# Social Media Text Processing and Semantic Analysis for Smart Cities

*João Filipe Figueiredo Pereira*

Master's thesis project supervised by *Prof. Rosaldo Rossetti* and *Pedro Saleiro* at *LIACC*

## 1. Motivation

With the rise of Social Media, people obtain and share information almost instantly on a 24/7 basis. Many research areas have tried to gain valuable insights into 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.

## 2. Problem Description

Mining Twitter data is 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 smart cities and transportation domains makes the task even more complex. The lack of gold standards datasets is the most disturbing problem since we are not able to benchmark any analysis performed to these aforementioned domains. The problem on focus in this dissertation is to find a way of demonstrating social media text analysis about cities/regions/countries that can be valuable for entities, governments or even ordinary citizens during decision-making policies, such as, for example, which city presents the most safety level or which mode of transport may an individual choose to travel across a city.

## 3. Goals

We address nine different goals in this dissertation to achieve our aim:

1. Continuous collection of geo-located tweets from multiple bounding boxes in parallel and in compliance with Twitter API usage limits
2. Tackling Twitter Geo API inconsistencies and filtering noisy tweets
3. Implement standard text pre-processing methods for social media texts
4. Content analysis using topic modeling and comparative characterization among different bounding boxes (e.g. cities)
5. Travel-related classification of tweets using supervised learning
6. Train word embeddings from geo-located tweets
7. Study the impact of word embeddings in travel-related classification
8. Creation of gold-standard data for travel-related supervised learning
9. Aggregation and visualization of results

## 4. Proposed Solution

In this section it is described the problem to be tackled in this dissertation as well as the designed and implemented framework that we proposed to solve it and the core modules composing it.

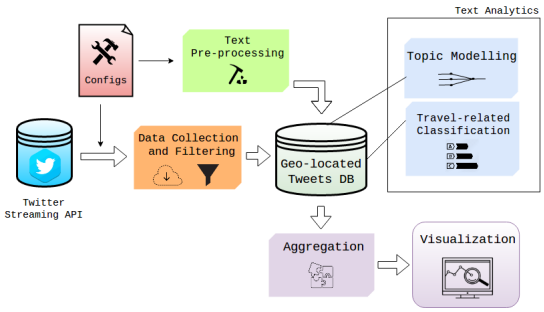


**Fig. 1 – Framework Architecture Overview**

### 4.1. Data Collection and Filtering

The data collection module was built using Tweepy, an open-source Python library to access the Twitter APIs. We explore the Twitter Streaming API using the locations heuristic that allows the retrieving of geo-located tweets through a bounding-box matching. The filtering task is due to the large amount of tweets in different languages comparative to the one spoken in the target city and to tweets retrieved by the API that are actually outside of the searching bounding-box.

### 4.2. Text Preprocessing

We apply a considerable group of text preprocessing operations to the messages such as lowercasing, lemmatization, tokenization, transformation of repeated characters, punctuation removal, cleaning of *metadata*, numerical symbols in the text as well as removal of stop and short words.

### 4.3. Topic Modelling

Social media, more specifically, microblog services are platforms where people publicly share their opinions and due to that they are seen as a rich source of content to explore. In order to mine such information, we implemented in our framework a generative module using topic modelling techniques. We chose to use Latent Dirichlet Allocation [1] in our module and tested it in two Brazilian cities, Rio de Janeiro and São Paulo, using a total of 9.5M of geo-located tweets. In the end 29 different topics (Figure 2) characterize both cities, being only 2 of them unique for each city.


**Fig. 2 – Latent topics of Rio de Janeiro and São Paulo**

### 4.4. Travel-related Classification

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. The transportation entities may take advantages from this kind of information since human mobility can be study, as well as citizens' opinions regarding the transportation services.

We conducted experiments 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. Word embeddings [2], 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 established the use of different groups of features to train our text classifiers, 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).

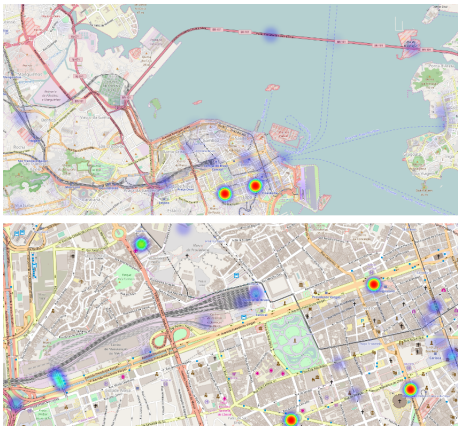

**Fig. 3 – Agglomeration of tweets classified as travel-related by our module for Rio de Janeiro**

## 5. Conclusion

In this dissertation, 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.

## References

[1] David M Blei, Andrew Y Ng, e Michael I Jordan. Latent dirichlet allocation. volume 3, páginas 993–1022, 2003.

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