twitter, have a huge affluence in a daily-basis, where people publicly share around 500 million messages about a diverse set of current themes, expressing their opinions and feelings [?]. For this reason, technological projects developed in some cities has seen social media streams as a potential resource to extract knowledge, i.e. the satisfaction of people regarding some services in the cities, such as the urban transportation service [?]. The collection of this participative activity of people through online platforms, named Crowd Sensing, emerged to replace the traditional way of sensing data capture and tracking in physical infrastructures, allowing a considerable reduction of economic costs [?].

The information extraction from social media streams is a hard task. For example, tweets besides being text messages and, consequently unstructured data. There's also some extra particularities, as, for example, the limited length (140 characters per message) which restricts the amount of information in its content, the informal language used and there are also many spell-mistakes, abbreviations, ambiguity and special mentions (e.g. URL references, hashtags, users) and the presence of a high variety of emoticons [?].

Many research projects have been conducted in order to extract the sentiment present in opinions through text mining techniques. Sentiment analysis is the field that focuses on this task. The detection of the sentiment polarity in messages generated by citizens, whether at a specific level (relative to certain aspects) or at a general level (general sentiment of the message) can allow companies or even ordinary citizens the possibility of identification city's services problems and may be assimilated to a kind of sensor for the issue of quality awareness, providing, at least, some help in decision-making processes.

1.2 Problem Statement and Goals

The problem around this dissertation establishes in the analysis of a continuous flow of social media streams, in particular from Twitter, about a target scenario, as, for example, the quality of the urban transportation services in Porto. Hence, it will be necessary to filter the relevant messages, extract sentiment aspects/topics with polarity and create useful aggregated data visualizations. The problem presented can be divided in five distinct points:

1. **Data collection for our target scenario**

   At this point, a real scenario must be chosen so that a case study can be produced.

2. **Named Entity Disambiguation and content filtering**

   The identification of the entities present in the messages is a important point in order to disambiguate some mentions that could not be related with target scenario, making the filtering task easier, since only messages that are related must appear in the final dataset.

3. **Identification of aspects/topics in Twitter messages**

   Since each opinion usually has a target aspect/topic about an entity, or even a service of a city, the recognition of it is relevant so that the sentiment present in the message has an orientation.

4. **Sentiment polarity classification**

   The polarity of a sentiment may have three different types: positive, negative or neutral. At this point, it will be necessary to estimate the level of polarity expressed in the message.

5. **Data aggregation and visualization**

   The aggregation of the results provide by all other tasks is needed. Some messages could refer to the same aspect through different ways, so it will be necessary aggregate messages that present this characteristic. Another important task of the aggregation is the continuously calculation of the results So that when an user access the analytics UI, the results are presented immediately, without waiting time. At the visualization of the results, some qualitative and quantitative indicators may be presented to the end-user to make the analysis easier.

Taking into account all the aforementioned points, the final goal of this dissertation is to create a framework, based on the concept of analysis. The framework should be capable of automatically processing social media texts, regarding semantic processing, topic detection and sentiment analysis. An user interface will be provide to the end-user, illustrating a set of qualitative and quantitative indicators to analysis. This knowledge can be relevant both to users of a particular service or to the responsible entities in order to improve the decision-making process.

1.3 Structure of this Dissertation

This report covers a great diversity of points and because of that, its structure is divided in three different sections.

The Section 2 starts with a brief contextualization in the Smart Cities and Intelligent Transportation System fields. After that, it was made a review on Social Media Analytics, specially about Twitter, and what benefits its exploration can provide to the civilization. To explore the information from Twitter messages, an intensive study regarding the Text Mining area was made, in particular Information Extraction, Topic Modeling and the various Sentiment Analysis fields, as well as related works that already explored similar problems with social media data.

The proposed solution for the aforementioned problem and its methodology are referenced in Section 3. The final section of this planning report is composed by some conclusions relatively to the studied works and what benefits and risks our solution may have - Section 4.

Chapter 2

Background and Literature Review

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

2.1 Smart Cities and Intelligent Transportation Systems

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

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

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

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

2.2 Social Media Analytics

In the last few years, the Social Networks have generate an incredible impact in the business communication because of the rise of Web 2.0 where the users assumed the role of customers of content through high levels of interactivity [?]. 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 [?]. Facebook, Twitter and other social networking sites are nowadays used as business tools by companies aiming the efficient use of digital marketing techniques to publicize their products [?]. Besides the business field, the population turn into this new communication technologies in a intensely way, where they publicy share real-life events, their opinions about certain topic, their on-time feelings in the network through a simple message [?]. Social Media Analytics can be describe as a type of digital analytics to study the people interaction with others, or their opinion about companies, its products and services through the social media data. This study provides important information to "analysts, brands, agencies or vendors", and its analysis could facilitate the generation of economic value to many organizations [?]. To achieve the main goals of the SMA, the companies focus their effort in the development and evaluation of frameworks, to make possible an easy collection, analyse, summarization and visualization of processed social media data. Hence, the companies can establish specific points about what to improve in their products [?]. To create a significantly value regarding the SMA, J. Philips in [?] enhance some important factors: users permissions, the listening of real-time information, the search mechanism, the data access and integration, and others, before the choice of a tool that allows the information collection. Besides the tool, is also important have an idea of what is need to explore because the use of a wrong technique of SMA could have bad business impact for the company. The majority of SMA techniques focus on modeling in order to understand the large range of data collected and support techniques, such as sentiment analysis, trend analysis and topic modeling, are the most commonly used [?].

2.2.1 Twitter

Twitter is a social network where people freely micro-blogging about any topic and, like any other social network, makes possible the connection between users around the world [?]. This social network has faced an exponential growth since its inception, and nowadays its users, which surpass 200 millions, produce around 500 millions tweets daily, performing a massive bunch of information that could be an ideal testbed for research projects on big data [?, ?]. The Twitter has been seen as a micro-blogging service since the length of a message never exceed the 140 characters limit [?] allowing better efficiency results in traditional extraction algorithms [?]. Overall this characteristic, micro-blogging text analysis has several attributes that challenges the classification tasks, as N. Banerjee et al. refers in their work. One of these attributes, like *Limited Context Information*, was already described below. The *Richness of Exchange* symbolizes the different kinds of posts in the micro-blog network, i.e. daily activities, individual or group conversations. The last one is the content *High Dimensionality* because of the informality and ambiguity characteristic present in these messages [?].

Twitter has not only evolved in terms of usability, but also in the purpose of its use. Before, Twitter could be seen as a personal diary, but nowadays people focus its use to share more news about real-time events than their own mood [?].

One of the advantages of Twitter compared to other social networks, are the tools that give developers access to the entire set of data inserted in the network. While Facebook, for example, does not provide private information about its users unless there are permissions to do so, Twitter allows the collection of all tweets from channels or directly from people in order to be analysed in any kind of project [?, ?]. Although this freedom in the collection of data, Twitter has also a an ethical perspective and a regulation must be accomplish by developers or researchers. The TOS (Twitter Terms of Service) was created in order to make known what can be done with the data and protect the users' rights [?].

2.2.2 Specificities of Twitter Services

Twitter provides to the developers/researchers community two distinct kinds of application programming interfaces (API) for data tracking, REST (Representational State Transfer) APIs and the Streaming API, where four types of objects can be collected. The communication with the APIs is based on TSL or HTTP protocol, depending on the defined specification for the project and its requirements. The types of the objects returned by the API to the client requests on the server could be:

Tweets representing the Twitter "basic atomic building block", or also known as the users update status and its content is limited by a 140 characters length. This object can be embedded with photos or videos, responses to other tweets, liked or unliked and deleted [**?**]. A Tweet has special characteristics as the possibility of reference other users by using the character @ followed by the intended user name (in example, @joaopereira) and can also have several keywords that distinguish the tweet content from others using the character #, performing the concept of *hashtag*. The hashtag system contribute to the good performance of the Twitter searching tools [**?**]. The API provides a lot of information about a certain tweet, since its inception data, location, text and the entities attached in it (like hashtags, users and URLs mentions), the detected language in the text, the owner user of the account, among other information [**?**].

Users normally represents a person or an entity. A User can perform several actions, such as "tweet", follow, create lists, be mentioned and also be a target of a massive search [**?**], like celebrities or politicians. Through the APIs, this object can provide information about the profile details, the list of recent tweets and its identified entities in the text, statistical information, like the number of tweets he has published, the list of followers or even the list of users he is following.

Entities represents metadata and additional information about the content of each tweet. This objects are entirely connected with the field description they assume and could range from hashtags, media elements, *urls* or user mentions [**?**].

Places is the last type of object that the Twitter APIs could retrieve and are related to the tweet location, through its corresponding geo coordinates [**?**].

Regarding the two main types of APIs that Twitter provides, the REST-type should be used when the developers need to access the core data, i.e. updating timelines, checking the status or look up to the user profile information. Unlike the REST API which needs a connection in every client request, the Streaming API keeps the connection persistent, reducing the latency in the requests due to the absence of multiple endpoints and for this reason the Streaming API is a good choice when the main goal is to extract massive bulk of data [**?**]. Some authors in their works adopted a hybrid approach and made a combination between this two types in order to have both volume and specificity of information [**?**, **?**].

2.2.3 Text Mining

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

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

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

Text mining shares some of the issues presented by the Natural Language Processing field. Once texts are usually performed by humans some associated problems can appear, such as spelling mistakes, wrong phrasal construction, slang among other. Before the "mining" of a text, it's important to apply some pre-processing steps in order eliminating noisy data from the primary analysis process. A. Stavrianou et al. cite this issues very well in they work and it can be seen in Table 2.1.

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

[1] https://google.com

Table 2.1: Text Mining Issues by A. Stavrianou [?]

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

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

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

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

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

[2] https://en.wikipedia.org/wiki/Levenshtein_distance

[3] https://wordnet.princeton.edu/

Tokenization serves to pick up all the terms presented in a text document and to achieve this it's necessary the split of the document into a stream of words implying the removal of the punctuation marks and non-text characters [?]. some authors also see tokenization as a text representation form since one of the most used models to represent texts is *Bag-of-words* (BoW). This model broke down texts into words and stores it in a vector being also presented the word frequency occurrence in the text. Hence, each word may represent a feature [?]. 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 [?].

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

2.3 Information Extraction

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

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

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

2.3.1 Name Entity Recognition and Name Entity Disambiguation

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

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

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

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

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

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

Y. Li et al. [?] define Named Entity Disambiguation as a "process of associating an entity name mentioned in a text to an entry, representing that entity, in a knowledge base". In the last few years, NED has been a target for a considerable number of research projects. The majority methods implemented to tackle this task are focused in three main features. Y. Li et al. [?] enhance *entity popularity* as a statistical one where there is a "assumption that the most prominent entity for a given entity mention is the most probable underlying entity for that mention". At this feature, a link between the term and its Wikipedia page reference is established. The *context similarity* is another feature and aims the complementarity of the *entity popularity* feature. This feature centers on similarity measures between the text in analysis and the content text of the Wikipedia page. Y. Li reveals that this feature is word-dependency since it's necessary that both texts shares identical words in order to produce expected results. *Topical Coherence* is the third feature and solves the emerged problem of the second one. This feature uses the Wikipedia cross-page links mechanism in order to look up for related-topics of other entities and makes a connection with the target entity in the disambiguation problem. Through this process the domain text is expanded, decreasing the word-dependency problem appeared in the second feature.

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

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

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

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

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

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

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

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

2.3.2 Content Filtering

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

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

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

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

The RepLab 2012 follows an identically problem as the WePS-3, the unknown-entity scenario. Some research teams follow the same approach as S. Yerva et al. [**?**] where the use of profiles describing each company mention to correctly filter the content. DAEDALUS [**?**] and OXYME [**?**]

[4] `http://nlp.uned.es/weps/`

[5] `http://nlp.uned.es/replab2013/`

[6] `https://wordnet.princeton.edu/`

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

tackle a manually exploration, such as the development of dictionaries and rules sets to the detection and the classification task, and the selection of feedback terms about the entities, respectively. The automatically methods were explored mainly with external resources. CIRGDISCO [?] and ILPS [?] used the Wikipedia, while BMEDIA [?] combined it with Freebase to extract related and unrelated concepts. CIRGDISCO proposed a two-step algorithm to solve the filtering task. The first step involves the extraction of the entity related-terms from the Wikipedia and further calculus of the IDF (Inverse Document Frequency) score for each term founded. The second step focus on the idea of concept term score propagation, i.e. to propagate the labels of the high-precision classified tweets to the remaining, in order to increase the recall measure. ILPS tackle the filtering task by using semanticising, where two probabilities are verified: Link Probability and Commonness. The first one represents the probability that an n-gram is linked to an Wikipedia page, while the second is the probability of an n-gram is linked to a certain concept. The ILPS group also used list aggregation and disambiguation techniques to carry out this task.

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

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

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

2.3.3 Topic Modeling

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

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

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

W. Zhao et al. [?] proposed another extension of the LDA model and named it Twitter-LDA [8]. Their model follows the idea that each tweet is about some topic, so instead of grouping the tweets into schemes and than extract some topic, they tackle each tweet as a singular problem and extract the target of the content. In their work, the evaluation of the model was made by comparing its effectiveness against the standard LDA model and the Author-topic model. The Twitter-LDA

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

results, obtained from a small set of topics in a preliminary test, have surpass the performance of the others models (standard LDA and Author-topic models).

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

2.4 Sentiment Analysis

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

2.4.1 Lexicon-based vs Machine-learning based

Sentiment analysis in Twitter can be divided in three different approaches relatively to the sentiment classification: lexicon-based, machine-learning based or even a hybrid approach between the previous two. In the first place it's necessary to talk about what features are relevant or not in order to tackle this problem. Aggarwal et al. [**?**] in his work refer some of the common features used in this problem:

- ***Term presence and frequency:*** groups of words, named *n-grams*, and the frequency they occur in the document;

- ***POS Tag***: the existence of adjectives can be relevant indicators to determine the opinion polarization;

- ***Opinion words and phrases:*** words that usually transmits some polarity, such as *good and bad*, or even whole phrases that don't have this type of words, e.g. "cost me an arm and a leg";
- ***Negation***: the existence of negative words that may change the opinion orientation, such as "I don't like apples" which means the same as *hate*.

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

- ***Lexicon-based:*** It's necessary human annotation. Starts with a small set of seed words and then a bootstrapping methodology is applied to expand the lexicon domain through the discovery of synonyms in external resources;
- ***Point-wise Mutual Information***: It's a statistical method where the co-occurrence level between a given word w and a class c is computed in order to see if a feature is or not correlated with the class. The formula of calculus is given in the equation 2.1,

  $$M_c(w) = \log(\frac{F(w).p_c(w)}{F(w).P_c}) = \log(\frac{p_c(w)}{P_c}) \tag{2.1}$$

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

- ***Chi-square***: It's another statistical method used to measure the correlation between the features and the classes 2.2,

  $$X_c^2 = \frac{n.F(w)^2.(p_c(w) - P_c)^2}{F(w).(1 - F(w)).P_c.(1 - P_c)} \tag{2.2}$$

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

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

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

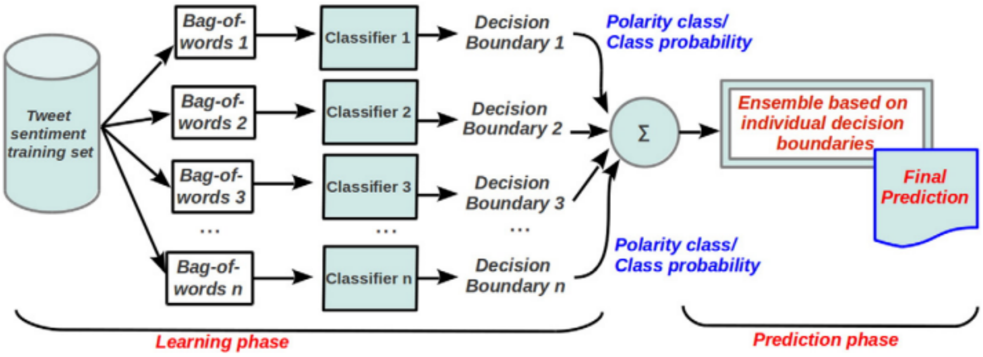
A. Giachanou et al. [**?**] divided the **machine-learning approaches** into three different categories: supervised learning, the classifiers ensembles and deep learning.

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

The Classifiers Ensembles approach is based on the combination of several classifiers to improve the performance of the classification task. The workflow of this approach can be verified in the figure 2.1. da Silva et al. [?] have explored the sentiment analysis in Twitter using a combination between Random Forests, Support Vector Machines, Multinomial Naïve Bayes and Linear Regression. The final classification decision in this kind of approach is usually made by majority voting between the models. In their work, da Silva et al. decided to calculate the average between all classification probabilities given by the models and applied that resulting value to the decision task.

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

Figure 2.1: The workflow of the classifiers ensembles by da Silva et al. [?]



their framework, the problem of imbalance classes was reduced which could be a great advantage to explore in our approach.

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

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

Contrary to the machine-learning approaches, the **lexicon-based approaches** doesn't depend on training data and features to classify the sentiment as positive, negative or neutral. In this approach the final sentiment classification is given by measuring the sentiment score of each term using external resources, such as dictionaries with a large number of previously evaluated terms (SentiWordNet [9], SenticNet [10], LIWC [11]). At 2010, Thelwall et al. [?] developed a lexicon-based algorithm, named SentiStrength [12], capable of detecting the sentiment value in messages that usually have informal language, such as tweets. The algorithm have access to 298 positive terms

[9] http://sentiwordnet.isti.cnr.it/
[10] http://sentic.net/
[11] http://liwc.wpengine.com/
[12] http://sentistrength.wlv.ac.uk/

and 465 negatives and to a list of emoticons, negations and booster words to increase or decrease the sentiment value of derived words. The authors have compared their algorithm with machine-learning approaches and the results were very interesting since in terms of accuracy as evaluation metric, SentiStrength has surpass the others.

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

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

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

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

[13] http://multiwordnet.fbk.eu/english/home.php

[14] http://wndomains.fbk.eu/wnaffect.html

by the linear equation of the model, which was enough to prove the efficiency of the approach explored by the author relatively to the polarity of a tweet.

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

2.4.2 Aspect-based Sentiment Analysis

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

In particular to the aspect extraction, an overview was already done in the subsection 2.3.3 where the topic-based approach was mentioned and described using the LDA model as the most used model. There are three more approaches that we can follow to discover relevant aspects in a document: frequency-based [?], ruled-based [?] and supervised learning [?, ?].

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

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

Supervised learning approaches are based in sequential learning methods, such as Conditional Random Fields (CRF) and Hidden Markov Models (HMM). The mentioned methods are similar because both of them attempt to discover patterns relative to an input data set. These types of methods are often used in the aspect extraction task of opinion mining. N. Jakob et al. [?] has built a classification model using the CRF algorithm in order to enhance the target of each opinion in the reviews. The domain of their dataset was constituted by four independently categories: movies, web-services, cars and cameras. Since the approach taken by the authors was through machine learning classifiers, they had the need of establishing a set of features to train the model. The features used for the classification vary from the string of each token, the POS tag for each token, the level of dependency between each token and the opinion expressed as well as its word distance, and, finally, the last feature is the opinion sentence itself to allow the CRF algorithm the ability of distinguish when a token is present or not in a sentence that is an opinion. Regarding the system validation, the authors applied also a 10-fold cross-validation to see if the model performance improves or not. As performance metrics to evaluate the system, they used the precision (Equation 2.3),

$$\frac{TP}{TP+FP} \tag{2.3}$$

the recall (Equation 2.4)

$$\frac{TP}{TP+FN} \tag{2.4}$$

and the F-measure (Equation 2.5) that is the harmonic mean between the previous two metrics.

$$2.\frac{precision.recall}{precision+recall} \tag{2.5}$$

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

Regarding the polarity sentiment classification, the majority of the works studied appeals to one of the approaches described in the section 2.4.1.

2.5 Conclusion

This chapter had the objective of review some basic concepts that may be relevant to contextualize the reader about the problem of performing analysis in social media streams, e.g. Twitter messages. Hence, the literature studied was divided in several points in order to have a overview about what is already done and what are the approaches that some author proposed to tackle each of the sub-problems that composed the main problem in this dissertation work. After a careful research, it was possible to identify that there are a great diversity of approaches to each sub-problem, whether it be disambiguation, filtering, topic detection or even sentiment analysis.

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

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

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

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

Chapter 3

Proposed Solution

In this chapter a description about the methodology to approach the problem presented in section 1.2 will be made, as well as the architecture designed to the proposed solution and the different modules which composed it.

3.1 Methodology

Once the problem to be solved has a scope of data analysis, in this case unstructured data like texts, it makes sense the use of an appropriate methodology for the field of interest.

The CRISP-DM (Cross Industry Standard Process for Data Mining) is the standard model used in the Data Mining field, providing several steps that the specialists may follow to tackle the problem of data analysis. The workflow of this methodology is presented in figure 3.1. The different model phases are:

1. **Business Understanding:** In this phase there are a definition of a work plan and what are the goals to achieve. At this dissertation case, the final goal is to provide a tool capable of helping entities or even ordinary citizens in the decision-making process relatively to some service in the city.

2. **Data Understanding:** The collection of the dataset is made in this phase as well as the identification of interesting patterns in it.

3. **Data Preparation:** Dataset cleaning and features engineering are made at this phase. The cleaning task in the text mining field can be seen as the removal of some words, the conversion of words to their root form, and the grouping of two or more words because, sometimes, there are mentions of technical terms or even negations (*n-grams*). The features engineering task is more complicate in the text mining field than in data mining. At common data mining analysis, the data is structured but, in texts, the data is unstructured which make more hardly the discovery of relevant features.

4. **Modeling:** The modeling phase represents the development of models that will be used in the classification tasks. At this point, it's made, also, some tuning parameters to optimize

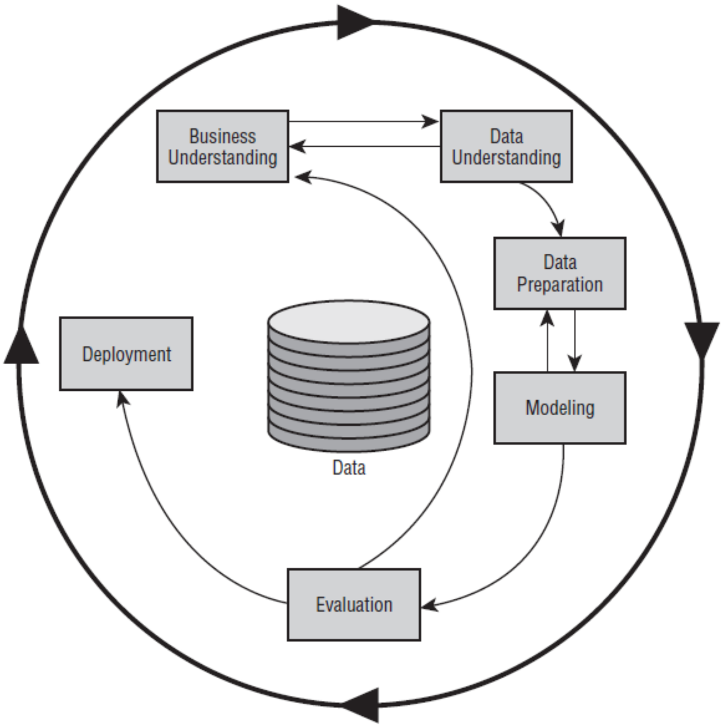


Figure 3.1: CRISP-DM workflow.

the models and, where appropriate, the returning to the Data Preparation phase in order to find other features.

5. **Evaluation:** The results given by the developed model may be evaluated is made in this step. To do that, some evaluation metrics are available (e.g. F1 measure, Area Under the ROC curve, Root Mean Squared Error, etc). The results should be good enough in order to achieve the defined objectives.

6. **Deployment:** The knowledge acquired by the model must be harnessed and, in this particular case, must be integrated with the rest of the system to take advantages of its functionality.

   One of the main advantages of this methodology is the independence it provides for the development of each of the modules that constitute the desired framework. Its last step allows an iterative integration of the developed models in order to produce the final tool.

3.2 Framework

In order to solve the problem described in Chapter 1, the proposed solution is the development of a framework whose usability will be directed to companies or even ordinary users and should be able to provide relevant information about a specific study scenario, namely the transportation area of a city.

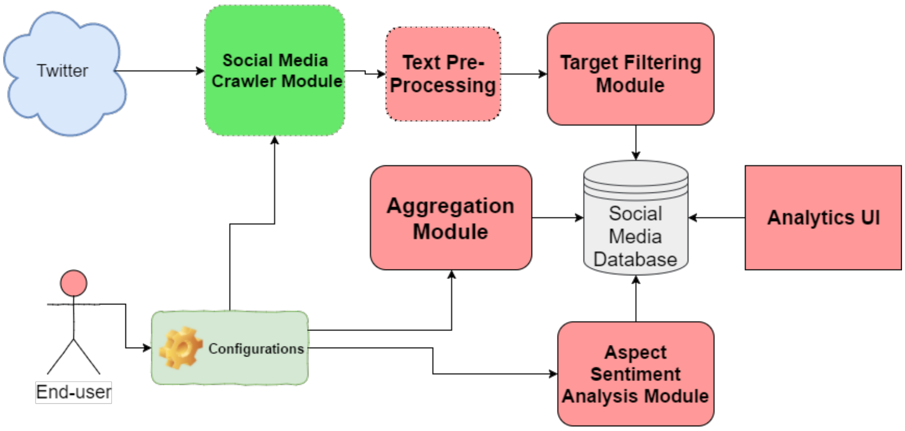


Figure 3.2: Framework design.

3.2.1 Architecture Design

The framework architecture was designed taking into account all the tasks that have to be performed to solve the different points in which the main problem can be decomposed. A visualization of the architecture can be seen in the figure 3.2.

Firstly, the data collection is made through a crawler that have connection with the open-sourced APIs provided by the social networks services, in this particular case, the Twitter Streaming API. The data must then be submitted to a preprocessing module in order to clear any noise that may exist in its content. After that, the data collection must be filtered since only messages related to the study scenario should be considered. After this primary three steps, the data is stored in a non-relational database, e.g. MongoDB. The Aspect-based Sentiment Analysis Module can access can access the stored data in order to extract the different aspects present in the messages and to classify their sentiment polarity, i.e. if a message has a positive, sengative or even neutral sentiment. The aggregation module also can access the data warehouse to perform continuously calculations of the already processed messages, grouping the results according to the configuration parameters stipulated by the user of the tool/framework. This step will be important to UI analytics, allowing the user immediately visualization of the analyzes perfomed by the aggregation module.

Apriori, there is no need to exist a communication between the different modules, with exception to the aggregation and the aspect-based sentiment analysis module. This is due to the need of the data be analyzed, firstly, with orientation to the aspects and the sentiment polarity and only then will it make sense to be aggregated. Therefore, to deal with this problem two solutions may be considered. The first one is a communication using REST services between the two modules. The second one can be the existence of two data warehouses, i.e. two different databases, in which the ABSA module accesses one of them and store its results in another that will be used by the aggregation module.

3.2.2 Social Media Crawler Module

The Social Media Crawler is already developed and it will be used to collect data regarding two different scenarios to test the final product of this dissertation, that is the proposed framework. The crawler is capable of collect data through a set of heuristics defined by the user:

- *Search terms*: Extracting all the messages that match with a specific term;
- *Users poll*: Extracting all the messages posted by a user, being the collection process made regarding its user name on Twitter;
- *Geo Reference*: Extracting all the messages inside a geolocalized area.

3.2.3 Preprocessing Module

This module should be able of clean the data collected by the crawler module and treat some special particularities of the messages. The cleaning can be the exclusion of some attributes that each tweet may have and are not necessary for the processing task. The particularities mentioned can range from stemming, that is the conversion of each word in the message to its root form (e.g. "running" is converted to "run"), removal of some stop words, such as abbreviations (e.g. "lol") or pronouns and determinants, and the classification of the part-of-speech tag to each word, i.e. if a word is an adjective, verb, adverb or even a noun.

3.2.4 Content Filtering Module

The filtering task of the framework module vary depending the study case scenario. If the target entities that composed our study case scenario were known, then only the classification of the messages that are related should happen. In this known scenario, an approach already proved and validate should be followed, as, for example, the approach performed by the best group at the RepLab 2013 filtering task. Therefore, it will be need the development of a binary classification model capable of automatically filter the relevant messages from our final datasets. Once there is no dataset collected for the final study case scenario, the development process will appeal to some golden-standard datasets in order to build the desired model. Apriori, the datasets provide by the RepLab 2013 contest will be chosen to support the development and posterior validation.

If the entities were not stipulated, then the first task should be the linking of the entities to a external knowledge base, like Wikipedia or Freebase to identify the real entity mentioned in the message and only after that the filtering process will be consumed.

3.2.5 Aspect-based Sentiment Analysis Module

The aspect-based sentiment analysis module of the framework is the most important module that will need more focus. Two tasks may be performed in this module: topic/aspects detection and the classification of the sentiment polarity.

Different approaches were identified in the literature review: ruled-based, sequential learning and topic-based. The results obtained by the authors demonstrate that the two last approaches can provide advantages since they don't depend on human-defined rules, like the ruled-based approach, and the aspect detection is made in a automatic way.

For the first case, the researches studied cover three different machine learning models that are commonly used. Two of them, follows the sequential learning (supervised) and the other topic-based: CRF (Conditional Random Fields) and HHM (Hidden Markov Model), LDA (Latent Dirichlet Allocation). Once Twitter messages have limited length (140 characters), the use of LDA models may be difficult since LDA models required a large ammount of information to work properly. Polling the data can be an alternative solution for the problem, but, since there is a short time to development of the framework, the chosen approach for this task will be the sequential learning using the CRF model.

The use of a model classifier in topic detection task implies the existence of a valid dataset. Like in the filtering task, there is none. Hence, the development of this model will explore golden-standard datasets, as, for example, the dataset from SemEval 2016 competition, in particular, the task of aspect-based sentiment analysis.

Lastly, the sentiment polarity classification could be based on one of three different approaches: lexicon-based, supervised learning or a hybrid between both. Tests will be made in all of them to see which can provide better results.

3.2.6 Aggregation Module and Analytics UI

The aggregation module is directly related with the analytics UI. The aggregation functions that will be used in this task should correspond to the configuration made by the user in order to provide the set of visual indicators he pretends. This visual indicators may be qualitative or quantitative statistics, such as line charts, pie charts or even KPIs (Key Performance Indicators) making the UI a kind of dashboard where the end-user can perform some coherent analysis of the results.

MongoDB provide several operations that make possible the group of values from multiple documents, returning a single one. Hence, it will be possible aggregate the sentiment present in the messages by days, weeks, entity names and aspects. It would be interesting the visualization of a heat map when the data collection configurations were by geo reference, allowing the companies an identification of possible existing problems in their services in specific areas of a city.

Chapter 4

Conclusions and Future Work

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

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

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

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

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

1. **Strengths**

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

2. **Weaknesses**

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

Opportunities

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

Threats

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

4.1 Expected Contributions

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

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

4.2 Task Planning and Scheduling

The tasks to be undertaken are mostly based in the modules described in Section 3.2.1 for the proposed framework architecture. The first task is to choose what are the specific scenarios that will serve to test the developed framework. A priori, two different scenarios will be enough to prove the good functionality and usability of the tool. Hence, the crawler module will be used to collect social media streams from the middle of February until, approximately, the ending of May. Meanwhile, the setup of the framework environment needs to be done. After this first step, the development of the modules will occur. The first module to tackle is the aspect-based sentiment analysis and the sub-module of preprocessing. With the estimation of a possible margin of error, these tasks are ready to employment in the begin or middle of April. The target filtering module will be developed, if everything is going as planned, between the beginning/middle of April until the middle of May. The remaining month of May will serve to work on the aggregation module

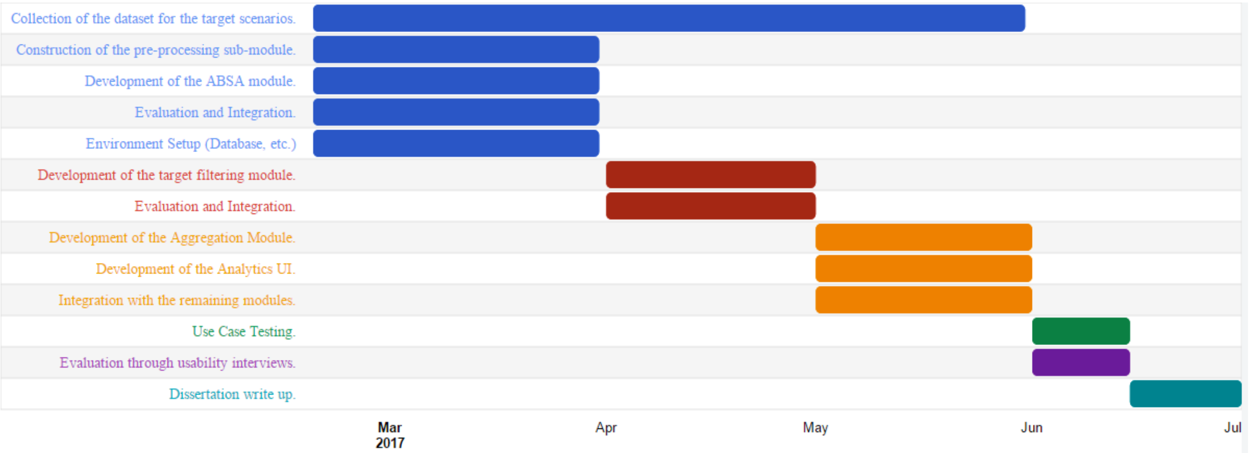


Figure 4.1: Dissertation working plan.

and the analytics UI. The month of June will be to test the final framework into the collected dataset about the two different scenarios. In order to evaluate the usability of the framework, it's planned the existence of a bunch of interviews to see if it's really possible that users of this tool are capable of immediately identify some conclusion from the analysis presented. This evaluation step will occur in the first two weeks of June, being the remaining two to the final dissertation report write up.

In the Figure 4.1 it's possible to visualize a Gantt chart scheduling according the mentioned tasks and the ideal scenario in case there are no delays.