TRABAJO FIN DE GRADO

Título:	Prototipo de sistema de predicción de la bolsa de valores basado en el análisis de la emoción y el sentimiento de Twitter
Título (inglés):	Prototype of stock prediction system based on Twitter emotion and sentiment analysis
Autor:	Marcos Torres López
Tutor:	Carlos A. Iglesias Fernández
Departamento:	Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL CALIFICADOR

Presidente:	Gregorio Fernández Fernández
Vocal:	Mercedes Garijo Ayestarán
Secretario:	Carlos Ángel Iglesias Fernández
Suplente:	Tomás Robles Valladares

FECHA DE LECTURA:

CALIFICACIÓN:

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN

Departamento de Ingeniería de Sistemas Telemáticos
Grupo de Sistemas Inteligentes

TRABAJO FIN DE GRADO

PROTOTYPE OF STOCK PREDICTION SYSTEM BASED ON TWITTER EMOTION AND SENTIMENT ANALYSIS

Marcos Torres López

Junio de 2014

Resumen

Esta memoria recoge los resultados de un proyecto cuyo principal objetivo es intentar predecir la bolsa de valores a través del sentimiento y la emoción de Twitter.

Se introduce en primer lugar la posibilidad de predecir el mercado de valores debido a la refutación de la Hipótesis del Mercado Eficiente por parte de las Finanzas Conductuales. Éstas últimas explican que los movimientos del mercado son motivados por factores sentimentales y emocionales y, por tanto, la medición de los mismos ayudaría en la predicción de la bolsa. Se elige Twitter para medir estos factores por ser una red social donde las opiniones se expresan de forma rápida y concisa.

Se presentan después las diferentes tecnologías utilizadas para la realización del proyecto, destacando la aplicación de Sefarad, que es la utilizada para mostrar los resultados de nuestro trabajo. Esta aplicación resulta idónea ya que permite mostrar de la forma más óptima los datos que se tienen a través de la creación de widgets propios.

Se presenta también la arquitectura del sistema basada en diferentes módulos encargados de la recolección, análisis, agregación, experimentación y visualización de los datos. Se tiene un capítulo específico dedicado a la experimentación donde se muestran los diferentes tests estadísticos realizados, así como los resultados de los mismos. Se encuentran ciertos valores (como son Google y Vodafone) donde los resultados son mejores cuando se introduce la emoción y el sentimiento en los tests.

Se presentan finalmente las conclusiones extraídas de este trabajo así como las posibles líneas de acción para mejorar el trabajo en el futuro.

Palabras clave: Sentimiento, Emoción, Análisis semántico, Predicción, Bolsa de Valores, Finanzas

Abstract

This work collects the results from a project whose purpose is to try to predict the stock market through the sentiment and emotion of Twitter.

First, we introduce the possibility to predict the stock market due to the refutation of the Efficient Market Hypothesis by the Behavioural Finance. This last explains that the movements of the market are motivated by sentimental and emotional factor, therefore, the measurement of those factors would help to predict the stocks' movements. We chose Twitter to measure the emotion and the sentiment due to its characteristic of being a social network where opinions are expressed in a fast and concise way.

In addition, we present the different technologies used in the project, standing out the Sefarad application, which is used to show the results of the work. Sefarad is the ideal application because it allows us to present the obtained data in the best way thanks to the possibility of creating our own widgets.

Furthermore, we present the architecture of the system, which is based on different modules whose purpose are the recollection, analysis, aggregation, experimentation and visualization of the data. We have a specific chapter for the experimentation where it is presented the statistical tests executed as well as the results of them. We found some stocks (as Google and Vodafone) whose results are better when we introduce the emotion and sentiment in the tests.

Finally, we present the conclusions of the work and the possible future work that could be done in order to improve the project.

Keywords: Sentiment, Emotion, Semantic Analysis, Prediction, Stock Market, Finance

Agradecimientos

A mis padres.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Overview

The main objective of this work is to study the possible correlation that exists between the stock market and the sentiment and emotion in Twitter through its tweets, so it will be possible to predict the future movements of the stocks' prices. This work is done on the context of the R&D Financial Twitter Tracker project whose main objective is enriching content with information extracted from financial social media like Twitter and detecting financial demand for new content on certain topics. The coordinator of the Financial Twitter Tracker project is Paradigma Tecnológico.

The possible correlation between stock markets and sentiment and emotion is based on different studies [1] which show that investors are not only affected by rational reasons but also by emotions when they are looking for the better place to put their money. This fact is widely studied in the Behavioural Finance which bases its assumptions on the falsity of the Efficient Market Hypothesis. The study of this possible correlation will be done firstly, by retrieving a large amount of tweets from two different sources (an Infochimps dataset[1], and tweets provided by Paradigma Tecnológico) and secondly by making a semantic analysis

[1] http://www.infochimps.com/datasets/twitter-census-stock-tweets

to obtain the sentiment and emotion of each tweet. Once we have this data available, we will pass it through different statistical tests to get the correlation and the accuracy of the prediction.

We will also try to predict the future risk of the market by trying to predict the Volatility Index of the Nasdaq 100 market[2] (VXN), which is fully related with the existence of risk in the market. We will do that by relating the VXN with fear emotion as it is that emotion which can be related to the risk of the market: the bigger the risk, the bigger the fear.

In addition, the results obtained in the work will be presented and displayed using the Financial Twitter Tracker application. The presentation is a big challenge due to the high amount of data that we are trying to manage, so we are going to use MongoDB as it allows to store a lot of information in an efficient way. We will also create new widgets for the mentioned application in order to display our results in the best possible way.

1.2 Methodology

The methodology of the work consists in the following steps:

First, we have to recollect all the tweets, obtained from an infochimps dataset and Paradigma Tecnológico. The infochimps dataset was focused on financial related tweets and from public companies that would be useful in our work. This dataset did not have the tweet itself but his identifier, so we had to recover the tweets through the Twitter API, selecting a period from 01/05/2009 to 30/03/2010. The tweets from Paradigma Tecnológico were from a period from 12/12/2013 to 13/03/2014.

Later, we begin to analyse semantically the tweets to obtain the sentiment (positive, neutral or negative) and the emotions (anger, disgust, fear, happiness, surprise and sadness) saving them in MongoDB.

Once we have the sentiment and emotions from the tweets, we pass through the statistical tests. First, we aggregated all the tweets by day, obtaining a positive index, a negative index and an index for every dimension of the emotion. From here, we follow two studies: the Granger Causality Test, which determines if given two sets one of them causes the other; and the second one is the prediction of the stock market based on the models proposed in [2] and later the comparison between the prediction and the real value.

Finally, we build the visual environment by configuring the widgets of the application,

[2]http://www.cboe.com/micro/vxn/

so we can display the most interesting data.

1.3 Description of the following chapters

Chapter 1 provides an introduction of the work, the main objectives of it and the methodology followed to obtain the results.

Chapter 2 explains why the prediction of the market could be possible by analysing the sentiment and emotion of the people. It explains the behavioural finance which refutes the Efficient Market Hypothesis and concludes that investors are influenced by psychological factors when they invest their money.

Chapter 3 describes the available technologies that are going to be used in the project.

Chapter 4 describes the full architecture of the project itself, built it in different modules that are deeply explained in the chapter.

Chapter 5 conforms the most important chapter of the work since it is composed by the results of the carried tests. It is formed by the correlation and the predictions obtained as well as the prediction of the risk of the market through the study of the correlation between the VXN and the fear emotion.

Chapter 6 sums up the conclusions reached during the work and gives some future work that could be done to improve the features and results of the project.

CHAPTER 2

Stock Prediction Based on Sentiment and Emotion Analysis

We cannot talk about predicting the stock market without taking into account the Efficient Market Hypothesis. This efficiency would make impossible to beat the market and because of it, to predict it. The refutation of the hypothesis will come with Behavioural Finance which adopts a different vision of the market in which investors are not completely rational people. Finally we will describe how new technologies and specifically Twitter will lead us to measure the investor sentiment so we can try to predict the market with it.

2.1 Efficient Market Hypothesis

The Efficient Market Hypothesis states that in an efficient market all prices should rapidly reflect all available information, which comprises past events and future expectations [3]. This hypothesis is based on the Expected Return Model and Random Walk Model described below [4].

The Expected Return Model explains how the prices of a security move to a specific price depending on the new public information [5]. We understand security as a financial

instrument representing the ownership of a company (stock), the right to be paid a specific amount of money (bond) or any rights of ownership (futures or options). The model states that future prices are function of expected returns. The return of a security is defined as follows:

$$r_{j,t+1} = \frac{p_{j,t+1} - p_{j,t}}{p_{j,t}}$$

where $r_{j,t+1}$ is the return of the security j at time t+1 and $p_{j,t}$ is the price of security j at time t. The future price of the security is related with the expected return as follows:

$$E(\widetilde{p}_{j,t+1}) = [1 + E(\widetilde{r}_{j,t+1}|\Phi_t)] * p_{j,t}$$

where $E()$ is the expected value operator, $\widetilde{p}$ and $\widetilde{r}$ are random variables and Φ_t is the symbol to represent the general information assumed to be "fully reflected" in the price of the security at time t. As it could be seen, the expected return is projected on the information Φ_t. The equilibrium expected return should be determined by the particular return theory at hand, but what this expression shows us is that the term Φ_t is always used to determine the future price.

The Random Walk model explains that the price of a security in an efficient market will follow a random walk where successive price changes (or successive one-period returns) are independent and also identically distributed [5]. Formally:

$$f(r_{j,t+1}|\Phi_t) = f(r_{j,t+1})$$

In the random walk model, the Φ_t is supposed to include just the past return history. The distributions of returns are assumed to be the same for all t and independent of available historic information. Of course the model does not say that past information is of no value in assessing distributions of future returns, instead, the model says that the sequence (or order) of the past return is of no consequence in assessing distributions of future returns.

The implications of the market efficiency depend on the grade of efficiency. In a weak-form efficient market, it is impossible to predict future return since the current prices include past prices and volumes. Extra returns can be obtained by future or privileged information. In a semi-strong efficient market, all public information (plus past prices and volumes) are reflected in current prices. This public information is comprised by annual reports, newspaper and magazine articles and new contracts among others, and extra returns could be only obtained by privileged information. In a strong efficient market, all available information (news and past prices and volumes) is reflected on the security price and also, investors with privileged information are forbidden to invest in the markets. In this form of efficient market, nobody is able to beat the market and to have a greater return than the market

return itself.

2.2 Behavioural Finance

One of the assumptions of the Efficient Market Hypothesis is that markets are formed by rational investors. Due to the failure in the developed models based on this hypothesis, researchers focused its attention to Behavioural Finance. This field is built on two blocks: On one side, limits to arbitrage which explains why there could be difficulties for rational traders to undo the dislocation caused by the less rational traders; and on the other side, human psychology, understanding that investors are not as rational as the efficient market hypothesis states [6].

Limits to Arbitrage: in efficient markets, prices tend to go to the fundamental value, which is all the cash flows that will generate a company discounted to the present by an interest rate. When there is a deviation in the price of a company (the price is higher or lower than the fundamental value) the Efficient Market Hypothesis states that it will have been created an incredible opportunity to investors and a lot of them will tend to make profit from it. Because a lot of rational investors will try to end with the deviation, the price will return to the fundamental value. Instead, Behavioural Finance has a different point of view. Although it is true that if an incredible opportunity is presented to investors they will make profit out it, they should question if a deviation is such a good opportunity that they will rush to make the price go again to the fundamental value. There are some factors that should be addressed to understand why a deviation in a price is not always the perfect opportunity:

- Fundamental risk refers to the risk which makes the deviation from the fundamental value get worse due to unexpected bad (or good) news about the company. This news could lead to a greater deviation and the investor may be aware of it. Therefore, the investor, fearing that the deviation worsens, will not buy (or sell) the security and the price will not return to the fundamental value. Although the risk could be reduced by selling (or buying) a substitute security, there is never the perfect security to hedge the risk.

- Noise trader risk is the risk caused by the worsening of the deviation in the short run due to the action of more investors producing a bigger gap. This risk matters because it could make the arbitrageurs (or rational investors trying to reduce the deviation) close their positions with losses. In addition, the risk is even worse when there is

another person managing the money of other people, because these people could ask to the manager to close the position due to this noise trader risk, causing big losses in their accounts.

- Implementation costs are some costs that the investor has to face and which can make the investment less attractive. Commission costs, in which you have to pay a certain amount to enter in the market, bid-ask spread, which is the difference between the price paid when you buy and when you sell any security, short sale constraints, as the borrowing costs to short sale a security. Sometimes you are not going to find any person who will lend you the securities to sell them. There could be also legal constraints as some managers are not able to make some operations in the markets due to the regulation. All of these issues obstruct to close the gap of the deviation from the fundamental value.

Psychology: investors, far from being rational, make investments driven by sentiment and emotion. They are affected by cognitive biases and they make decisions basing them on incomplete and noisy information. This was one of the biggest errors in the Efficient Market Hypothesis, because it states that investors are always rational. Behavioural Finance states that capital markets are formed by people and therefore the decisions made on them follow the human psychology [7].

This idea is supported by the numerous bubbles formed over time in stock markets. The technological bubble in the late 1990s is a paradigmatic example of this fact. Investor sentiment pushed the prices really high, following a mimic behaviour among investors. This bubble constitutes an example to noise trader risk, as even arbitrageurs investors could not do anything to solve the deviations on prices.

Another characteristic of the human psychology that affects markets is the overreaction to news, which leads to a volatility increase in the market. We can observe this every time the chairman of the FED or the chairman of the ECB gives a speech. The movements are really aggressive when one of them talks in public. Also, as another proof of the human psychology in markets, there have been studied some kind of relations between the public mood and the return on the stock exchange. For example, the Dow Jones Industrial Average has a worse performance when it rains in Central Park [3], as well as, previous high levels of geomagnetic activity have a negative impact on today returns.

The question now is not if sentiment affects stocks, but rather how sentiment affects stocks and how can we measure its effects. Measuring investor sentiment is not straightforward due to the amount of investors in the markets and the constant change in them. We

can find several examples trying to accomplish this task, for example, we have the investor surveys in which an institution asks several investors for its sentiment or the forecast of the market. There are also different attempts to measure the sentiment by exogenous changes in human emotions such as weather. Another way to complete the task is by looking to the individual inexperienced investors' trades because these inexperienced investors are more affected by sentiment than expert ones (as example, in the technological bubble, young investors bought more than old investors).

2.3 Twitter Sentiment

The traditional methods which try to measure investor sentiment (like the explained before), have several problems that should be addressed so we can improve their measurement. These problems are related with the cost and frequency of the collection of the data. First, in terms of costs, there are both monetary costs and time costs. In the investor surveys it is really costly to ask each investor its opinion, and it consumes a lot of time to recollect the data. Also, this kind of surveys are made as much as one per day, which is a problem considering the fast pace of capital markets. This fast pace could lead in a sentiment change in less than hours, due to any positive or negative news [7].

These two problems have been solved during the past years thanks to the development of new communication technologies. This development has led to more communications among people and specifically investors, and also has permitted to obtain more easily the sentiment of them [8]. Diverse studies has shown the correlation between the opinions or sentiments expressed through these technologies and the stock markets return [9]. Significant progress has been made in sentiment tracking techniques in the last years. These techniques allow us to extract indicators of public mood from social media content such as blogs or Twitter feeds [10]. Although each tweet does not contain enough information due to the short length of it, aggregating millions of tweets can give us an accurate representation of public mood and sentiment [1]. Finally, another study suggests that microblogs are an ideal early warning about company prices as they are freely, rapidly updated and provide spontaneous glimpses into the opinions and sentiments of consumer [11].

CHAPTER 3

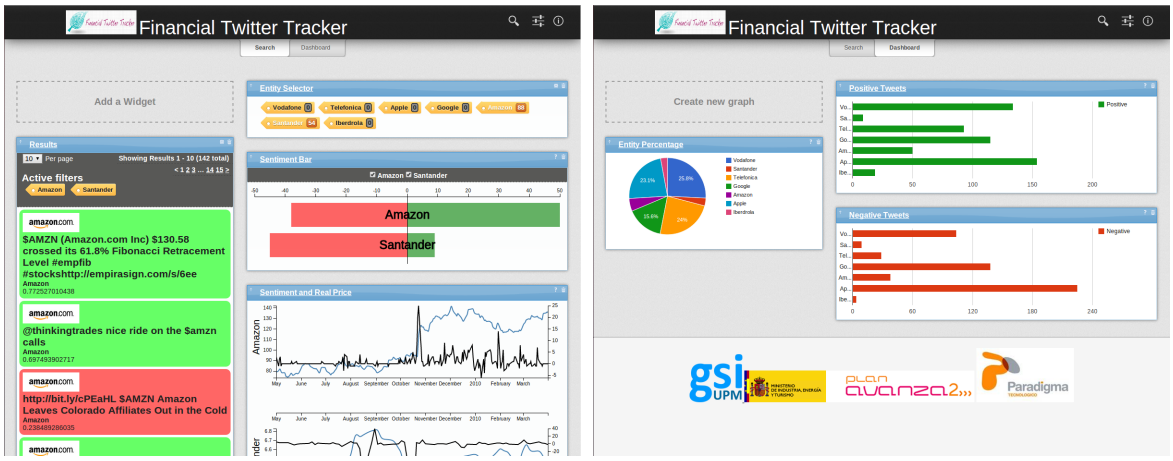
Enabling Technologies

We will describe in the following sections the different technologies that we have used during the progress of the project. All of them were chosen because of their particular characteristics and features, which helped us to develop the project or the different parts of it.

3.1 Sefarad

Sefarad[1] is a web application whose purpose is providing a semantic front end to Linked Open Data (LOD) [12]. It allows the user to configure a dashboard to visualize different perspectives of such datasets. Two screens are defined, (i) search and (ii) control panel. It is in the search screen (Figure 3.1a) where semantic faceted search can be carried out with the results shown in the widgets. It is in the control panel (Figure 3.1b) screen where statistics about the dataset are visualized. Developed in HTML5, Sefarad follows a Model View View-Model (MVVM) pattern with the Knockout[2] framework. This framework is specially appropriate for dynamic user interfaces due to the dependency tracking which automatically updates the correct parts of the interface after a change in the data model.

[1] http://demos.gsi.dit.upm.es/ftt/index.html
[2] http://knockoutjs.com/



(a) Search screen (b) Control panel screen

Figure 3.1: Sefarad Application

To do these updates, the framework allows us to easily connect parts of the interface with the data model through declarative bindings.

This application was developed to provide the user with the capability of creating his own widgets using the powerful framework of Knockout.js. Thanks to that framework it is really easy to develop the widgets without being worried that they are going to display the correct information The creation of new widgets is an incredible feature of Sefarad because it allows us to visualize the data in the most adequate way. For this purpose, the application specifies how to create a new Javascript file in which it should be placed a Javascript object. This object should be as the following example:

```
var widgetName = {
  // Widget name.
  name: "Name of the widget",
  // Widget description.
  description: "Description of the widget",
  // Path to the image of the widget.
  img: "path/to/img",
  // Type of the widget.
  type: "widgetType",
  // [OPTIONAL] data taken from this field.
  field: "field",
  // Category of the widget (1: textFilter, 2: numericFilter, 3: graph, 5:
      results)
  cat: 3,

  render: function () {
    var id = 'D3' + Math.floor(Math.random() * 10001);
```

```
    var field = widgetD3.field || "";
    vm.activeWidgetsLeft.push({"id":ko.observable(id),"title": ko.observable(
        widgetD3.name), "type": ko.observable(widgetD3.type), "field": ko.
        observable(field),"collapsed": ko.observable(false)});

    // widgetD3.paint(field, id, widgetD3.type);
    widgetD3.paint(id);
  },

  // paint: function (field, id, type) {
  paint: function (id) {

    // Code to paint

  }
};
```

Those characteristics of Knockout are really useful in our application as you could select the stocks that you want to visualize and all the widgets will update the information accordingly to it.

Sefarad uses two external modules: a semantic repository and an indexer based on Solr technology [13]. It also uses the Sefarad Packager Tool through a grunt application.

3.2 D3.js

The library D3.js[3] is a visualization library that allows us to create complex graphics with little effort. It is based on binding data to the Document Object Model (DOM) so it could apply data-driven transformations to it.

3.3 MongoDB

MongoDB[4] is a document-oriented database written in C++. MongoDB is a free and open-source software distributed under the GNU Affero General Public License and the Apache License. The database is structured in collections and the collections are structured in documents which are BSON objects (binary JSON) which makes ideal the mapping to different programming language data types. These documents are structured in single key-

[3] http://d3js.org/
[4] https://www.mongodb.org/

value pairs conforming a really versatile structure for every need. To retrieve the documents from a collection, MongoDB specifies the different queries allowing the user to select the documents by conditional statements such as the existence or not of a key, or that the values from the document are greater than a given value.

MongoDB allows us to replicate all the information of the database in a slave so if the master fails, all the operations will be handled by him, providing high availability to the data. It also allows to have multiple machines working together as one database, which makes scalability really easy and powerful as you just have to increase the number of machines. All these characteristics and the ones said before, make MongoDB the ideal database to store all the information that we have.

3.4 Prediction Algorithms

We cannot talk about prediction in stock prices without taking into account Granger Causality Test [4]. This test tries to check if Y time series does not strictly cause X time series, which means that if changes in past values of Y do not induce changes in present values of X. As [1], we use this test to try to find a relation between past values of sentiment and emotion in Twitter and the stock market. The results obtained in [1] state that there is such relation and therefore sentiment (emotion is not studied in this research) does Granger cause stock market movements.

In a formal way [14], we can say that time series Y does not cause time series X if the following expression holds:

$$F(X_t|I_{t-1}) = F(X_t|I_{t-1} - Y^{Ly}_{t-Ly})$$

where $F(X_t|I_{t-1})$ is the conditional probability distribution of X_t given I_{t-1} which is the bivariate information set consisting of an Lx-length lagged vector of X_t, say $X^{Lx}_{t-Lx} \equiv (X_{t-Lx}, X_{t-Lx+1}, ..., X_{t-1})$, and $Y^{Ly}_{t-Ly} \equiv (Y_{t-Lx}, Y_{t-Lx+1}, ..., Y_{t-1})$. If the equation does not hold, then lagged past values of Y help to predict current and future values of X. Therefore Y is said to Granger cause X.

In order to try this Granger Causality Test, we will try to find as in [1] the "p-value", where "p-values" under 0.1 indicates a Granger cause between two time series. To obtain that "p-value" we will use the statsmodels[5] library from Python which has a function in which you specify the two time series and the number of lagged steps you want to test.

[5]http://statsmodels.sourceforge.net/devel/index.html

We also followed another methodology used in [2] to try to predict the stock market. The model used here tries to find if the use of sentiment or emotion analysis improve the results of prediction. It uses the following multiple linear regression models:

$$\hat{R}_t = f(R_{t-1}) \qquad \text{(M1, baseline)}$$

$$\hat{R}_t = f(R_{t-1}, ln(TIS_{t-1}) \qquad \text{(M2)}$$

$$\hat{R}_t = f(R_{t-1}, ln(RTIS_{t-1}) \qquad \text{(M3)}$$

$$\hat{R}_t = f(bind_{t-1}) \qquad \text{(M4)}$$

$$\hat{R}_t = f(ln(TIS_{t-1}) \qquad \text{(M5)}$$

where R_t is the return on day t, $bind_t$ is the Bullishness Index on day t, TIS_t is the Twitter Investor Sentiment on day t, and $RTIS_t$ is a ratio of TIS:

$$bind_t = ln\Big(\frac{1 + Positive_t}{1 + Negative_t}\Big)$$

$$TIS_t = \frac{Positive_t + 1}{Positive_t + Negative_t + 1}$$

$$RTIS_t = \frac{TIS_t}{TIS_{t-1}}$$

Because of the rapid changes of the characteristics of the stock market due to its high pace, we made the prediction of one day with the latest values from the previous 20 days, so we assured that the predictions are made with the latest characteristics of the market. Therefore, for every prediction, the values used in the regressions are updated by taking just the values from the last 20 days.

We will try to find if the models from M2 to M5 (which contain sentiment analysis) can improve the results from the baseline. In order to find if those models improve the baseline and the accuracy of the prediction, we use two different statistical error tests, the Root-Mean-Squared Error (RMSE) and the Mean Absolute Percentage Error (MAPE). These two tests have the following equations:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \big(y_i - \hat{y}_i\big)}{N}}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_1}{y_i} \right| 100\%$$

where N is the number of values of the time series that we are testing, y_i is the real value of the time series at time i, and $\hat{y}_i$ is the predicted value of the time series at time i. The lower the RMSE and MAPE values, the better the predictions, and although both

of them compute the mean error, RMSE is more sensitive to high individual errors than MAPE.

CHAPTER 4

Architecture

The architecture of this project is composed by the following modules:

- **Recollection:** used to obtain the tweets needed in the project. To achieve this task, we used two different sources, Paradigma Tecnológico and Infochimps.
- **Analysis:** to analyse of both the sentiment and the emotion of all the tweets.
- **Aggregation:** we have to aggregate the sentiment and emotion values of the tweets and store them in MongoDB.
- **Experimentation:** used to obtain the correlation between the stock market and the sentiment and emotion as well as to obtain the predictions. This module will be fully described in the following chapter.
- **Visualization:** we have to show all the data obtained in the project so we use the Sefarad application creating new widgets developed specifically to show our results.

We can see in Figure 4.1 a broad view of the architecture.

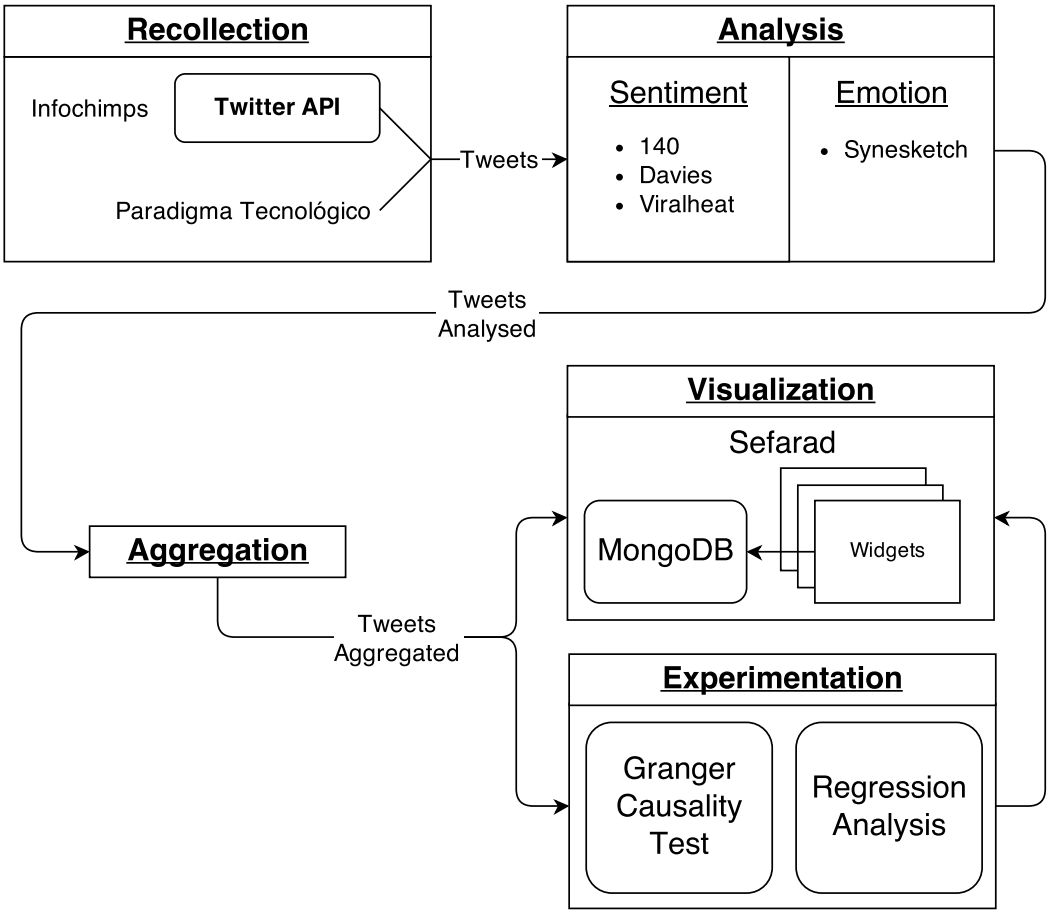


Figure 4.1: Architecture

4.1 Tweet recollection

To start our work we have to recollect the tweets. They were obtained from two different sources, Paradigma Tecnológico and a financial dataset from Infochimps[1]. Paradigma Tecnológico provided us with tweets that talked about the following companies: Telefónica, Iberdrola, Santander and Vodafone. These tweets were extracted from day 12/12/2013 to day 13/03/2014, corresponding to a 91 days period and the sentiment value was already analysed. The tweets obtained from the Infochimps dataset did not have the text of the tweets, it just had the date and time of them, the ID, and the entity of which the tweet was talking about. Because we were interested in the text of the tweets so we could analyse them and obtain the sentiment and emotion values, we had to use the Twitter REST API[2] to retrieve those tweets. First, we created a Twitter application to get access to this

[1] http://www.infochimps.com/datasets/twitter-census-stock-tweets
[2] https://dev.twitter.com/docs/api/1.1

API, specifically to the call *GET statuses/show/:id*, which was the necessary call to retrieve tweet by tweet from its ID. This call returned a JSON object with all the information of the tweet (text, user, number of retweets, location, language...). Then, we created a little script program in Python to make automatic the process, considering the limit of calls to the API per day. The period of time obtained with those tweets goes from day 01/05/2009 to day 31/03/2010, corresponding to a 334 days period. The companies studied from this dataset were Apple, Google and Amazon, although we could have taken more, we selected those three companies because of the high number of tweets per day due to their popularity.

Once we had all the text of the tweets and the different characteristic of them, we store them all in MongoDB. The database contained one collection per company studied and all the documents contained in these collections have the same fields such as tweet (text of the tweet), date, author, ticker of the company and the identifier Twitter gives to its tweets.

4.2 Tweet analysis

At this point we could begin with the analysis of the tweets. We had to analyse the sentiment of the tweets obtained from the Infochimps dataset and the emotion of the tweets obtained both from the Infochimps dataset and "Paradigma Tecnológico".

Sentiment analysis was carried out by three different services so we could try the prediction with all of them and discover which was the best. Those services were Alex Davies[3], Sentiment140[4] and Viralheat[5]. Alex Davies provided a sentiment analysis word list and a script written in Python as a example of how his method worked. We used this example by adapting it to take the tweets from MongoDB and to store the sentiment value in it again. The values obtained by this method were "1" for positive tweets and "0" for negative ones. Sentiment140 provided a REST API were you could send them a CSV file with all the tweets in it and the service returned the file with the tweets analysed. The values returned were "4" for positive tweets and "0" for negative ones. Finally, we used the Viralheat platform to analyse the tweets. It provided a REST API where you could send a tweet in a specific query and ViralHeat would return it with the sentiment. Because this service allowed us to obtain the sentiment tweet per tweet, we created another Python script to make automatic calls to it, we had to consider both the limit of number of call per day and the limit of making to many calls in a short time period. The values obtained were "positive" or "negative" and the probability that the value was correct. As we said

[3] http://alexdavies.net/twitter-sentiment-analysis/

[4] http://www.sentiment140.com/

[5] https://www.viralheat.com/

before, the sentiment of Paradigma Tecnológico was carried by them, so we did not have to analyse those tweets, the numeric values correspond to a negative sentiment if the value was below zero, and positive sentiment if the value was over zero.

Because the different services returned different values for same sentiment, after storing the sentiment in the database we converted those values into fixed values for all services. These fixed values were between 0.5 and 1 for positive tweets and between 0 and 0.5 for negative ones.

Emotion analysis was carried out by the synesketch[6] framework, a free open source software to analyse the emotion of any text. This tool analyses the emotion in 6 categories (anger, disgust, fear, happiness, sadness and surprise), giving a single value for every category corresponding to the probability that the emotion could be found in the tweet.

The database at this point was structured as it can be seen in Figure 4.2, as we said, there is one collection for every company analysed, and at this point, the documents contained in the different collections had the same fields as the ones shown in the "AAPL" collection. We added three fields of the sentiment analysis and another one with six parameters corresponding to the emotions analysed.

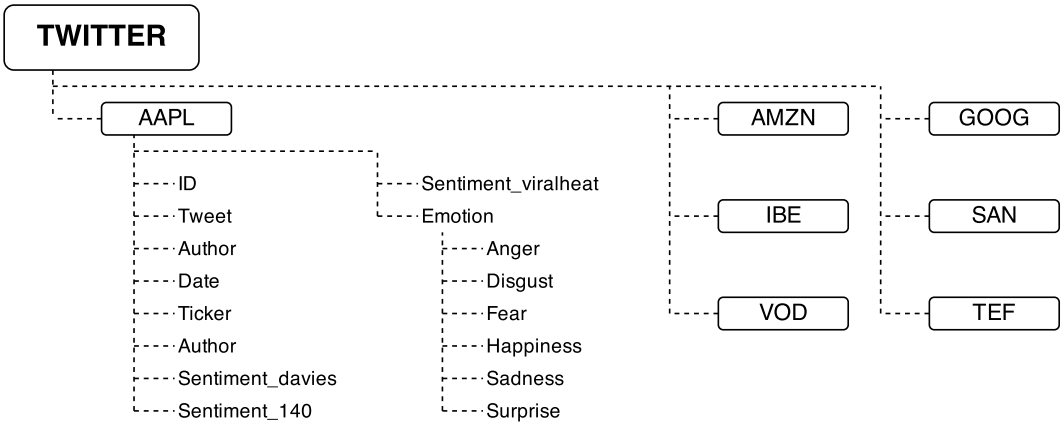


Figure 4.2: Database's structure after recollection and analysis

We can see the number of tweets per day in Figures 4.3 and 4.4. We can appreciate a few days where there are some peaks in the number of tweets corresponding to important events of the companies. For example, we can see a peak on the number of tweets of Apple on the 27th of January of 2010, this date corresponds to the first iPad launch. If we look for Amazon, we can see a high point on the 22nd of October of 2009 corresponding to a results announcement. Also, we can see a peak on the number of tweets of Telefónica on the day

[6] http://synesketch.krcadinac.com/blog/

Figure 4.3: Number of tweets (Paradigma Tecnológico)

5^{th} of March of 2014 corresponding to the launching of the new Service "Movistar Fusión".

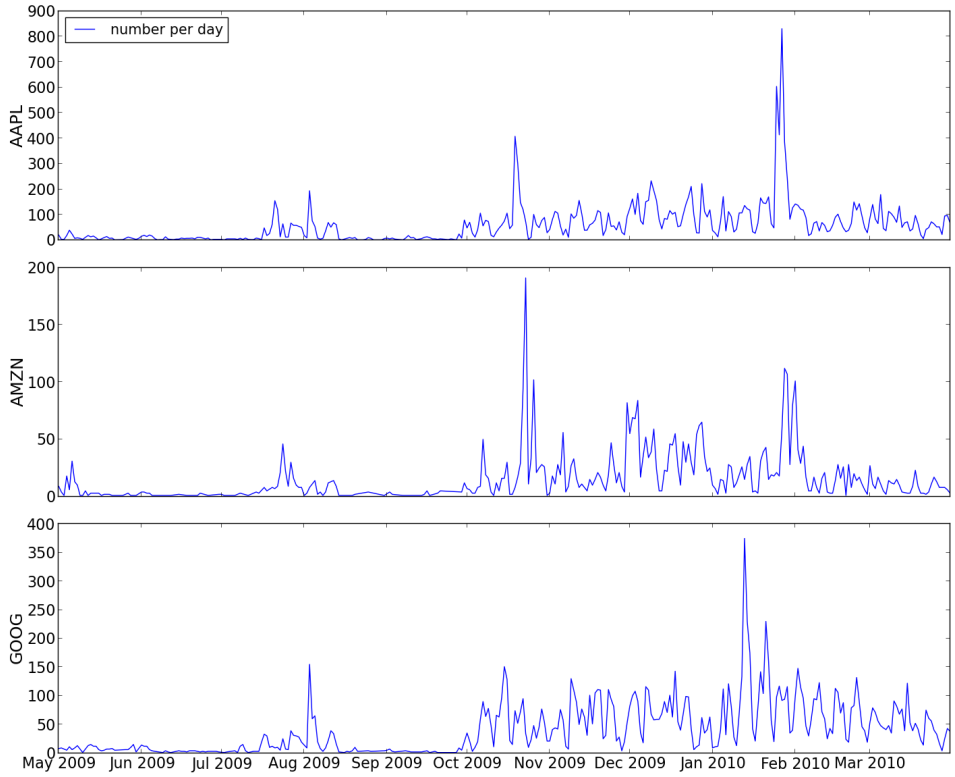


Figure 4.4: Number of tweets (Infochimps)

4.3 Tweet Aggregation

In order to be able to experiment with the data, we had to aggregate all the tweets in different values or parameters depending on the day the tweet was posted. So for every single day from the dates stated before, we had different values for different parameters that we will describe later. We should take in consideration that we only took days when the stock exchanges were opened, so the weekends and the holidays were not taken into account in the aggregation.

On the one hand we had the sentiment aggregation, although we had three different services, we made the same aggregation for all of them. We took all the tweets from a single day and we saw the sentiment of them, then we look for the positive ones and for the negative ones and we made the sum of each, obtaining the number of positive and negative number of tweets per day. Once we had those sums, we calculated the indicators explained on Chapter 3: Bullishness Index (bind), Twitter Investor Sentiment (TIS) and a ratio of TIS (RTIS).

On the other hand we had the emotion aggregation. It was made also by taking all the tweets from a day, looking for the values of the different emotions and summing all of them. Therefore, we obtained the number of angry tweets (and the rest of emotions analysed) per day.

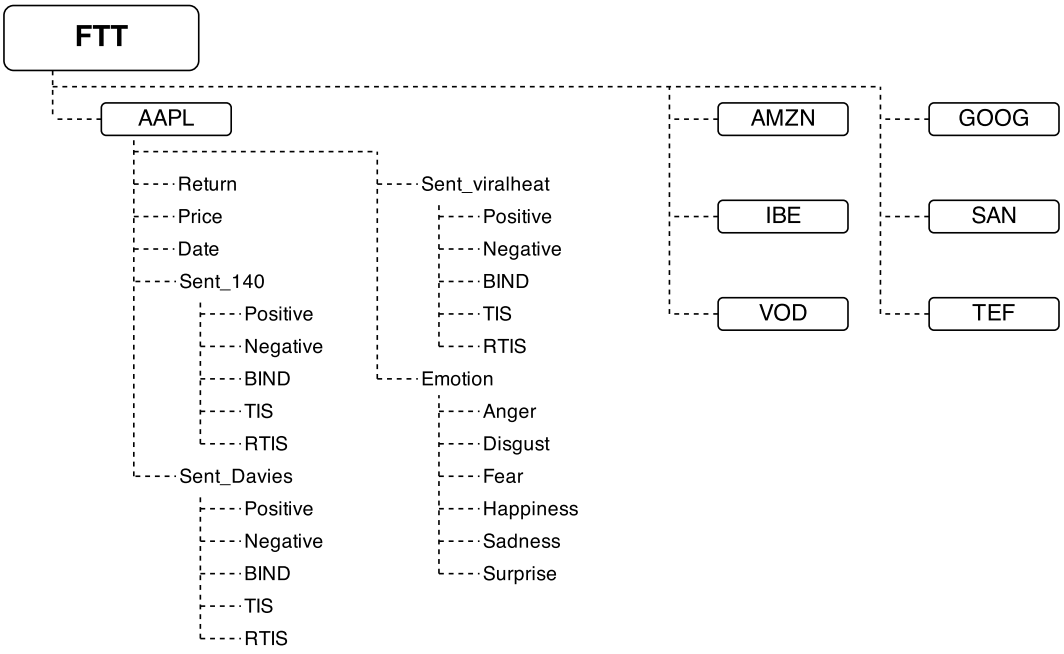


Figure 4.5: Database after aggregation

After the aggregation process, we stored all the values calculated in a new MongoDB database called "FTT". We can see in Figure 4.5 the structure of this new database, as in the "Twitter" database showed earlier, we have here one collection per company, and in all of them we have several documents (one per day of study) with the same fields.

In Figures 4.6, 4.7, 4.8 and 4.9 we can see the positive and negative distribution of the tweets over time after the aggregation process. There is one figure per service used in the analysis module and we can see several differences between them. We can see that the Davies tool does not give almost any negative tweet, the Sentiment140 tool marks more positive than negative tweets and the Viralheat service gives more or less the same number of positive tweets than negative ones. Furthermore, the tweets analysed by Paradigma Tecnológico has more negative than positive tweets or positive than negative ones depending on the company analysed. All these considerations should be kept in mind in the experimentation part in order to achieve the best possible results.

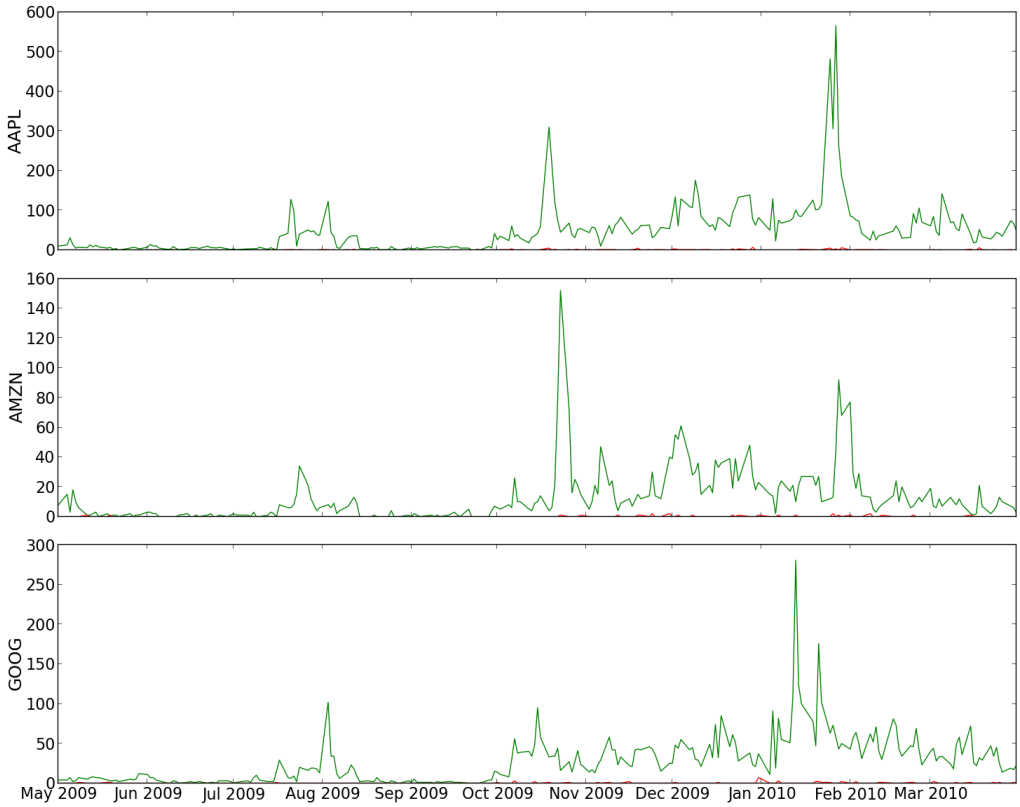


Figure 4.6: Sentiment Davies distribution

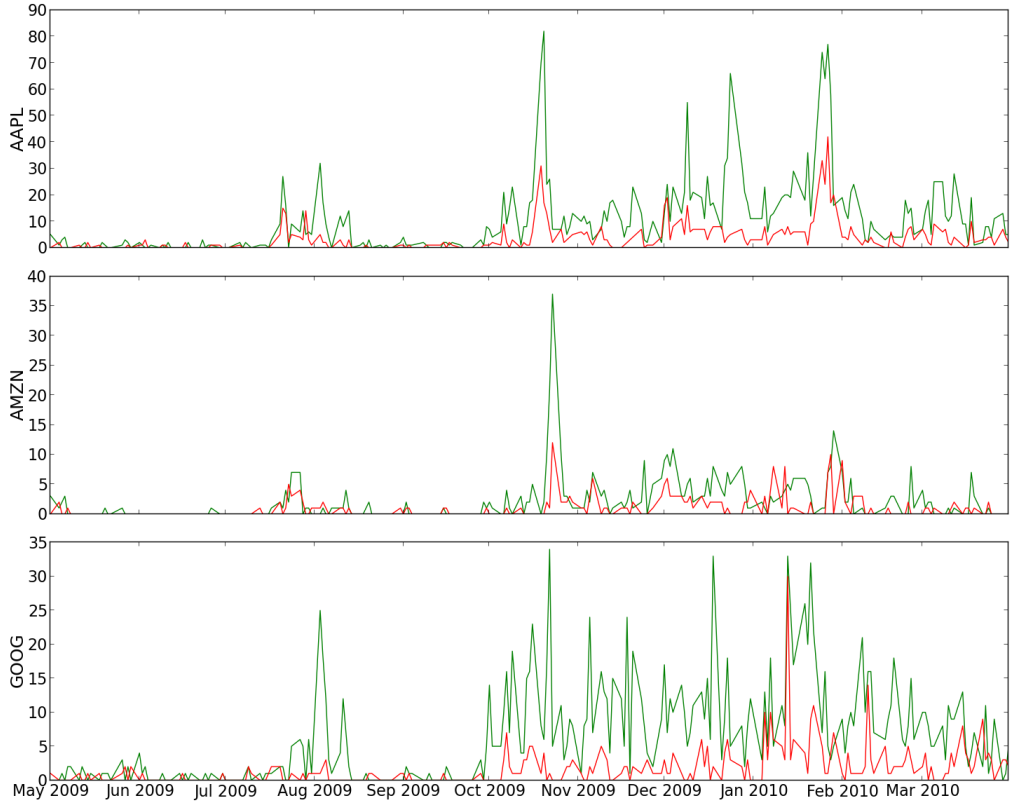


Figure 4.7: Sentiment 140 distribution

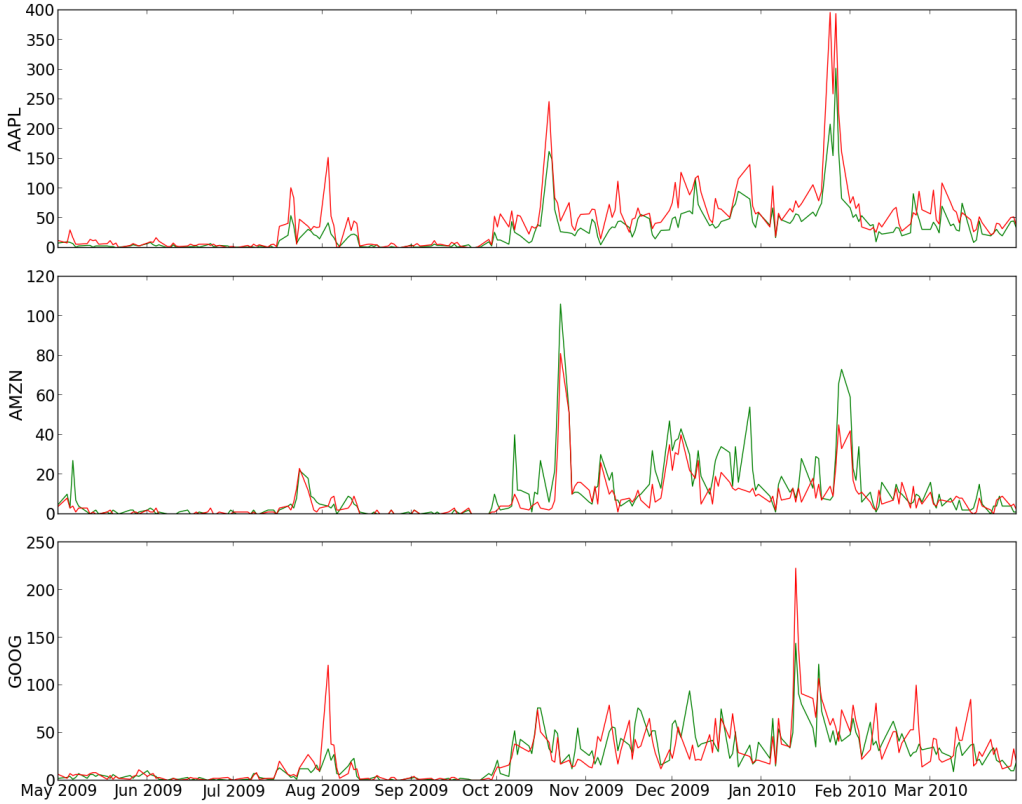


Figure 4.8: Viralheat distribution

Figure 4.9: Paradigma Tecnológico distribution

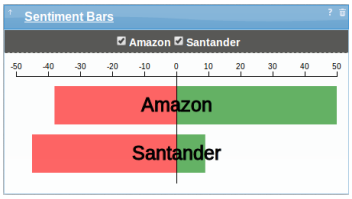
4.4 Visualization

To visualize the results we used the Sefarad application. We created our own widgets in order to show the data as best as possible. We maintained all the data in MongoDB, and we used php to retrieve all those data from the database to the application. Those data was sent in JSON objects so we could easily manipulate them with Javascript. Here are the different widgets created specially for this project:

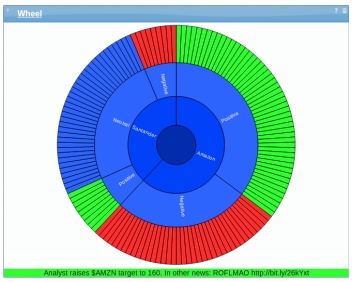
- **Results:** this widget shows a fixed number of tweets with the brand logo of the company and the sentiment of the tweet. Although this widget was already created in the Sefarad application, we customized it by changing the background colour of each result. So in our case, the background was set as green if the tweet had positive sentiment or red if the tweet had negative sentiment. The widget has a navigator bar to allow users to change the results that it is displaying.

- **Entity selector:** this tweet was already created for the Sefarad application and it allows the user to select the company or companies that wants to visualize. So for example it can choose only companies from the financial sector or from the technological sector. With this widget the user will select the companies that will be shown in the rest of the widgets since it works as a filter, so for example if the user selects Telefónica and Vodafone, all the information shown in the rest of the widgets will be with just the data from Telefónica and Vodafone.

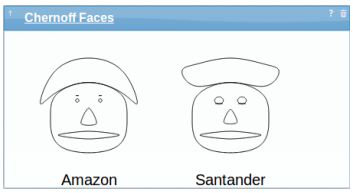
- **Sentiment Bars:** this widget shows in a bar graph the number of positive and negative tweets that the selected companies in the entity selector have. There is one bar per company selected and there are two sides, the right side constituting the positive tweets (coloured in green) and the left side constituting the negative tweets (coloured in red). The higher the number of positive or negative tweets is, the bigger the right or left bar will be.



- **Wheel:** this widget presents the tweets of the selected companies in the entity selector in a wheel graphic with different levels inside it. The tweets are grouped in the first level (corresponding to the inner place of the wheel) by the company that they are talking about. In the next level, those groups are split in two different new groups, which are the positive group and the negative group. In the outer place of the wheel, the positive and negative groups are again split in a number of pieces corresponding to the tweets that are in each group. If you move the mouse over one of these pieces it will appear at the bottom of the widget the text of the tweet. Finally, to improve the visualization, the background in the positive groups is green and in the negative groups is red.



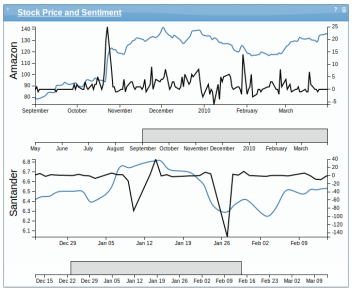
- **Chernoff faces:** this widget presents all the information available of the companies selected in a single face. The different facial characteristics of these faces depend on different measurements of the company. These facial characteristics and the measures related to them are: facial line (brand strength), eye size (positive sentiment), pupil size (negative sentiment), mouth (passion), eye brows (reach), nose (unique authors) and hair density (relative frequency results) [15]. The implementation of the Chernoff faces comes from an extension[7] of the D3.js framework which allows to store all these characteristics in a Javascript object and the extension will automatically create the face with all the shapes correctly painted.



  The characteristics stored in the Javascript object were calculated for every day as it follows: the positive/negative sentiment were calculated as the number of positive/negative tweets over the total of tweets per day; the unique authors was calculated as the number of different authors over the total of tweets that we had each day; and finally, the relative frequency results was calculated as the number of tweets per day over the average number of tweets per day of the period studied. The other values will be fixed at a medium value for every day. As other widgets, it will be presented to the user one face per company selected in the entity selector.

[7] http://bl.ocks.org/larskotthoff/2011590

- **Stock and sentiment widget:** this widget presents in a line graphic both the stock price and the sentiment value over time. By default, it will show all the time period available in the database, but you can adjust this period with a selector at the bottom of the widget which allows you to select the beginning and end date. It will appear one graphic per company selected and in all of them there will be two scales, the left scale will be the scale corresponding to the stock price and the right scale which corresponds to the sentiment scale.



CHAPTER 5

Experimentation

In this chapter we will describe the different experiments created to try to predict the stock market, basing the prediction on the sentiment and emotion analysis of the tweets. First, we will describe the Granger Causality Test and its results, and later we will try to find a prediction with a model based on linear regressions. Finally, we will try to find a correlation between the volatility of capital markets and Twitter emotion and sentiment. We will do that by executing the same tests explained before to the VXN index, also called fear index, which measures the current volatility in the market in a given moment.

5.1 Granger Causality Test

As we have stated in earlier chapters, the Granger Causality Test is a statistical test that tries to discover if a time series does not strictly cause another time series. This test is used to try to find a relation between past values of a time series and current values of another. For our experiment we will look if there is a relation between past sentiment and emotion values and current stock prices. We carried this experiment through a Python script, the first thing to do was to extract the sentiment and emotion of all the tweets from MongoDB, and also the close prices of every day studied.

The statistical test was carried by the statsmodels library from Python. That library has an implementation of the Granger Causality Test in a method, so we only had to call the method with a matrix of two columns, the first corresponding to the stock prices (the time series that is supposed to be caused) and the second corresponding to the sentiment or emotion values (the time series that is supposed to cause). The method should be given a parameter with the number of lags, in our case we used 4 days of lag as we though that number of days were enough for the stock to incorporate new information in its price.

The number in which we were interested on was the "p-value", which the lower the value the more related the time series. The method from statsmodels return 4 different implementations of the Granger Causality Test, so we chose the "ssr based F test" (all of them gave similar results).

For our experiments we differentiated between the sentiment and emotion. Beginning with the sentiment, as we have said before, we had three different services to analyse the sentiment of the tweets from Infochimps, but here we are going to expose only the one that gave the best results in the experiments, which was Viralheat. For our experiments we considered in a separate test both the positive and the negative sentiment series.

Starting with the positive sentiment, the results obtained from the Granger Causality Test are shown in Table 5.1. Moreover, the results obtained with the negative sentiment can be seen in Table 5.2.

	1-lag	2-lag	3-lag	4-lag
AAPL	0.6038	0.8226	0.7635	0.6576
GOOG	0.2238	0.4825	0.1707	0.1478
AMZN	**0.0219****	**0.0065****	**0.0313****	**0.0567***
IBE	0.8934	0.5790	0.5382	0.3494
SAN	0.5823	0.3791	0.4632	0.5003
VOD	**0.0823***	0.1693	0.1097	0.2119
TEF	0.1353	0.3009	0.1572	0.2259

Table 5.1: Positive sentiment results

We were looking here if there was any company with a "p-value" less than 0.1 (one

	1-lag	2-lag	3-lag	4-lag
AAPL	0.6384	0.7833	0.5474	0.3404
GOOG	0.3768	0.5942	0.6456	0.5174
AMZN	**0.0232****	**0.0018****	**0.0049****	**0.0045****
IBE	0.9711	0.8282	0.2427	0.3941
SAN	0.2380	0.4841	0.5149	0.6610
VOD	0.1161	0.2308	0.3380	0.6957
TEF	0.1425	0.1263	0.0866	0.1153

Table 5.2: Negative sentiment results

asterisk in the tables), which is the minimum value taken in [1] to accept a correlation between sentiment and stock market. A "p-value" less than 0.05 (two asterisks in the tables) indicates a stronger correlation than the 0.1 value. We can see from Tables 5.1 and 5.2 that Amazon obtains in the majority of the cases a "p-value" of less than 0.05, which would indicate a strong correlation between the sentiment and its stock price, but since the "p-value" does not descend with the number of lags we cannot consider Amazon as a representative case of correlation. The cause of this low values could be the few number of tweets per day from Amazon that we had in the database. Furthermore, we can see that only Vodafone has a "p-value" (in the positive tweets) below the 0.1 limit and this value increases over time. Therefore, we could say that there is a correlation between positive sentiment of Vodafone and its stock price, and that sentiment lagged one day cause the variations on the price. We could also see that the rest of the companies do not achieve "p-values" below 0.1, so there is no correlation for that companies.

Continuing with the emotion, we followed the same experiment as before, we used the Granger Causality Test with the different emotion categories and the correspondent stock price for every company. The results can be seen in the Tables 5.3 to 5.8.

As it could be seen, we had almost the same results for the emotion, Amazon gave us low values in all the categories except one (fear), and the emotions of Vodafone presented a strong correlation with the stock market. We had other correlations here as the disgust emotion and Apple, which had a "p-value" of 0.0583 for the 4th lag, and also the happiness emotion and Telefónica with a "p-value" of 0.0848.

	1-lag	2-lag	3-lag	4-lag
AAPL	0.1209	0.2737	0.3896	0.2449
GOOG	0.4437	0.6958	0.7214	0.7188
AMZN	**0.0537***	**0.0235****	**0.0497****	**0.0945***
IBE	0.9472	0.5643	0.4465	0.3685
SAN	0.9987	0.9588	0.3139	0.1344
VOD	**0.0644***	0.2029	0.4041	0.6962
TEF	0.4920	0.3390	0.5278	0.6132

Table 5.3: Anger results

	1-lag	2-lag	3-lag	4-lag
AAPL	0.3289	0.3643	0.1936	**0.0583***
GOOG	0.2587	0.6410	0.6878	0.3513
AMZN	**0.0555***	0.1797	0.2882	0.3904
IBE	0.5279	0.7929	0.5015	0.6883
SAN	0.8395	0.5341	0.4759	0.4942
VOD	**0.0938***	0.2864	0.4314	0.6990
TEF	0.4410	0.2013	0.5970	0.7326

Table 5.4: Disgust results

	1-lag	2-lag	3-lag	4-lag
AAPL	0.6027	0.8561	0.5512	0.2004
GOOG	0.8936	0.3888	0.4365	0.1445
AMZN	0.4740	0.6985	0.6977	0.8037
IBE	0.6317	0.7763	0.8380	0.8767
SAN	0.5786	0.8154	0.4216	0.4624
VOD	**0.0704***	0.1980	0.2372	0.5289
TEF	0.9616	0.3680	0.6284	0.5852

Table 5.5: Fear results

	1-lag	2-lag	3-lag	4-lag
AAPL	0.7174	0.1185	0.2409	0.1263
GOOG	0.6116	0.4059	0.6014	0.5864
AMZN	**0.0010****	**0.0000****	**0.0003****	**0.0004****
IBE	0.9886	0.9954	0.3294	0.2475
SAN	0.3746	0.6750	0.8396	0.6861
VOD	**0.0331****	0.1147	0.2265	0.4414
TEF	**0.0848***	0.1936	0.2619	0.3032

Table 5.6: Happiness results

	1-lag	2-lag	3-lag	4-lag
AAPL	0.5543	0.5354	0.5750	0.5579
GOOG	0.3635	0.2921	0.4223	0.3526
AMZN	**0.0307****	**0.0178****	**0.0593***	0.1028
IBE	0.7772	0.9828	0.7380	0.8118
SAN	0.5590	0.8648	0.6072	0.7977
VOD	**0.0513***	0.1397	0.1487	0.3644
TEF	0.5006	0.4159	0.3126	0.5426

Table 5.7: Sadness results

	1-lag	2-lag	3-lag	4-lag
AAPL	0.9881	0.2140	0.3178	0.2803
GOOG	0.5054	0.6350	0.4346	0.2557
AMZN	**0.0012****	**0.0010****	**0.0055****	**0.0133****
IBE	0.6647	0.9222	0.5421	0.3048
SAN	0.0182	0.0679	0.1243	0.1256
VOD	0.1707	0.4538	0.5346	0.7833
TEF	**0.0008****	**0.0079****	**0.0176****	**0.0299****

Table 5.8: Surprise results

5.2 Prediction based on linear regression

After the Granger Causality Test we tried to obtain a prediction of the different prices of the stocks. We used the regressions explained on Chapter 3 to obtain that prediction and the two error tests to measure its accuracy. Following the methodology of [2], we looked first for the accuracy of the baseline (a regression without sentiment) and then we tried with several regressions which included different forms of sentiment. Those forms were the ones explained also in Chapter 3 (TIS, bind, RTIS). With the results of all the regressions we looked if one of the last regressions improves the results of the error tests concluding then, that the sentiment helped to improve the prediction of the stock market prices. We made also the same experiment with the emotion values, using other regressions but keeping the basis of looking for the accuracy of a regression without emotion and then with it. Starting with the sentiment, we can see on Table 5.9 the results of the RMSE test and on Table 5.10 the results of the MAPE test. In both of them we can see in bold the values that are below the baseline.

	M1	M2	M3	M4	M5
AAPL	0.0182	0.0186	0.0188	**0.0180**	**0.0178**
GOOG	0.0137	0.0138	0.0142	**0.0132**	**0.0130**
AMZN	0.0309	0.0311	0.0314	0.0309	**0.0308**
IBE	0.0090	0.0092	0.0091	0.0090	**0.0089**
SAN	0.0136	0.0140	0.0143	0.0137	0.0137
VOD	0.0203	0.0216	0.0217	0.0206	0.0208
TEF	0.0125	0.0125	0.0124	0.0127	0.0125

Table 5.9: RMSE Results (Sentiment)

All in all, we could see that we did not obtain an improvement from the baseline. But for some specific companies and regressions it exists that improvement. For example, Google, which could be considered the company that obtains the better results, has a decline of 3.65% in the RMSE result for the M4 regression and a decline of 40.45% in the MAPE results for the same regression. With those results we created a graphic (Figure 5.1) with the real price of the stock (in green), the prediction of the price with the M4 model (in red)

	M1	M2	M3	M4	M5
AAPL	158.14	179.37	171.26	175.99	178.52
GOOG	637.39	**569.25**	**635.43**	**379.55**	**415.19**
AMZN	237.99	258.24	284.23	291.64	277.73
IBE	3001.67	4946.96	**2756.61**	**1864.52**	3022.20
SAN	2092.62	5864.80	4577.58	4812.48	5586.69
VOD	589.67	2132.76	921.30	2500.52	2304.62
TEF	1528.79	1451.05	2683.19	**1190.63**	**1481.84**

Table 5.10: MAPE Results (Sentiment)

and with the prediction of the price with the M1 model (in blue). We selected those three variables in order to compare how better the M4 is over the M1 model. Although the M4 model is not always really near to the real price we can see that it actually improves the M1 model as in the majority of the days all the values from the M4 are nearest to the real price than the values from the M1 model. Of course there are days when that is the opposite, which could be explained by the intrinsic characteristics of the markets as for example any news at night could condition the whole trade of the next day.

Continuing with the emotion, here we have the two linear regressions that we used in the experiment. We used only one regression which contains emotion due to the fact that it does not exist the concepts of bind or TIS for emotion.

$$P_i = f(P_{i-1}) \qquad \text{(E1, baseline)}$$

$$P_i = f(P_{i-1}, S_{x,i-1}) \quad \text{(E2)}$$

Where P_i denotes the price of the stock on day i and $S_{x,i-1}$ denotes the value of emotion x (one among the six categories) on day $i-1$. As we did with the sentiment, we tried both regressions with emotion and without it, to prove if adding the emotion improves the prediction. We can see in Table 5.11 the results for the RMSE test, and in Table 5.12 the results for the MAPE test.

It could be seen that except for Vodafone and Google or Amazon in a lesser extent, the emotion does not improve the predictions from the baseline. Considering Vodafone, the

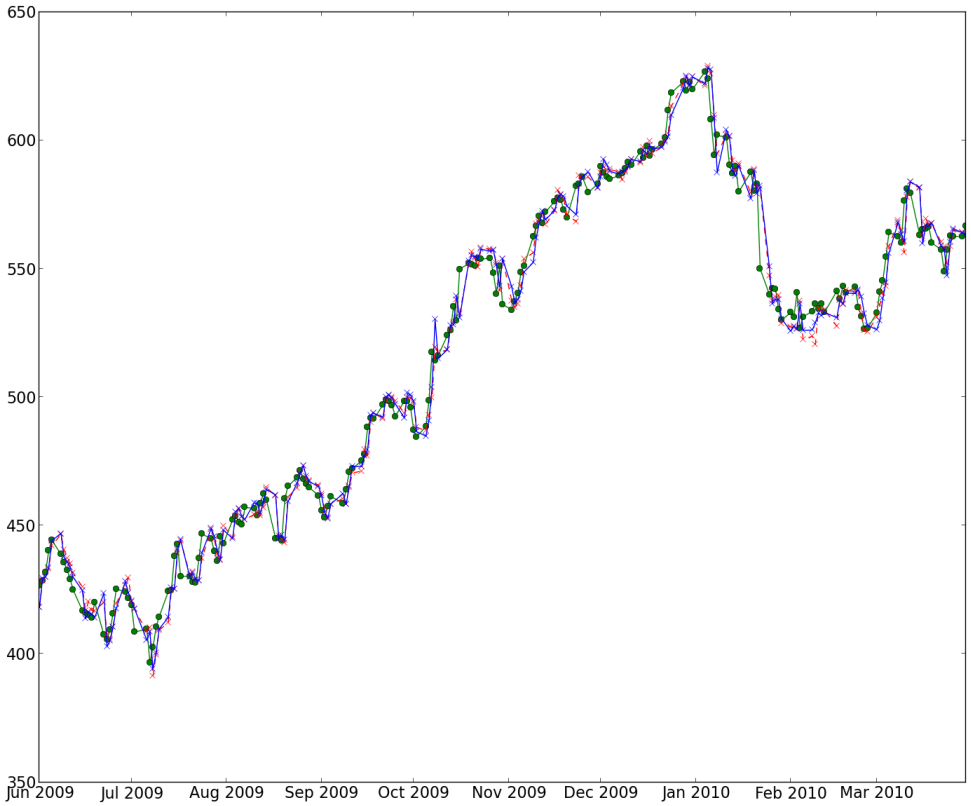


Figure 5.1: Google real price (green), prediction with (red) and without (blue) sentiment

anger emotion improves the RMSE result a 18.53% and the MAPE result a 11.97% so we used that emotion to predict the prices that are shown in Figure 5.2. We have here the real price of the stock (green), the predicted price using the anger emotion (red) and the predicted price using just the price of the previous day (blue). It can be seen that for the first part of the graphic both predictions are almost the same, but after 15/02/2014 there is a change in the behaviour of the predictions. That change lead the prediction with the anger emotion to be more accurate (being nearest to the real price) than the prediction without it.

	AAPL	GOOG	AMZN	IBE	SAN	VOD	TEF
E1 - Base	3.7066	7.5633	3.4171	0.0431	0.0837	0.8705	0.1478
E2 - Anger	3.8741	**3.5345**	7.9641	0.0436	0.0894	**0.7092**	**0.1452**
E2 - Disgust	3.9109	**4.8477**	7.8060	0.0534	0.0907	**0.8158**	0.1621
E2 - Fear	3.9327	**3.6171**	7.7921	0.0478	0.0905	**0.8124**	0.1616
E2 - Happiness	3.7580	**4.9816**	8.3134	0.0531	0.0877	**0.7590**	0.1514
E2 - Sadness	3.8131	**4.0154**	7.7697	0.0453	0.0865	**0.8036**	0.1511
E2 - Surprise	3.8373	**3.5235**	11.7688	0.0647	**0.0821**	**0.8119**	0.2140

Table 5.11: RMSE Results (Emotion)

	AAPL	GOOG	AMZN	IBE	SAN	VOD	TEF
E1 - Base	1.4004	1.0201	1.7327	0.7088	0.9639	1.5820	0.9232
E2 - Anger	1.4537	1.8917	**1.0812**	0.7124	1.0392	**1.3927**	0.9447
E2 - Disgust	1.4895	2.1775	**1.0633**	0.8173	1.0550	**1.5261**	1.0377
E2 - Fear	1.4788	1.8198	**1.0583**	0.7632	1.0307	**1.5460**	1.0348
E2 - Happiness	1.4221	2.1701	**1.1081**	0.7732	0.9884	**1.4406**	0.9721
E2 - Sadness	1.4230	2.0621	**1.0443**	0.7481	1.0410	**1.4977**	0.9481
E2 - Surprise	1.4521	1.9078	**1.2032**	0.8571	0.9655	**1.4780**	0.9722

Table 5.12: MAPE Results (Emotion)

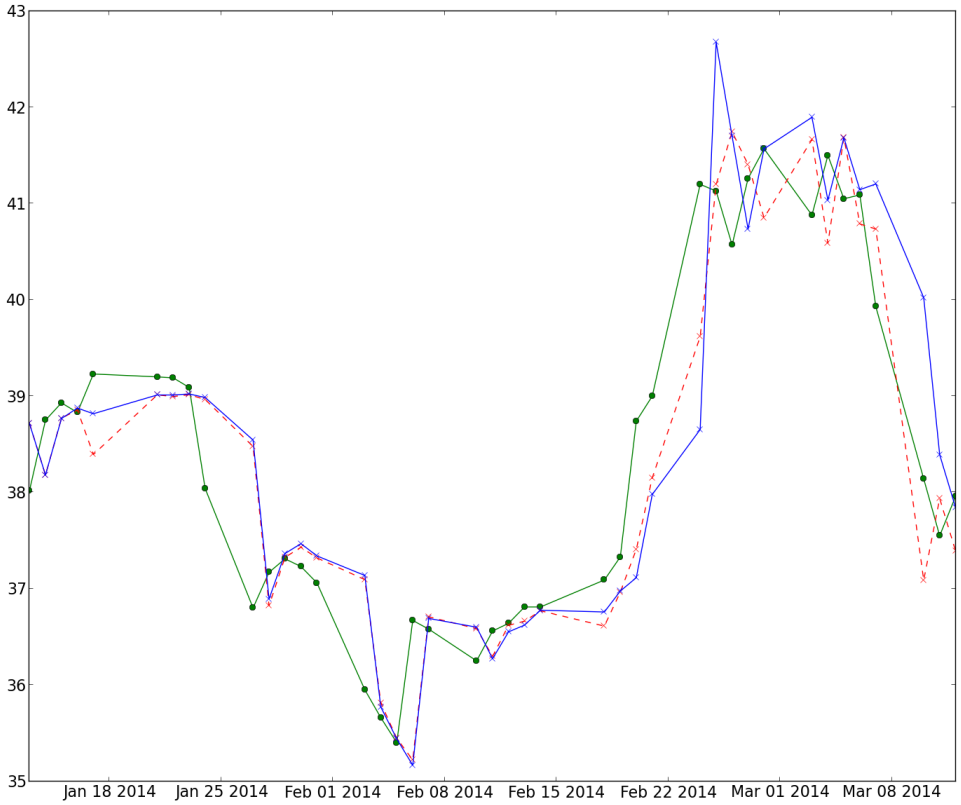


Figure 5.2: Vodafone real price (green), prediction with (red) and without (blue) emotion

5.3 VXN

The VXN is the volatility index for the Nasdaq 100 index. It is a financial index that measures the volatility of the Nasdaq market. It is more common the Volatility Index of the S&P market, or VIX, but we had to use VXN since we were studying Apple, Google or Amazon, which are traded on the Nasdaq market. The volatility is the statistical dispersion of the return of a given security, and it is related with the risk of that security: the higher the volatility, the higher the risk of the security. VIX, or in our case VXN, is called the "fear index" due to its relation with the risk, so for example the VXN climbed to its highest level with the default of Lehman Brothers or with the peripheral debt crisis in Europe.

Because of the importance of this indicator or index, we tried to predict it with the fear emotion, as they were supposed to be related due to its nature. In order to assess that, we used the Granger Causality Test first, and another proposed linear regression model afterwards. Beginning with the Granger Causality Test, we obtained the results shown in Table 5.13. We took a 7 days lag period in order to look if there was probable to find a correlation among more days than what we used with the stocks.

Lag-1	Lag-2	Lag-3	Lag-4	Lag-5	Lag-6	Lag-7
0.8111	0.8348	0.9504	0.4289	0.3136	0.4381	0.1990

Table 5.13: VXN Granger Test Results

We ended the experimentation trying to predict VXN index with the following linear regressions. If the fear emotion helps to predict VXN, the results of the RMSE and MAPE for the V2 regression are lower than for the V1 regression.

$P_i = f(P_{i-1})$ (V1, baseline)

$P_i = f(P_{i-1}, F_{i-1})$ (V2)

Where P_i is the value of the VXN on day i and F_{i-1} is the aggregation value for the fear emotion on day $i-1$. The aggregation value is calculated by adding all the values of every single day of the fear emotion for the companies Apple, Google and Amazon.

The results of the RMSE and MAPE test are shown on Table 5.14.

By looking at the results from the Granger Causality Test and the regression model, we could conclude that the fear emotion is not correlated with VXN index. We could observe that there is no lagged day with a "p-value" lower than 0.1 for the Granger Causality Test,

	V1	V2
RMSE	1.0903	1.1804
MAPE	2.8632	3.1137

Table 5.14: VXN RMSE and MAPE Test Results

so there is no correlation between them. In addition, the results from the regression model are equally bad, there is no improvement in the V2 regression in either of the RMSE or MAPE tests.

To sum up, we have seen the correlation between the stock market and the sentiment and the emotion from Twitter trough the Granger Causality Test. Although we did not obtain a strong correlation for the majority of the companies selected, we have seen that 1-day lagged values of the negative sentiment as well as emotion of Vodafone are highly correlated with current prices.

Looking at the predictions obtained with different values we cannot observe good predictions for the majority of the companies but for Google and Vodafone. We obtained a better result in the prediction of Google prices using the sentiment than in the prediction without it. Both of the error tests used returned a better result with a 40,45% of improvement for the MAPE test. In Vodafone, the improvement was a 18,53% for the RMSE result for the anger emotion. We can see in both Figures 5.1 and 5.2 the corresponding graph with the real prices and the predictions.

Finally, we studied the correlation and tried to predict the VXN without achieving good results as we did not find any correlation with past values of sentiment or angry emotion, and the predictions were not improved in any of the models.

CHAPTER 6

Conclusions and future work

6.1 Conclusions

We have seen how investors do not base their investments only in rational reasons as they are also influenced by sentiment and emotion factors. That is widely explained by Behavioural Finance [6], which is a refutation of the Efficient Market Hypothesis with the consequence of being possible of getting better returns than the market, and an explanation of how psychology intervenes in economy. That fact has allowed us to measure those factors in order to predict future changes in the market with current sentiments or emotions. We chose a social network as Twitter to collect those sentiments and emotions because we considered it the perfect tool for people to spread their ideas and opinions [16].

Once we had the tweets and their sentiment and emotion analysed, we began to look if there was a relation between them and the stock market by using two different statistical tests. Looking for the results of the first one, the Granger Causality Test, we could say that there was no such strong correlation between the two variables. Although for few companies that correlation existed, we cannot conclude this statement in the general behaviour. Thinking about the possible causes of those results, we thought that the main one was the number of tweets and the quality of them. So for example, as we have seen in the case

of Amazon, a low number of tweets leads to a low "p-value" for all the lags. In addition, the opposite (having a big number of tweets) could cause a no-correlation effect due to the possible noise inside some of the tweets, as it could happen in the case of Apple in which, although we had a lot of tweets, we did not see good results.

Furthermore, looking at the results of the linear regressions, we could be able to conclude that in general it does not exist an improvement in the prediction of the price or return when we add the sentiment or the emotion. Referring to sentiments, there is no regression proposed that improves the prediction for all the companies; although we had the M5 regression with an improvement in 4 out of 7 companies, it is so little that is not strong enough to conclude anything. It happens the same with emotion; there is no emotion that improves the results for all the companies, but the improvements are for all the emotions of a company.

To sum up, we could conclude that although it is difficult to find a correlation or a good prediction of the stock market, we can find some examples such as Vodafone or Google that give good results. In order to improve those results, the recollection of the tweets from Twitter should be perfected to obtain both a great number of tweets per day and a high quality in content.

6.2 Future work

Here we expose some recommendations that could be made in order to improve the features and characteristics of the work:

- The main problem of the work is the low number and quality of the tweets that are used, so in order to improve that, future works should take care of it and try to improve the recollection by being more specific with the tweets (and not just taking all the tweets that talk about a company) filtering them previously. That would probably improve the correlation with the stock market and the predictions obtained by the different regression models.
- One of the biggest problems of the project is that it is based on data from the past, which could lead to different results as if we used the same methods with current tweets, due to the fast changeable characteristics of the markets over time. Therefore, we propose to probe those methods with current tweets and data.
- The ideal use of this system would be using real time data with it. By building that system with the resources needed to collect, analyse and test tweets in real time, we

could achieve a powerful tool. The system would be able to deliver instantaneously any change of the sentiment or emotion of any company to investors or traders.

- The interface could be improved by creating new widgets in which you select the sector you wanted to be informed about, and the widget automatically updates the content with the companies from that sector.

- We propose to adapt the interface to new devices that allow those investors or traders to develop their work without needing to be looking to a web page. Thus, we suggest to develop an application for Google Glass which would be the ideal device for an application with this characteristics.

- Finally, we recommend building a new system based on the results obtained from this work, which studies if it is possible to obtain real profits in the real market considering the commissions and the different costs of investing.

Bibliography

[1] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[2] N. Oliveira, P. Cortez, and N. Areal, "On the predictability of stock market behavior using stocktwits sentiment and posting volume," in *Progress in Artificial Intelligence*, pp. 355–365, Springer, 2013.

[3] Y. Le Fur, M. Dallochio, and A. Salvi, *Corporate finance: theory and practice*. John Wiley & Sons, 2011.

[4] A. Mittal and A. Goel, "Stock prediction using twitter sentiment analysis," *Standford University, CS229(2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 2012.

[5] B. G. Malkiel and E. F. Fama, "Efficient capital markets: A review of theory and empirical work*," *The journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[6] N. Barberis and R. Thaler, "A survey of behavioral finance," *Handbook of the Economics of Finance*, vol. 1, pp. 1053–1128, 2003.

[7] A. Logunov, *A Tweet in Time: Can Twitter Sentiment analysis improve economic indicator estimation and predict market returns?* PhD thesis, The University of New South Wales Australia, 2011.

[8] J. F. Sánchez-Rada, M. Torres, C. A. Iglesias, R. Maestre, and R. Peinado, "A Linked Data Approach to Sentiment and Emotion Analysis of Twitter in the Financial Domain," 2014.

[9] P. D. Azar, *Sentiment analysis in financial news*. PhD thesis, Harvard University, 2009.

[10] C. Oh and O. Sheng, "Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement," 2011.

[11] T. T. Vu, S. Chang, Q. T. Ha, and N. Collier, "An experiment in integrating sentiment features for tech stock prediction in twitter," in *24th International Conference on Computational Linguistics*, p. 23, 2012.

[12] R. Bermejo, "Desarrollo de un framework html5 de visualización y consulta semántica de repositorios rdf," Master's thesis, E.T.S.I. Telecomunicación, Universidad Politécnica de Madrid, 2014.

[13] D. Smiley and D. E. Pugh, *Apache Solr 3 Enterprise Search Server*. Packt Publishing Ltd, 2011.

[14] C. Hiemstra and J. D. Jones, "Testing for linear and nonlinear granger causality in the stock price-volume relation," *The Journal of Finance*, vol. 49, no. 5, pp. 1639–1664, 1994.

[15] M. Farshid, K. Plangger, and D. Nel, "The social media faces of major global financial service brands," *Journal of Financial Services Marketing*, vol. 16, pp. 220–229, 2011.

[16] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining.," in *LREC*, 2010.