

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.
- Modularity, adaptability and openness should be the design principles on which the platform should be based. New functionality should be easy to add, and existing one should be easy to expand or modify.

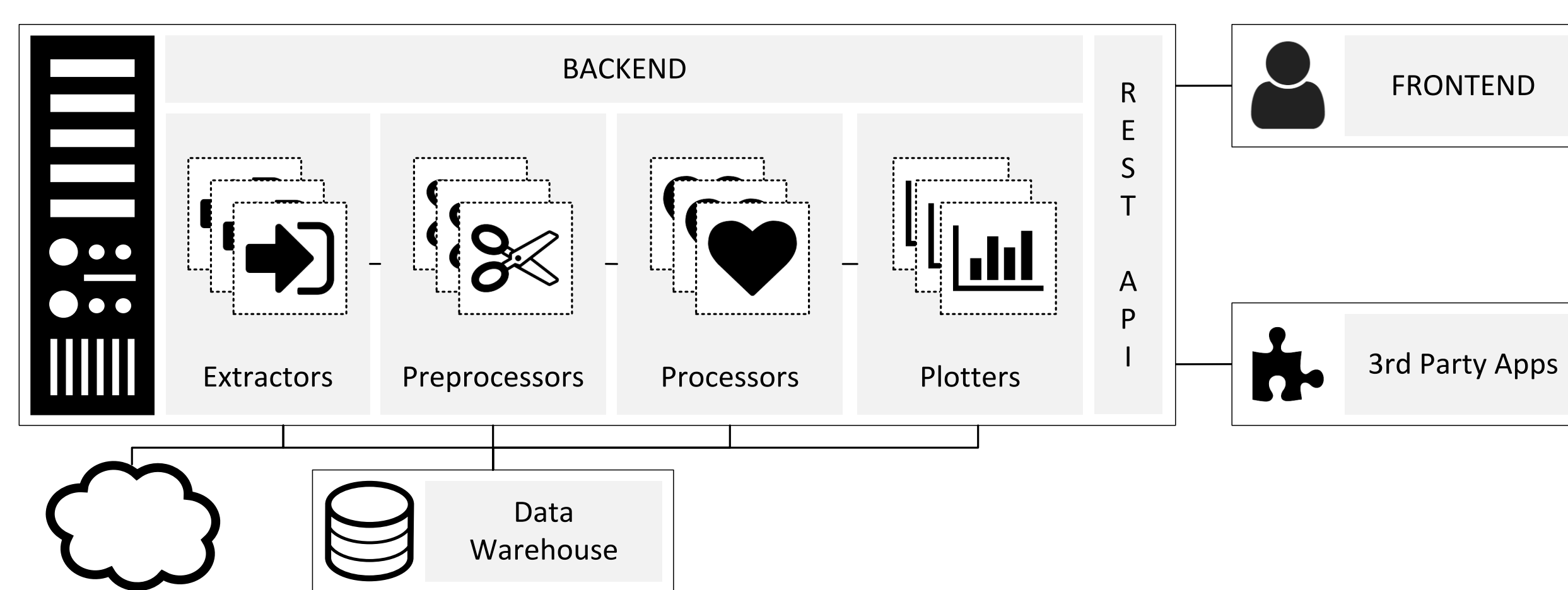
Research Plan

- 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 techniques is being performed continually 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:** Design a Java enterprise application model, using MongoDB as a data warehouse, 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 (Completed) - Advanced Modules:** Integration of several state-of-the-art techniques for each area, specially in preprocessing and sentiment analysis, external API support and live extraction.
- **Phase 5 - Experiments and Validation:** The platform will be used to perform analysis over different topics like music charts or top lists, and their correlation with the social network results.


- Validation and assessment of the results will be based on a statistical approach, and the project success will be evaluated using this criteria.
- The modularity of the platform is essential to provide an extensible framework for other projects on this area.
- Insights based on location, timing, or impact should be easily obtained from the platform, along with correlation conclusions. This can be provided in form of reports or visualizations.


Results




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

Extractor: A main 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: Modules 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 Analyzers: Modules in charge of extracting the opinion expressed in each message and define the polarity and intensity of the sentiment, using a combination of sentiment analysis techniques and heuristics, which will allow to identify specific characteristics of opinion on each user interaction. 

Plotters: Presentation modules, 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. 

- ▶ Marble is currently being used as a research media under the project "INRISCO: ANALISIS DE COMUNIDADES BASADO EN MINERIA SOCIAL (2015-2017). Ministerio de Ciencia e Innovación. Proyectos de I+D+I del programa estatal de investigación, desarrollo e innovación orientada a los retos de la sociedad (TEC2014-54335-C4-3-R).", specifically to extract useful data in collaboration with UPC and UC3M.
- ▶ The core of the platform was redesigned in order to separate the backend and the frontend. The backend is in charge of all the processing work, and exposes a REST API that is being used by the official frontend but is available to 3rd party applications.
- ▶ The platform has been containerized, in order to modularize processor and plotter parts, and enable extensibility without needing to recompile or restructure the core app.
- ▶ Every module of the platform is publicly available in Docker Hub for anybody to use it.
- ▶ New python-based processing algorithms are available to extract polarity ratings for establishing the polarity and intensity of sentiment of tweets: one using scikit learn [3] libraries and Support Vector Machines, and another one based on VADER model [4].
- ▶ The project was registered in the General Registry of Intellectual Property of Spain, Registry number 03/2018/392.
- ▶ Experiments on music chart positions are being run and tested for correlations between the tweets and the rankings. Other experiments for stock prices vs. tweets are also being run.

Next Year Planning

- **Experiments and Validation (Phase 5):**
The platform will be used to perform several experiments over real life subjects. Our focus now is on live events (like Eurovision contest) where the public influence directly the results of the show.
Processing connectors to cloud commercial tools like GCP Natural Language models.
- **Publish an article in the International Journal of Business Intelligence and Data Mining.**
- **PhD Document Preparation and Defence.**

References

- [1] J. Hirschberg and C. D. Manning, "Advances in natural language processing," Science, vol. 349, no. 6245, p. 261 LP-266, 2015.
- [2] M. Fernandes, R. Díaz Redondo, and A. Vilas, "Marble Initiative. Monitoring the Impact of Events on Customers Opinion", In Proc. International Conference on Knowledge Discovery and Information Retrieval (KDIR). 2014.
- [3] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, no. Oct, pp. 2825–2830, 2011.
- [4] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text." 2014.