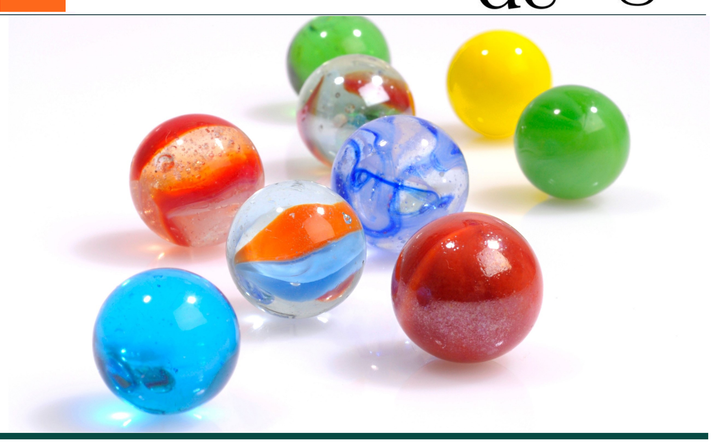
Spatio-temporal analysis of opinion in social media: outlier detection for the business intelligence area

Miguel Fernandes Caíña, PhD Student

Supervised by: Dr. Rebeca P. Díaz Redondo and Dr. Ana Fernández Vilas

Department of Telematics Engineering, University of Vigo



Motivation

- Internet has become a communication and expression platform, rather than just a static information source. Mailing lists, forums and chats have been part of it since the beginning, but over the last years, social networks have become the primary platform of communication for the majority of its users.
- The continuous flow of public information from forums and social networks makes possible to extract any kind of sentiment expressed about a product, service or brand. The aggregation of this data, including the impact of time and location, could be crucial in the success of a business decision.
- Natural Language Processing (NLP), defined as the ability of a system to process human language [1], is an artificial intelligence component that can be used to mine opinion and sentiment from social networks, and classify each post as being positive, negative or neutral towards a specific subject.
- This flow of opinions could be exploited by a Company, in order to verify the impact of a business decision or an external situation on the public's perception over its products, services or brands. Simply ignoring it could be harmful for the company's success.

Objectives

- The main objective of this PhD is to propose a general model applying different techniques (opinion mining, relevant topic identification, data and company connections, space-time scopes and real time analysis) that could be transformed into a solution that provides a high level view about products and services of interest, enabling decision making "as soon as possible", as well as post-mortem analysis of relevant events.
- A platform will be developed following the proposed model, and it should be able to provide insight about the impact of events on the public's perception over related brands, services or products.

Research Plan

- A comprehensive review of the relevant literature in the fields of opinion and sentiment mining, topic disambiguation, space-time scopes and real time analysis, outlining the state-of-the-art in the fields is being performed in order to gain critical insights.
- The development is following an iterative and incremental approach, and is divided in several phases:

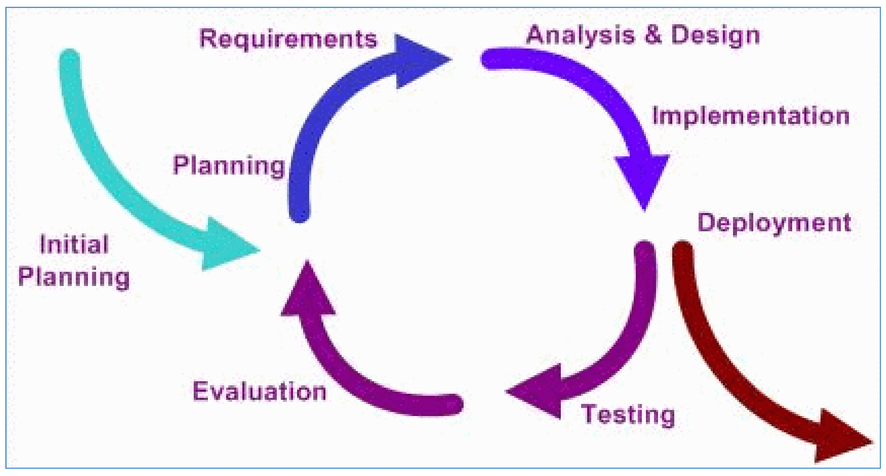
Phase 1 (Completed) - Conceptual Test: A basic approach using simple scripts written in Perl, to extract user's tweets, preprocess them and mine their opinions using a basic algorithm.

Phase 2 (Completed) - Platform Definition: Definition a Java enterprise application model, using MongoDB and PostgreSQL as data warehouses, and an extensible architecture to be able to accommodate the four main modules, and their subcomponents.

Phase 3 (Completed) - Basic Modules: Development of basic algorithms for each components, following the same principles as in Phase 1.

Phase 4 - Advanced Modules: Integration of several state-of-the-art techniques for each area, specially in preprocessing and sentiment analysis. Automatic anomalies detection, external API support and live extraction will also be part of this phase.

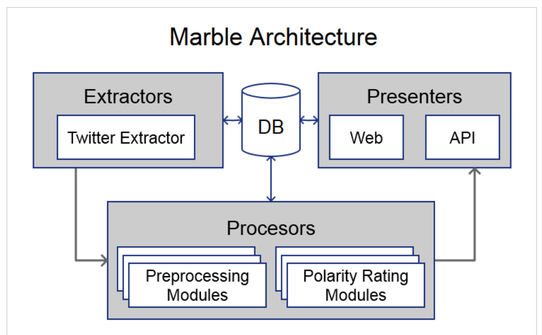
- Validation and assessment of the results will be based on a statistical approach, and the project success will be evaluated using this criteria.



Iterative and Incremental Development

Preliminary Results

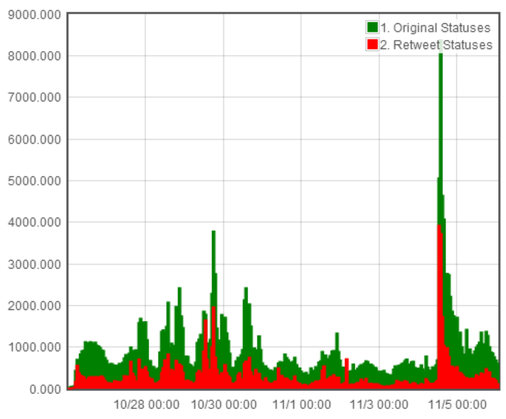
- A platform, named ***Marble*** has been developed, and integrates all the main components that could provide the achievement of the objectives. This model is flexible enough to allow an implementation based on its guidelines. The main components are:



Data Extractor: an extraction module capable of extracting information from Twitter related to a subject. If needed, the module will be expanded to cover other social networks.

Preprocessor: a module incorporating NLP processing techniques, stemming and lemmatization capabilities, synonyms recognition and disambiguation practices, that will be in charge of converting the raw data into information for the opinion mining module. This module would be configurable, in order to use different processing techniques, depending on the nature of the data.

Sentiment Analysis: a module in charge of extracting the opinion expressed in each message and define a polarity, using a combination of sentiment analysis techniques and heuristics, which will allow to identify specific characteristics of opinion on each user interaction.



Presenter: a presentation module, responsible for extracting relevant information from the mined opinion and correlating it to manually identified events. The module will also be able to detect "special situations" not related to any of the known events, in order to discover unidentified incidents

- Phases 1 to 3 have been completed, and a conceptual test was performed using shell scripts.
- The ***Marble*** platform have been implemented and tested with the data extracted in the conceptual test. Two subjects were selected for the experiment: *Blackberry* and *Whatsapp*. It is publicly accessible at http://marble.miguelfc.com, and the source code is hosted on github.
- Opinion and impact of events have been extracted, and the results of the experiment were presented in the **KDIR 2014 (Scopus Index) conference**, under the name: "Marble Initiative - Monitoring the Impact of Events on Customers Opinion" [2].
- The platform currently uses a unsupervised algorithm with limited functionality.
- The Validation results obtained using the Movie Review Dataset were not good enough, so there is a need to develop better and more efficient processing modules that could provide more precise and accurate results.
- The platform was adapted to allow several processing algorithms to be used in parallel, in order to use a combination of algorithms in the processing stage. A supervised and trained algorithm is needed to achieve better precision, and several techniques are being evaluated to be added in the next iterations, including Naive Bayes, Part of Speech tagging and Support Vector Machines.

Next Year Planning

- **Development of new processing modules (Phase 4):**

  The platform will be extended with several processing modules, including self-validation capabilities using different tagged datasets, in order to provide a fast method of evaluation in different fields of use. This modules will included supervised and unsupervised algorithms.

  The presentation layer will be expanded to support different approaches to the data, covering traditional statistics views as well as modern visual methods.
- **Evaluation and Analysis of the Processing Modules:**

  Each processing module will be evaluated and validated, in terms of its recall, precision and effectiveness. The results of the evaluation will be included in a paper that will be submitted to a journal (JCR indexed).

References

[1] Preeti & BrahmaleenKaurSidhu 2013, "Natural Language Processing", *International Journal of Computer Technology and Applications*, vol. 4, no. 5, pp. 751-758.

[2] Fernandes M., Díaz, R. & Fernández A. 2014, "Marble Initiative - Monitoring the impact of events on customers opinion", *In Proc. International Conference on Knowledge Discovery and Information Retrieval (KDIR).*