INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

# Analysis of Associations between Financial Time Series and Social Network Activity

## TESIS

QUE PARA OBTENER EL GRADO DE:

**Maestría en Ciencias de la Computación**

PRESENTA:
**Ing. Francisco Javier García López**

DIRECTORES DE TESIS:
**Dr. Ildar Batyrshin**
**Dr. Alexander Gelbukh**

Ciudad de México

Diciembre 2017

SIP-14 bis

# INSTITUTO POLITÉCNICO NACIONAL
## SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

### ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 12:00 horas del día 08 del mes de septiembre de 2017 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

***Centro de Investigación en Computación***

para examinar la tesis titulada:

**"Analysis of Associations between Financial Time Series and Social Network Activity"**

Presentada por el alumno:

| | | |
|---|---|---|
| **GARCÍA** | **LÓPEZ** | **FRANCISCO JAVIER** |
| Apellido paterno | Apellido materno | Nombre(s) |

Con registro:

| | | | | | | |
|---|---|---|---|---|---|---|
| B | 1 | 5 | 1 | 1 | 5 | 3 |

aspirante de: **MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***APROBAR LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

**LA COMISIÓN REVISORA**

Directores de Tesis

Dr. Ildar Batyrshin

Dr. Alexander Gelbukh

Dr. Sergio Suárez Guerra

Dr. Oleksiy Pogrebnyak

Dra. Olga Kolesnikova

Dr. Grigori Sidorov

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Marco Antonio Ramírez Salinas

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION
EN COMPUTACION
DIRECCION

***INSTITUTO POLITÉCNICO NACIONAL***

***SECRETARÍA DE INVESTIGACIÓN Y POSGRADO***

## *CARTA CESIÓN DE DERECHOS*

En la Ciudad de México el día **1ro** del mes **diciembre** del año **2017**, el que suscribe **Francisco Javier García López**, alumno del Programa de **Maestría en Ciencias de la Computación** con número de registro **B151153**, adscrito al **Centro de Investigación en Computación**, manifiesta que es autor intelectual del presente trabajo de Tesis bajo la dirección del Dr. Ildar Batyrshin y el Dr. Alexander Gelbukh y cede los derechos del trabajo intitulado **"Analysis of Associations between Financial Time Series and Social Network Activity"**, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección **f.javier.334@gmail.com**. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Javier García

Francisco Javier García López

Nombre y firma

# Resumen

Una aplicación interesante de la minería de datos de series de tiempo es en el área financiera. En esta área es importante entender la dinámica de los precios de las acciones ya sea para la creación de un portafolio que minimiza el riesgo para inversores o para entender la relación entre las compañías que las series de tiempo representan. Con la ubicuidad de la información estos días, es posible incorporar en el análisis eventos de las noticias y redes sociales para así entender su impacto en los precios de las acciones.

En este trabajo primero se estudia la interrelación de las series de tiempo, particularmente, la relación entre su comportamiento y la naturaleza de las compañías a las que pertenecen. Se proponen dos métodos, que a su vez se basan en una metodología que fue propuesta antes en la literatura. Se mejoraron los métodos anteriores al incluir un proceso para la selección automática de la ventana de tiempo, que no había sido considerado antes y una medida de similitud basada en la longitud de los patrones que se obtienen utilizando dicha metodología. Los métodos que se proponen tienen una interpretación natural en las respectivas industrias analizadas.

En segundo lugar se estudia la relación que tienen las asociaciones de las series de tiempo financieras con los eventos en las noticias de las compañías analizadas. Fue propuesto un método para visualizar dinámicamente estos cambios. Los eventos de las noticias fueron contrastados visualmente con las asociaciones de las series de tiempo y se puede apreciar que sí existe un impacto de éstos en dichas asociaciones.

Finalmente se analiza la relación entre algunos indicadores de los precios de las acciones y las redes sociales usando la plataforma Twitter. Estos experimentos se dividen en dos partes. Primero, las medidas usadas anteriormente y un análisis de correlación fueron usados para medir la relación entre la cantidad de tweets y los indicadores financieros. Después se analizó el texto de los mensajes en función de las tendencias del precio que tuvieron lugar cuando los mensajes fueron generados, esto se hace usando clasificadores del área de Aprendizaje Automático. Se encontró una relación entre los indicadores financieros y la cantidad de mensajes en las redes sociales y también se encontró que los clasificadores fueron capaces de distinguir la tendencia durante la cual se produjeron los mensajes.

# Abstract

An interesting application of time series data mining techniques is in the area of finance. In that area it is important to understand the dynamics of the stock prices for several reasons, it can be used for the creation of a portfolio that minimizes risk for investors and it can also be informative about the relationship of the companies that the time series represent. With the ubiquity of information it is possible to incorporate into this analysis events reported in news and social media, this is done with the aim of understanding the impact they have on the stock price.

In this work it is first studied the relationship between the time series behavior and the nature of the companies they represent. Two methods are developed based on a methodology previously proposed in the literature. The previous methods were improved by including a process for the automatic selection of the window size, which was not considered before, and a similarity measure based on patterns. The methods that are proposed have a natural interpretation in the respective industries analyzed.

Secondly, it is studied the relationship between changes in financial time series associations and news events of the analyzed companies. A method to dynamically display these changes was proposed. The news events were visually contrasted and they seem to have an impact on the time series associations.

Finally it is analyzed the relationship between some indicators of the stock prices and social media using the platform Twitter. These experiments are divided in two parts, first, the previously described measures and the correlation analysis were used to measure relationship between the amount of tweets and price indicators. In the second experiment, the text from the messages in the platform were analyzed in function of the price trends in which such messages were generated. It was found a relationship between some financial indicators and social media using both the time series and the textual approaches.

# Acknowledgments

> "We shall not cease from exploration, and the end of all our exploring will be to arrive where we started and know the place for the first time."
>
> — T. S. Eliot

I would like to thank my two advisors Dr. Ildar Batyrshin and Dr. Alexander Gelbukh for their continuous support, advise and direction.

I would also like to express my gratitude to my professors in the *Centro de Investigación en Computación* (CIC) of the *Instituto Politécnico Nacional* (IPN) for their lessons and for sharing their knowledge, thus making a valuable contribution to my education.

Thanks to the students in the center and in the Natural Language Processing laboratory, for the environment of respect and cordiality.

I recognize the personnel in the center for their attentive support, professionalism and for their daily efforts to make this a great center.

Many thanks to the *Consejo Nacional de Ciencia y Tecnología* (CONACYT) for their financial support and to the IPN also for their financial support in the last stages of the Master's process.

I also greatly appreciate the *Sociedad Mexicana de Inteligencia Artificial* (SMIA) for their support that allowed me to go to important conferences and the *Red Temática en Tecnologías del Lenguaje* (Red TTL) also for their support to go to conferences and for the workshops that they organized, where it was possible to meet new people that work in my area.

# Dedication

To my parents for their guidance,
to my siblings for their love and
to family for their support.

# Contents

# List of Tables

# List of Figures

# List of Acronyms

| | |
|---|---|
| **BOW** | Bag of Words |
| **DJIA** | Dow Jones Industrial Average |
| **DT** | Decision Tree |
| **EMH** | Efficient Market Hypothesis |
| **EST** | Eastern Standard Time |
| **FTSE** | Financial Times Stock Exchange |
| **MV** | Majority Vote |
| **NLP** | Natural Language Processing |
| **RF** | Random Forest |
| **S&P500** | Standard and Poor 500 Index |
| **SGDC** | Stochastic Gradient Descent Classifier |
| **SVC** | Support Vector Classifier |
| **UTC** | Coordinated Universal Time |
| **VC** | Voting Classifier |
| **WE** | Word Embeddings |

# Chapter 1

# Introduction

## 1.1 Background of the Study

With the growth of Internet users the amount of available data is enormous. The data that users generate can be used for those purposes and there are platforms such as Facebook or Twitter that allow users to share their thoughts about any topic. Many companies rely more and more on information for decision making. However, getting the right information from the data is not a trivial task.

For that reason, data mining is an area that has attracted attention in the last years. New methods have been developed with the aim of extracting relevant information from the available data. This study intended to explore new methods in time series data mining, the new methods are tested on financial time series, and to understand the relationship between the financial market and the flow of information that users generate in textual form.

In finance there are two predominant explanations of the stock market behavior, the Efficient Market Hypothesis [1] (EMH) and the Random Walk Hypothesis [2] (RWH), neither of these well known theories supports the predictability of the stock market. Random Walk theory states that the stock market behaves randomly. The principal argument are the myriad factors involved in the price calculation that is not humanly possible to consider them all.

Efficient Market Hypothesis states that all information is reflected immediately in the stock market prices and therefore it is not possible to use it for prediction. The theory can be understood at three different levels or variants depending upon which information is reflected in the price. The weak variant includes only financial indicators such as the expected earnings or the historical prices of a company, the semi-strong variant besides financial indicators includes public information, usually information on the news, and the strong variant includes all others and the private information, reports from the internal organizations, corporate insiders, private consulting firms, etc. The last variant is the most severe on stating how unpredictable the prices are.

Besides such theories there have been different efforts in stock market forecasting. The two main paradigms in stock price forecasting are technical analysis and fundamental analysis. The first approach considers historical time series, and methods such as ARIMA, state-space models and neural networks are used. The second approach relies on financial indicators at large scale (overall economy, unemployment , welfare, etc.) and from company itself (expected earnings, balance sheets, reported revenue). In real case scenarios both approaches must be considered. In recent years there has been a large effort in the inclusion of news in an automated way using the advances in the computational field of Natural Language Processing (NLP).

There are three target market levels that have been studied: global market, sector or industry and the level of individual companies. There are efforts at the three levels, using more or less similar techniques. In the current study the level of individual companies is addressed.

## 1.2 Statement of the Problem

For financial or any type of time series, point to point analysis, as used traditionally in correlation, may lead to misleading results. In the first part of the study new methodologies that consider time series trends are proposed to analyze relationships in the financial domain. Further, the trends that appear in the relationships between two companies are used as patterns, and such patterns are used to determine the time series similarity. These methods also consider the changing nature of the financial time series and how these changes are related to news events.

In the second part it is addressed the problem of understanding the relationship between social media and the stock market. Most works in the literature treat this as a forecasting[1] problem. In this part, the problem of including text to the financial analysis is explored. Two textual representations and different machine learning classifiers are tested. In the literature many

[1] Although the word forecast as a noun is used to refer to the estimation of future events especially for weather or finance, in this study the two words, predict and forecast, are used interchangeably.

approaches were found, however different methodologies have been used to measure the results which make the studies difficult to compare.

## 1.3 Significance of the Study

The significance is that the proposed methods are tested with real data from the financial field, also, that news events are integrated in the analysis and the relationship of these time series with social media is analyzed.

In the part of the study that deals with the relation between text features of social media and individual stocks the particular contributions are, first, the review the different approaches that have taken place in the analysis of the stock market using social media, second, the comparison of textual representations and machine learning methods that are used in the literature, and third, the relationship analysis with social media using traditional approaches, the methods proposed in this work and also machine learning approaches. It is important the understanding of how the two fields, the financial market and social network activity, relate to each other, this will provide empirical evidence in favor or against the EMH [2].

## 1.4 Objectives

### 1.4.1 General Objective

Test association measures on the stock market time series to verify their interrelationship, and the impact of the news and social media on these time series.

### 1.4.2 Specific Objectives

- Propose new similarity measures that consider the time series trends and that have a natural interpretation.

- Understand the relationship between the data on social media and the price of financial stocks of individual companies.

- Test experimentally what is the best language representation for those purposes.

- Understand what are the best machine learning methods for those purposes.
- Use social media to test predictability on the stock market.

## 1.5 Limitations, Delimitations and Assumptions

The social media data is obtained from Twitter and the financial data from Google Finance. It is assumed that the time stamp used by Twitter is accurate and that Google Finance provides accurate information as well. In the data obtained from Twitter fake news, bot messages or publicity is collected.

A few companies are chosen based on their popularity. It is assumed that tweets are from everywhere in the world and that the data collected from Twitter is representative of the all messages in that social network. The methods can be extended to other companies or industries. Because of the nature of the financial data collected the price changes are analyzed daily.

Finally, the time span for the tweet collection do not include periods of major world financial crises.

## 1.6 Organization of the Study

The dissertation has the following structure: In Chapter 2 it is presented the theoretical framework and the methodology used through the experiments. In Chapter 3 it is presented a method to visualize the relationship between financial time series and also a method for clustering them based on length of association patterns. In Chapter 4 a method to visualize the dynamic associations of financial time series and the relationship that these associations may have with events. In Chapter 5 the relationship between financial time series and social media is analyzed. For this purpose the methods from previous chapters are used. Finally in Chapter 6 the conclusions and future work are presented.

# Chapter 2

# Methodology

## 2.1 Introduction

In this section it is presented the theoretical framework used on Chapters 3 to 5. The Moving Approximation Transform (MAT) will be used through Chapters 3 to 5. For Chapter 5 two different textual representations, the machine learning classifiers and statistical tests are discussed.

## 2.2 The Moving Approximation Transform (MAT)

A time series of length $n$, where $n$ is a positive integer, is a sequence of real numbers $x = (x_1, x_2, \ldots, x_n)$ that correspond to points in time $t = (1, 2, \ldots, n)$. A time series can be denoted simply as $x$. A time window $W_i$ of length $k > 1$ is a sequence of indexes $W = (i, i+1, \ldots, i+k-1)$, $i \in 1, \ldots, n-k+1$. The sequence $x_{W_i} = (x_i, x_{i+1}, \ldots, x_{i+k-1})$ is defined as a sequence of values of the time window corresponding to indexes of $x$. A sequence $J = (W_1, W_2, \ldots, W_{n-k+1})$ of all possible windows of size $k$ for $1 < k \leq n$ is called a sliding window.

A function $f_i = a_i t + b_i$ with parameters $a_i, b_i$ that minimize the equation

$$Q(f_i, x_{W_i}) = \sum_{j=i}^{i+k-1} (f_i(t_j) - x_j)^2 = \sum_{j=i}^{i+k-1} (a_i t_j + b_i - x_j)^2 \tag{2.1}$$

is a linear regression of $x_{W_i}$, that satisfies the least squares criterion. The values $a_i, b_i$ are calculated as follows:

$$a_i = \frac{\sum_{j=i}^{i+k-1}(t_j - \bar{t}_l)(x_j - \bar{x}_l)}{\sum_{j=i}^{i+k-1}(t_j - \bar{t}_l)^2} \tag{2.2}$$

$$b_i = \bar{x}_l - a_i\bar{t}_l \tag{2.3}$$

where

$$\bar{t}_l = \left(\frac{1}{k}\right)\sum_{j=i}^{i+k-1} t_j \tag{2.4}$$

and

$$\bar{x}_l = \left(\frac{1}{k}\right)\sum_{j=i}^{i+k-1} x_j \tag{2.5}$$

The Moving Approximation Transform (MAT) [3] is the transformation of time series values $x = (x_1, x_2, ..., x_n)$ into a sequence of slope values (local trend) calculated in sliding window of length k:

$$MAT_k(x) = (a_1, a_2, ..., a_{n-k+1}) \tag{2.6}$$

For the time series $x$, its MAT transform is denoted as $MAT_k(x) = (a_{x_1}, a_{x_2}, ..., a_{x_m})$, where $m = n - k + 1$. The local trend association measure $(LTAM)$ is calculated for time series x and y of the same length n as the cosine of their corresponding MATs:

$$LTAM_k(x,y) = cos(MAT_k(x), MAT_k(y)) = \frac{\sum_{i=1}^{i=m} a_{x_i} \cdot a_{y_i}}{\sqrt{\sum_{i=1}^{i=m} a_{x_i}^2 \cdot \sum_{i=1}^{i=m} a_{y_i}^2}} \tag{2.7}$$

The dynamic local trend association measure $(DLTAM)$ is defined as a sequence of $LTAMs$ calculated for subsequences of $MAT_k(x)$ of length $L < m$:

$$DLTAM_{k,L,s}(x,y) = \frac{\sum_{i=s}^{i=s+L-1} a_{x_i} \cdot a_{y_i}}{\sqrt{\sum_{i=s}^{i=s+L-1} a_{x_i}^2 \cdot \sum_{i=s}^{i=s+L-1} a_{y_i}^2}} \tag{2.8}$$

where $s = 1, ..., m - L + 1$

The $LTAM_k(x,y)$ calculates local trend associations between time series $x$ and $y$ for all time domain and gives as a result one number. The $DLTAM_{k,L,s}(x,y)$ calculates local trend associations for smaller time domains showing dynamics of these associations. For a better visual correspondence to the original points, the $DLAM_{k,L,s}(x,y)$ values are assigned to the time points $s + \frac{k+L}{2} - 1$ in the plots.

The patterns of two time series x and y, which MAT slopes are ($a_{x_1}$, $a_{x_2}$, ..., $a_{x_m}$) and ($a_{y_1}$, $a_{y_2}$, ..., $a_{y_m}$) respectively, for $m = n - k + 1$, are as follows:

$$\begin{split}A(a_x, a_y) = (sgn(a_{x_1}) \cdot sgn(a_{y_1}), sgn(a_{x_2}) \cdot sgn(a_{y_2}), ...,\\ sgn(a_{x_m}) \cdot sgn(a_{y_m}))\end{split} \tag{2.9}$$

## 2.3 Textual Representations

Natural Language Processing (NLP) is a field of computer science that develops algorithms for the automatic processing of human language, the different applications NLP addresses range from detecting spam in e-mails, or correcting misspellings in word processors to more complex tasks such as author identification in a document, author profiling (age, genre, background) or automatic translation. Yearly there are worldwide contests in which the latest advances in some of the NLP applications are tested, two of those contests are PAN[1] and SemEval[2]. In order for the applications to work some subtasks have to be solved first.

### 2.3.1 Bag of Words (BOW)

One of the key subtasks on NLP is the textual representation. BOW is a traditional approach that has been widely used in the literature, this representation consist of a two-dimensional array or matrix where one axis represents the documents and the other represents the features or words for each of the documents. Each cell in the matrix correspond to the number of times that the feature appears in the document.

BOW is a subclass on n-gram representation. In the n-gram representation the $n$ consecutive words form the features instead of only one word. This way context is better represented than BOW (two words in different sentences do not always convey the same meaning). Besides the word n-grams, the character n-grams are another representation and, in combination with the former, are specially useful in the author profiling and authorship attribution tasks [4]. Skip-grams are another extension of n-grams in which the features are not only created by consecutive words but also by words that are $k$ places ahead or behind the word, disregarding the consecutive ones.

[1] http://pan.webis.de/tasks.html

[2] http://alt.qcri.org/semeval2017/

Depending on the application a word is represented by a one if it appears in the document or a zero if it does not, or it can be represented by the number of times it appeared in the document. The matrix representation of a BOW is expensive in memory for its sparsity so it is usually computationally represented as a linked list of lists or as a dictionary.

Weighting schemes are used in together with BOW with the rationale that for some tasks certain words are more important than others. Term frequency (TF) is the approach where a word is represented by the number of times it appears in the document. It is usually used with Inverse Document Frequency (IDF), which is a weighting scheme that takes into account that if a term occurs in many documents no matter what their class is, then this term is less useful to distinguish what class a given document belongs to.

Leung [5] introduced transformation techniques of text representation in the forecasting such as TF, IDF and CDF. They found that CDF performs better than IDF, this means that assigning more weight to the terms that appear the most within one category is more important than assign it to the terms that are more rare and distinctive within the three categories.

### 2.3.2 Word Embeddings (WE)

The text representations discussed in previous paragraphs have the problem of sparsity, i.e. the matrix that represent the documents is filled with too many zeros. In recent years the word embeddings (WE) representation proposed by Mikolov [6] has gained acceptance for its efficiency and for the great results as compared to BOW [7].

Word embeddings is the representation each word as a vector, is a recently proposed word distributed representation [8]

In this work the *word2vec* model[3] is used, which is a distributed representation trained on google news and contains three million words, each word being represented by three hundred dimensions. The semantics captured by this model allows powerful arithmetic operations among words that are semantically meaningful, a popular example of such operation of the trained model is: “king” − “man” + “woman” = “queen”.

Word embeddings has been used the task of sentiment analysis on twitter [9–11]

[3] `https://github.com/mmihaltz/word2vec-GoogleNews-vectors`

## 2.4 Machine Learning Classifiers

Machine Learning is a multidisciplinary field that develops algorithms that learn from data. It intersects with many areas such as computer science, statistics, mathematics, neurology, etc. Through the years many paradigms and algorithms have been developed. In this chapter some of the most popular algorithms are discussed.

In this study a supervised learning approach is used, this means that the algorithm is provided with examples of correctly labeled data in order learn from it (create a hypothesis). With the learned hypothesis the algorithm must able to label correctly or classify new data.

For an algorithm the data is represented as features e.g. the features that could represent the object book are its length, width and depth. For the NLP field, the features are represented in the BOW or WE models.

The data available is split into a training set and a test set. The algorithm learns from the former set and its performance is evaluated with the latter set.

The simplest learning algorithm is the perceptron algorithm. The notation is as follows: $n$ is the number of training examples, $m$ are the features, $\mathbf{x}_i$ represents a single example of the data, $\mathbf{y}_i$ represents correct class for that example, the training set is $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), ..., (\mathbf{x}_n, \mathbf{y}_n)$ and there are two classes $y \in \{-1, 1\}$. The objective is to learn a hypothesis as a function of the input data $h(\mathbf{x})$. The McCulloch-Pitts model of the perceptron neuron is shown in Eq. 2.10:

$$h(\mathbf{x}) = sign\left(\left(\sum_{i=1}^{m} w_i x_i\right) + b\right) \tag{2.10}$$

The algorithm needs to learn the synapsis weights $w_1, ..., w_m$ and bias $b$ that best fit the training set. The function $sign(.)$ makes the classification decision, $sign(s) = 1$ if $s > 1$ and $sign(s) = -1$ otherwise.

It is often useful for illustration purposes the case where there are two features, $m = 2$, each dimension of the vector is a feature of the object.. In this case the weights and bias form a two-dimensional line $w_1x_1+w_2x_2+b = 0$ and the learning problem can be interpreted as finding the line that separates the two classes of the data. This is illustrated in Figure 2.1. The goal of the perceptron training algorithm is to learn the synapsis weights and bias of a line that separates both classes.

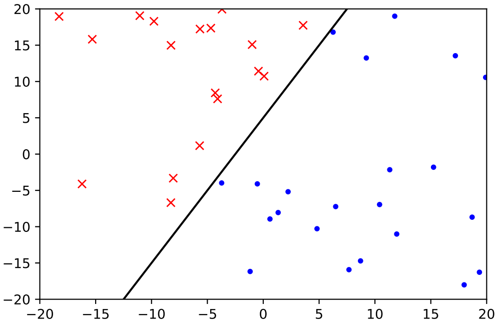


Figure 2.1: Example of two classes, red crosses and blue points, in a two-dimensional space

For notation purposes the bias is treated as a weight and incorporated in the vector of weights $\mathbf{w} = [w_o, w_1, ..., w_m]^\intercal$ where $w_0 = b$ and the symbol $^\intercal$ represents the transposed of a vector. In the same way a feature $x_0$ with value 1 is incorporated to the vector of features $\mathbf{x} = [x_0, x_1, ..., x_m]^\intercal$. The Eq. 2.10 can be rewritten as follows:

$$h(\mathbf{x}) = sign(\mathbf{w}^\intercal \mathbf{x}) \tag{2.11}$$

The algorithm consist of updating the weights through a series of $t$ iterations. This is done by picking one example $\mathbf{x}(t), y(t)$ that has been missclassified at a time and use it in the update rule shown in Eq. 2.12

$$\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t) \tag{2.12}$$

The perceptron training algorithm works well when the data is linearly separable, and although this is not the case in real scenarios the algorithm is useful to show the principle of learning from the data.

All the machine learning algorithms described below were implemented using the *Python's Sklearn Learn API* [12] [13].

### 2.4.1 K-Nearest Neighbors (k-NN) Classifier

k-NN is known to be a clustering algorithm however in this particular implementation it is a classifier. For this algorithm no model is constructed, just the training data is stored. For each point in the test set the euclidean distance is measured (although it is possible to use any other distance measure) to the $k$ nearest neighbors in the training set, then using a majority vote among the classes of the nearest neighbors the result will be the predicted class.

### 2.4.2 Naive Bayes (NB) Classifiers

The Naive Bayes classifiers use the Bayes theorem

$$P(y|x_1, ..., x_n)P(x_1, ..., x_n) = P(x_1, ..., x_n|y)P(y) \tag{2.13}$$

where $x_1, ..., x_n$ is a sequence of input features, and the naive assumption that all features are independent

$$P(x_i|y, x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) = P(x_i|y) \tag{2.14}$$

which can also be stated as

$$P(x_1, ..., x_n|y) = \prod_{i=1}^{n} P(x_i|y). \tag{2.15}$$

From (2.13) and (2.15) and the fact that $P(x_1, ..., x_n)$ is constant and it is possible to remove from the equation it follows that

$$P(y|x_1, ..., x_n) = P(y) \prod_{i=1}^{n} P(x_i|y). \tag{2.16}$$

Finally in 2.17 it is evaluated the class that has the maximum probability according to the probabilities it calculated on the training data.

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i|y) \tag{2.17}$$

Depending on the assumption over the distribution of data there are three Naive Bayes classifiers in the *Python's Sklearn Learn API*: Gaussian, Multinomial and Bernoulli. The distributions assumed for each classifier are shown on equations 2.18, 2.19 and 2.20 respectively.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}) \tag{2.18}$$

$$P(x_i|y) = \frac{\text{count}(x_i, y) + \alpha}{(\sum_{x \in V} \text{count}(x, y)) + \alpha n} \tag{2.19}$$

The function $\text{count}(x_i|y)$ counts the number of times that a feature $i$ appears on the class $y$. The function $\text{count}(x, y)$ counts the total number features of class $y$. The parameter $\alpha$ smooths the feature values so that if a words that was not seen in the training set appears in the test set the probabilities are not set to zero. $n$ is the total number features of the training corpus.

$$P(x_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i) \tag{2.20}$$

### 2.4.3 Decision Trees (DT) and Random Forests (RF)

In decision tree learning the goal is to create rules from the training set to make decision on the test set. The rules are created by evaluating how well a feature discriminates among classes. The better a feature can discriminate, the higher its hierarchy in the decision tree. This approach was used in the first articles documented in the literature review [5, 14, 15]. An advantage of this learning type is that the decision making process for classification is transparent to the user and the rules can be visualized in the decision tree. One of the disadvantages is that it can *overfit* the data, another disadvantage is that if the dataset is not balanced it may be biased towards the class with more samples. On Figure 2.2 it is shown an example on a decision tree on whether or not to assign credit to a person.

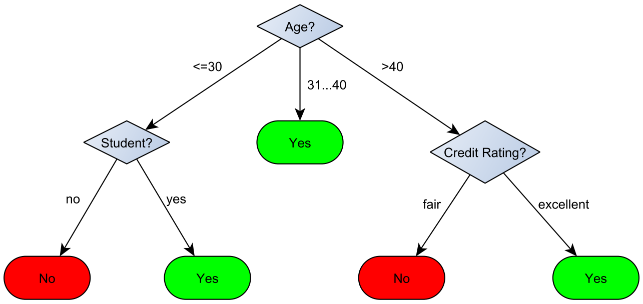


Figure 2.2: Example of a decision tree

Random forests are an extension of Decision Trees. In this algorithm many decision trees are created and a majority vote among them is implemented, using this technique the problem of *overfitting* is addressed. RF was reported to give best results among different classifiers tested on [16].

### 2.4.4 Stochastic Gradient Descent Classifier (SGDC)

In the introduction of this Chapter it was shown the perceptron model and its algorithm. In real case implementations the model is a function as shown in Eq. 2.21.

$$E(w,b) = \frac{1}{n}\sum_{i=1}^{n} L(y_i.f(x_i)) + \alpha R(w) \tag{2.21}$$

$L$ is a loss function that measures how well the hypothesis adapts to the data. $R$ is a regularization term a penalizes model complexity and $\alpha$ is a regularization parameter. This term is intended to reduce the ovefitting (when the hypothesis learns too well the training set and therefore is less able to do a good generalization on the examples that the model has not seen).

Some of the loss functions are:

- Least squares loss $L = (h(\mathbf{x}) - y)^2$
- Logistic loss $L = \ln(1 + e^{-h(\mathbf{x})y})$
- Hinge loss $L = \max(0, 1 - h(\mathbf{x})y)$

Some of the regularization functions are:

- L2 regularization $R(w) = \frac{1}{2}\sum_{i=1}^{n} w_i^2$
- L1 regularization $R(w) = \sum_{i=1}^{n} |w_i|$

Stochastic Gradient Descent Classifier is a linear classifier that implements the update rule in Eq. 2.22, it iteratively minimizes the error between the hypothesis $h(\mathbf{x})$ and the function that is being approximated.

$$w := w - \eta(\alpha\frac{\partial R(w)}{\partial w} + \frac{\partial L(w^\intercal x_i + b, y_i)}{\partial w}) \tag{2.22}$$

The algorithm learns iteratively the hypothesis as a function of the features $x$ that approaches the actual tags $y$ in the training set. One of the disadvantages of this classifier is the number of parameters, such as the learning rate $\eta$ (the step-size towards minimizing the error), the batch size (the number of samples it takes at a time), and a regularization parameter to reduce overfitting.

### 2.4.5 Logistic Regression (LR)

Logistic Regression Classifier uses the logistic loss function and, unlike the SDGC, it has available other solvers. The recommended solvers depend on the size of the dataset. For large datasets it is recommended the SGDC.

### 2.4.6 Support Vector Classifier (SVC)

The SVC, which is mostly known as Support Vector Machine (SVM) for classification (there are also SVM's for regression) maximize distance between the hyperplane that separates the data from different classes and for this purpose it calculates the support vectors. The support vectors are the examples that are closer to the hyperplane. The SVC is formulated as a maximization problem and is solved using quadratic programming and lagrangian multipliers which are outside the scope of this study. In this formulation it is possible to add kernels to modify the separating hyperplane so that it allows non-linear solutions.

### 2.4.7 Multi-Layer Perceptron (MLP)

The multi layer perceptron is a neural network model. It consists on many layers of perceptrons or neurons. The model of the MLP for two layers is shown in Eq. 2.23. $W_1$ and $b_1$ represent the weights and biases of all perceptrons in the first layer. $W_2$ and $b_2$ represent the weights and biases of all perceptrons in the second layer.

$$h(x) = W_2 g(W_1^{\intercal} x + b_1) + b2 \tag{2.23}$$

The activation function $g(z)$ defined as follows:

$$g(z) = \text{softmax}(z)_i = \frac{\exp(z_i)}{\sum_{l=1}^{k} \exp(z_l)} \tag{2.24}$$

The activation function combines the input features to a neuron into one output, the softmax activation function is useful for classification because it generates the probabilities that a given example belongs to each class and outputs the class with the highest probability.

## 2.5 Statistical Significance Tests

To test the statistical significance of the accuracy results with respect to a baseline the McNemar's test is applied with

$$\chi^2 = \frac{(|e_{01} - e_{10}| - \lambda)^2}{e_{01} + e_{10}}, \tag{2.25}$$

where

$$\lambda = \begin{cases} 1, & e_{01} + e_{10} < 25, \\ 0, & \text{otherwise.} \end{cases} \tag{2.26}$$

Here, $e_{01}$ is the number of examples that the baseline classified correctly and the classifier incorrectly and, correspondingly, $e_{10}$ is the number of examples that the classifier classified correctly and the baseline incorrectly.

McNemar's test is a measure of the significance between two classifiers considering the examples on the test set where the classifiers disagree. It can be the case that two classifiers have the same accuracy but their difference is statistically significant, this is because although the same number of samples were classified correctly, the samples were not the same ones.

## 2.6 Summary

In this chapter it was discussed the Moving Approximation Transform (MAT), the Local Trend Association Measure (LTAM), and the Dynamic Local Trend Association Measure (DLTAM) which are important for measuring the negative associations between time series. Two different word representation were reviewed, bag of words and word embeddings. The machine learning classifiers and its main parameters, and the statistical test to be used in Chapter 5 were also discussed. Also it was stated that the MAT will be used in Chapters 3 to 5 and that the textual representations, the machine learning classifiers and the statistical tests will be used in Chapter 5.

# Chapter 3

# Similarity of Time Series based on the Length of the Patterns of the Moving Approximation Transform (MAT)

## 3.1 Introduction

In the last years, one of the tasks that has attracted attention in data mining is the measure of similarity between time series [17]. For such purposes many methods have been proposed [18–20]. However there are cases in which two time series have an inverse behavior and as such, it is unlikely to be captured by a similarity measure. A real case scenario occurs in market share where there are two rival companies that are the dominant competitors, if one increases its market share, the rival company looses it.

Batyrshin et al. [3, 21] proposed a parametric method to analyze of associations between time series patterns based on subsequences of the time series (local trends). They developed an association measure that considers both positive and negative relationship between two time series, i.e. when one time series has positive trend while the other has negative trend. They considered real cases scenarios in different areas such as finance, economics or politics. A method that selects the size of the window used in the MAT was not addressed in previous work.

In this Chapter it is proposed a method to select the window size of the MAT. The window selection is useful to produce more significant results. In the method proposed the windows that provide the most information are selected.

Besides the window selection, since the positive and negative patterns that arise between two time series change in time, it is proposed a method to measure the associations based on the patterns.

The time series are presented in an association network. The edges of the network depend on how large the patterns are between two time series. In the last part, the time series with the largest length of continuous positive or negative patterns are clustered together.

As a result of the method two set of companies related to two industries are analyzed: oil and gas (Oil&Gas), and information technology (IT). For the first set it was found an association based on geographical differences between the companies. For the second set it was found an association based on the main field of the companies, software or hardware.

This chapter is structured as follows: The related work is presented in Section 3.2. The pattern based measures are presented in Section3.3. An automated method to select the window size of the MAT is presented in Section 3.4. The results are presented on Section 3.5. The conclusion is presented on Section 3.6.

## 3.2 Related Work

In this section some of the works on similarity measures between time series are briefly presented.

Das et al. [18] proposed an randomized algorithm based on dynamic-programming (DP) to measure similarity using the Longest Common Subsequence (LCSS). The algorithm considers similarity even if the time subsequence is not aligned in both time series, i.e. if the i-th point in one time series does not correspond to the i-th point of the other. This measure is invariant to amplitude shifting and uniform amplification. The is a threshold parameter that allows separation in the compared subsequences. The similarity measure is given by $\frac{l}{n}$, where $l$ is the LCSS and $n$ is the length of the time series.

Alcock et al. [19] considers the time series features. He divides the features into first and second order features. First order features are the mean, standard deviation, skewness and kurtosis. Second order features are the energy, entropy, inertial correlation and local homogeneity. These second order features were considered because they are widely used in image analysis. The time series are transformed to two-dimensional matrices. The transformation process is through the Q-quantization of the time series, which reduces the values of the time series to discrete levels, e.g. if the time series is $1, 2, 3, 4$, the two-level Q-quantization would be $1, 1, 2, 2$. The next step is the construction of the matrix $c(i, j)$, in which the point $(i, j)$ represent the number of times that a number in the time series with level $i$ followed by a point $j$ at a distance $d$. The distance $d$ is an adjustable parameter on their method.

Lin et al. [20] considers Dynamic Time Warping (DTW). The DTW besides considering the time shift as in the LCSS measure, also considers the "speed" at which the time series changes, i.e. If one time series changes in the same way as the other but in a smaller or larger interval. The implementation of DTW involves also DP. The DP matrix both time series are compared and a path that minimizes the accumulative distance between the time series is selected (the path that minimizes the distortion on one time series to match the other). There are some imposed restrictions in the search for the optimal path, which are the Sakoe-Chiba's band and the y Itakura's parallelogram.

## 3.3 Pattern Based Similarity Measures

The Moving Approximation Transform (MAT), as explained in Chapter 2, consist of replacing the time series values by the slopes of the lines that approximate their values using a moving window. A lineal regression is used to approximate the slopes using the least-squares criterion. Moving windows of a fixed size are used to determine the subsequence of points to approximate. In Figure 3.1 an example of the MAT of two time series is shown for a window $k = 5$. The positive associations occur when the slopes have the same sign and the negative associations when the slopes have the different sign. The measures of local trend association and local trend distances are obtained using the MAT. Those measures have the property of being invariant under normalization or linear transformations.

The patterns between two time series are obtained from the MAT slopes of the time series. First the sign of each slope is calculated. Then each point $i$ in the sequence of slopes of the time series $x$ is multiplied by the corresponding slope $i$ of time series $y$. This is shown on Eq. 2.9. A point in the resulting series of slopes, $A(a_x, a_y)$, is positive if the $i$-th slope in both time series have the same sign (positive association) or negative if the $i$-th slope in both time series have different sign (negative association).

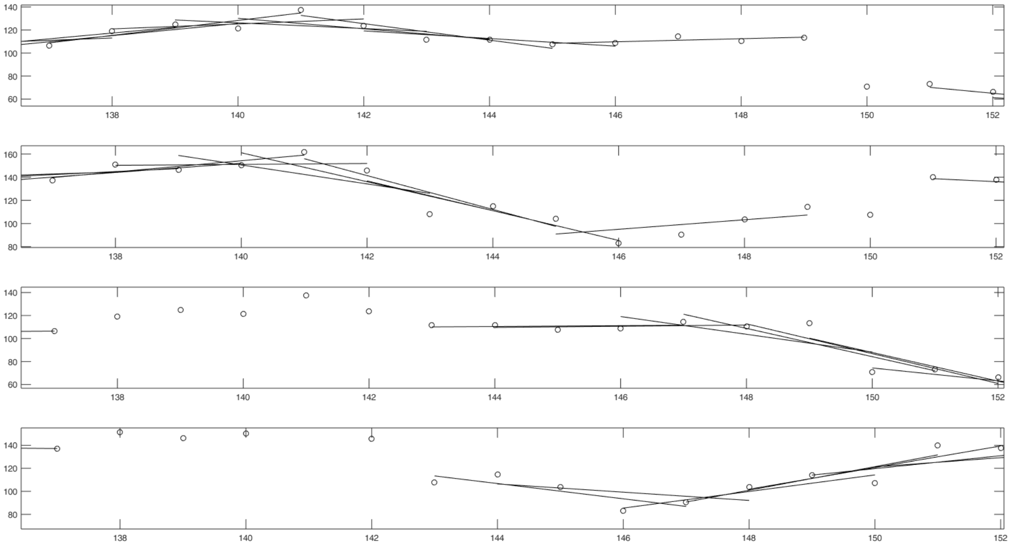


Figure 3.1: Positive (top) and negative (bottom) associations between two time series

In Figure 3.2 (bottom) it is shown the association patterns using the sign of the MAT. The top part of Figure 3.2 are the positive and negative slopes of the two time series. The continuous intervals $+1(-1)$ form the positive(negative) patterns. In the figure the sequence of patterns lengths is {10, -3, 57, -2, 10, -2, 97, -3, 15, -2, 22}. The patterns end in the month of November because the figure is aligned with the initial points of the slopes. In this section two methods are proposed to measure the similarity based on the length of consecutive positive or negative patterns. They are described below.

In the first method only the largest association pattern between two time series is considered. As for the visualization, only the associations greater than a threshold are shown. The threshold is calculated relative to the largest association in the set, e.g. if the length of the largest pattern found in all time series under analysis is 100, and a threshold of 85% is selected, then in the association graph only the time series whose largest patterns are larger or equal than 85 are connected by an edge.

The second method applies considers all patterns between two time series. The patterns are separated into positive and negative ones. The sum and length of the patterns is considered according to the next equation,

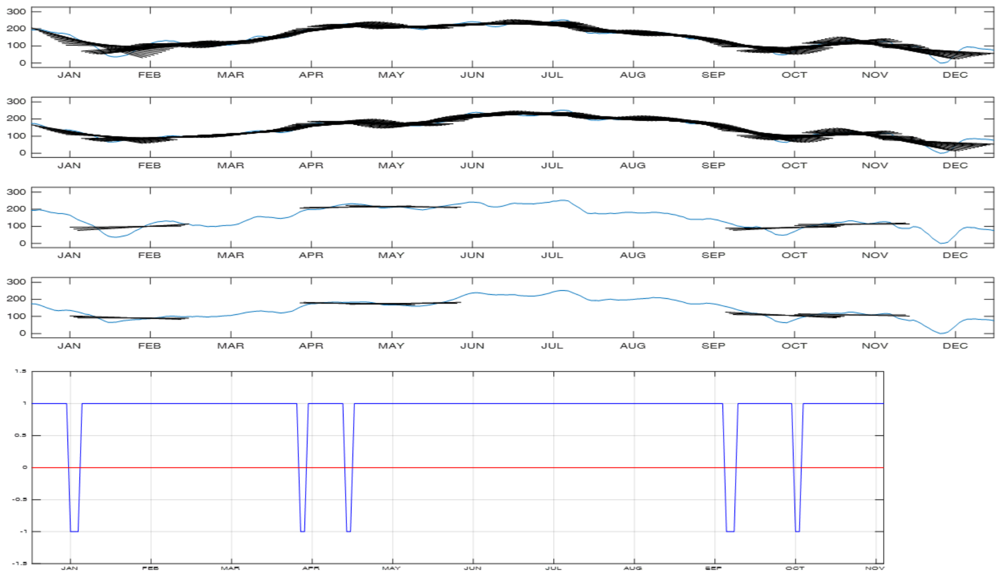


Figure 3.2: Association between Chevron and ExxonMobil with a window $k = 30$ (top) and the sequences of positive and negative patterns (bottom)

$$SIM(x) = \frac{sum(x)}{k_{max} - k + 1} \cdot \frac{\kappa - len(x)}{\kappa - 1} \tag{3.1}$$

where

$$\kappa = \frac{ceil(k_{max} - k + 1)}{2} \tag{3.2}$$

the *ceil* function rounds the argument to the nearest integer greater or equal than the argument, the window size is $k$, the maximum value that the window can take is $k_{max}$ , the sum of all patterns is $sum(x)$ and the length of the pattern list is $len(x)$.

In Equation 3.1 the first factor is the sum of all patterns normalized with respect to the greatest possible sum. The second factor represents the length of the pattern, it has a minus sign because more patterns convey less similarity, e.g. the positive pattern 5 implies more similarity between time series than the pattern 1,1,1,1,1, it means that the time series move together continuously. A threshold is used in this second method to display only edges above it in the association graph.

In both methods the positive and negative association are shown in different graphs.

## 3.4 Window Selection for the MAT

A window of size $k$ indicates how many points are considered in the linear regression. At most $n - k + 1$ points are allowed.

As it was mentioned in the introduction of this chapter, the window selection is important to determine what information is being retrieved. In [3] it was shown the effect of the window choice, smaller windows are more sensible to changes in the time series trends while bigger windows detect the overall trends, the latter trends are easily found by eye inspection.

In this window selection process, the window size is only explored in the range between the minimum possible, 2, and a quarter of the time series length. This is done to reduce the search space. It can be said that for smaller windows the similarity is more significant than for bigger windows. As the window size tends to the maximum value it will only measure how the first points in the time series are different from the last points.

The window selection considers the maximum similarity values between two companies versus the rest of the companies and the number of continuous windows in which these two companies are maintained as the ones with greater similarity. In Figure 3.3 the positive similarity values are plotted against the window size. Each of the plotted lines represent the similarity between two companies. Also in Figure 3.3 the similarity between two companies, MSFT and AAPL is shown. The similarity is pointed while these companies, maintain the highest value of similarity.

It is obtained the highest value for each window size and the companies pair that it belongs to. Only values of similarity greater than 0.5 are considered. It is also obtained the length of the pattern. These two quantities are multiplied and the greatest results correspond to the pair of companies where the window will be located. In can also be observed in Figure 3.3 that as the window size increases, the similarity values converge to 0 or 1. This is because, as it was explained before, the larger the window size the less information is obtained, only the global trend of the time series.

## 3.5 Results

The dataset for the experiments in this chapter consist on the daily close prices of oil related companies in the year 2014. The data was downloaded

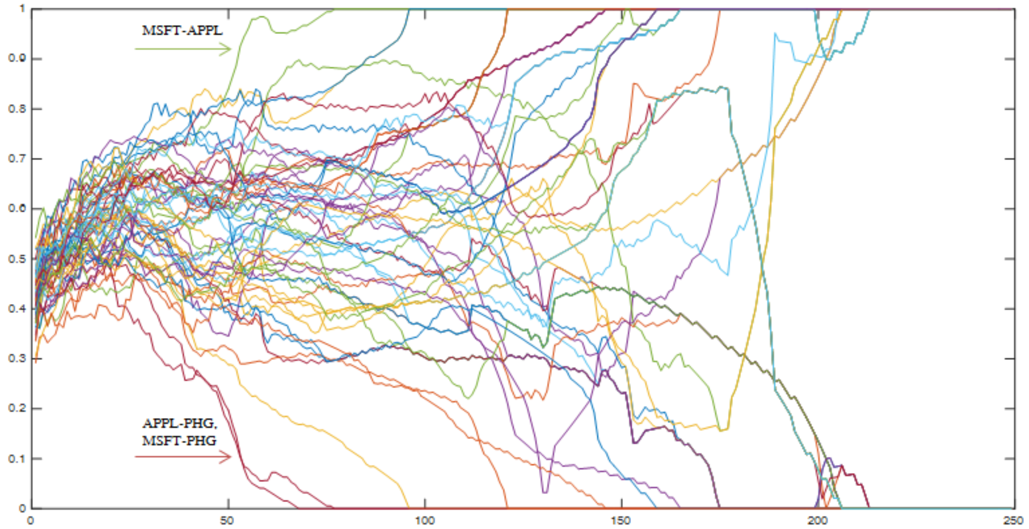


Figure 3.3: Positive similarity (vertical axis) of the close price of IT companies in 2014 using different window sizes (horizontal axis)

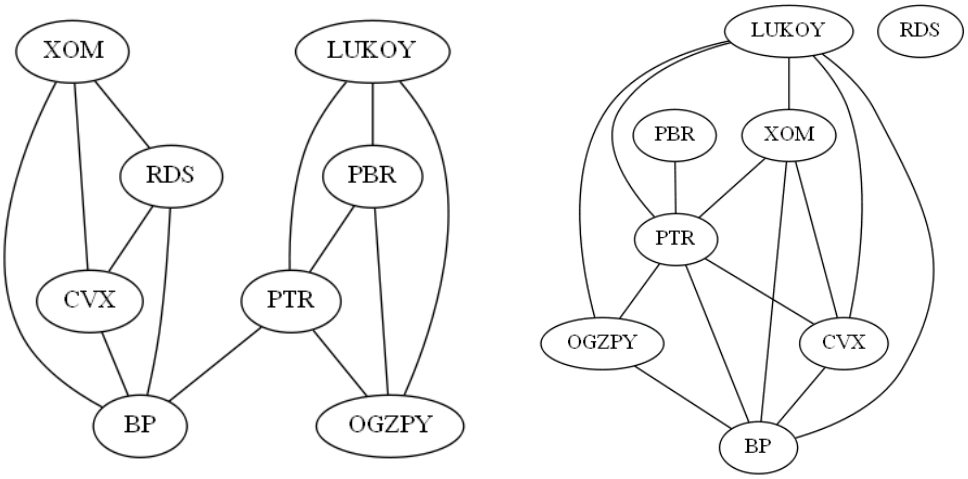
from *Google Finance*[1]. The company stock symbols and their time series can be seen in Appendix A.

The time series patterns that were at least 85% as large as the largest pattern in all time series are displayed. Only patterns which length is at least 40 were considered, this criteria was selected after looking at the pattern lengths in the data. The time series that are connected by edges in the association graph are those which lengths were at least by some proportion as large as the pattern with the largest length in all company pairs.

In Figure 3.4 the clustering of the positive patterns of oil related companies in 2014 and 2015 respectively is shown.

Using the first method for the Oil&Gas companies in the year 2014 Figure 3.4a it was found an association graph that almost completely separates the companies into two strongly connected clusters. The first cluster has some of the companies that belong to the seven sisters (whose countries belong to the Organization for Economic Co-operation and Development (OECD)) BP, Chevron, Royal Dutch Shell and ExxonMobil. The second cluster has companies which countries do not belong to the OECD, which are Gazprom, Petrobras, Lukoil and Petrochina. In the first cluster the companies are

[1] `http://www.google.com/finance`



(a) 2014

(b) 2015

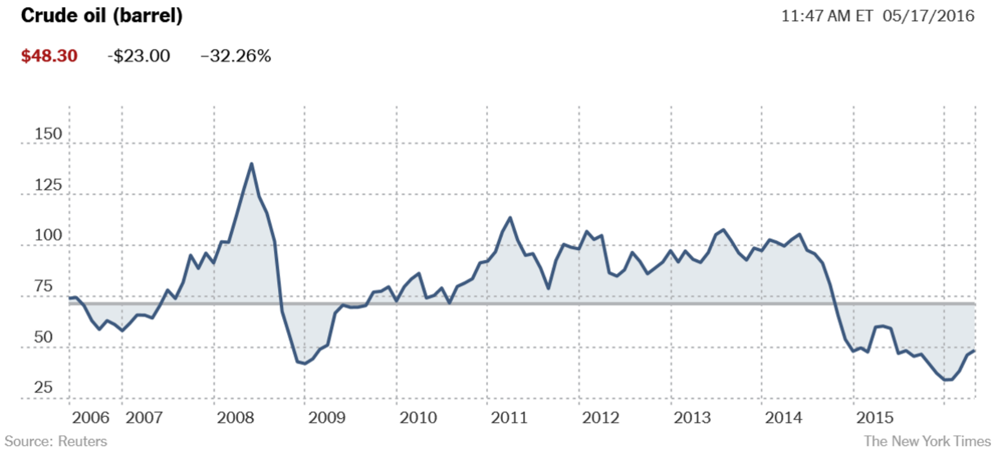
Figure 3.4: Clustering of oil companies using the first method using a window $k = 30$

| | | 2014 | 2015 |
|---|---|---|---|
| Oil&Gas | Positive | 55, 10, 37 | 34, 61, 24 |
| | Negative | 56, 41, 30 | 63, 57, 56 |
| IT | Positive | 59, 33, 45 | 40, 47, 12 |
| | Negative | 55, 24, 63 | 63, 27, 35 |

Table 3.1: Suggested windows

considered the seven sisters and the companies in the second cluster are considered the new seven sisters. The same method was applied to the time series in the year 2015 Figure 3.4b and the results were different, in this latter case, all companies show high positive association regardless of their geographic location. This may be explained in part because all time series showed a similar behavior in that year since at the end of 2014 the oil prices dropped as it is shown in figure 3.5. According to the initial criteria (only consider patterns grater than 40) no negative associations were found for this method and dataset.

With the proposed method for window selection and the dataset the suggested windows are shown in table Table 3.1. Using these windows and the



Source: Reuters

The New York Times

Figure 3.5: Historical price of the oil barrel

measure of the second method (Eq. 3.1) the companies were clustered, this can be seen on Figures 3.6 to 3.10.

## 3.6 Conclusion

In this chapter it is proposed the clustering of time series based on the association patterns. The patterns are obtained using the MAT transform and the association of their trends. Using the proposed method it is possible to display graphically positive and negative associations. Comparing the visualization method against the cosine similarity in [21], with the new method it is possible to display graphically the sequences of positive and negative associations. The results have a natural interpretation in the dataset tested of IT and Oil&Gas companies.

Previous work used window selection intuitively. In this chapter for the first time it is proposed a window selection method that obtains information of the time series with the greatest association compared against different window sizes. The similarity values are plotted with respect to the window size, and using this plot the time series pairs that exhibit greater similarity are identified for each window. The maximum values of similarity that are separated among each other are considered as candidate windows.

It is also proposed a visualization method of the time intervals that shows the association periods clearer than in the previous publications that consider

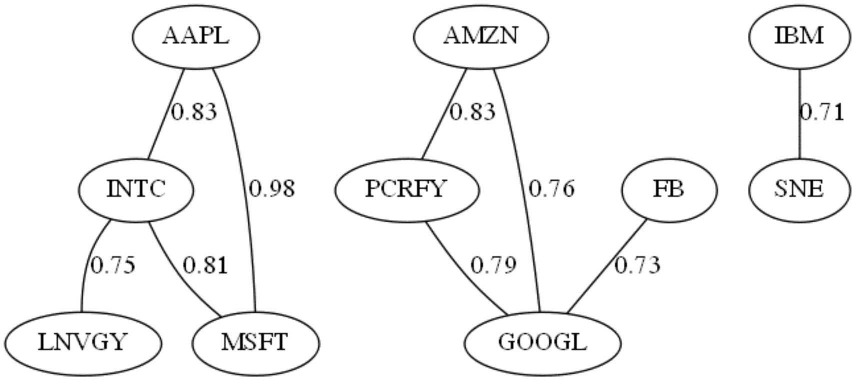


Figure 3.6: Positive association of IT companies in 2014, $k = 59$

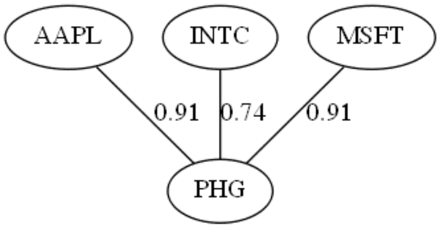


Figure 3.7: Negative association of IT companies in 2014, $k = 55$

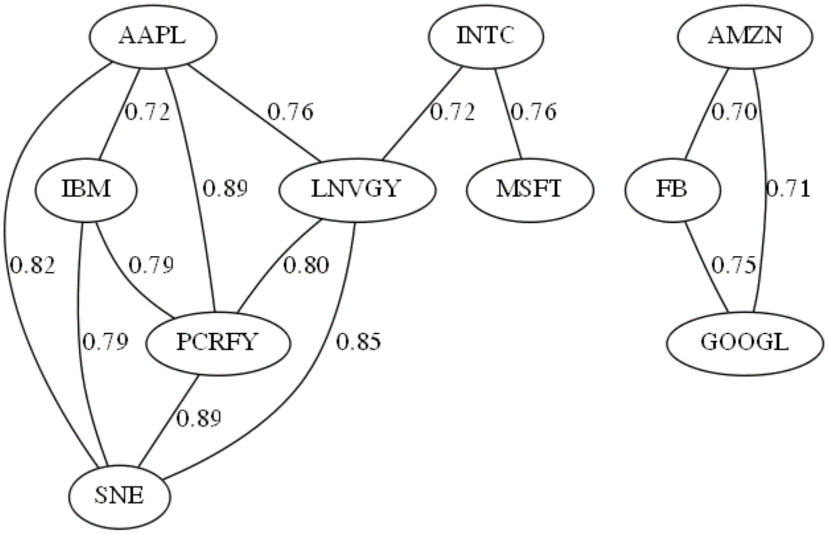


Figure 3.8: Positive association of IT companies in 2015, $k = 47$

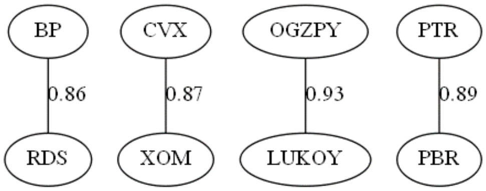


Figure 3.9: Positive association of Oil&Gas companies in 2014, $k = 33$

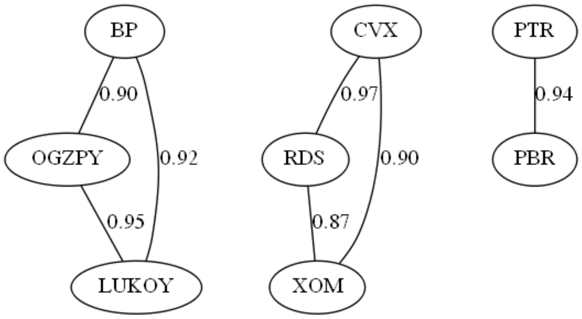


Figure 3.10: Positive association of Oil&Gas companies in 2014, $k = 55$

the MAT. The proposed methods also extend the possibilities of time series analysis considered previously [3, 21].

# Chapter 4

# Dynamic Local Trend Associations in the Analysis of Financial Time Series Co-movements

## 4.1 Introduction

Last years the problem of development of new methods of time series analysis has attracted much attention [3,17,21,23–32]. Many methods have been developed addressing the problem of time series similarity [17,25], however some applications and studies require not a measure of similarity but a measure of association between dynamics of time series. The task of analysis of dynamic associations of time series is to find time intervals where two time series move together or in inversely direction. Dynamic association analysis has different applications such as the identification of competitors in the stock market, that can be allies or enemies at different time periods [21], portfolio optimization [23], stock market forecasting [26], etc.

Different methods have been proposed to analyze the co-movements of financial time series [3,21,24,27–32] and many of such approaches are based on the concept of correlation. In [27] co-movement is considered as positive correlation of returns among different traded securities. In [32] correlation is used to analyze co-movement of commodity prices. Local correlations are considered in [31].

Most of the methods applied to analyze financial time series co-movements are usually based on traditional statistical or signal processing methods, including correlation analysis. But often there is no rationale for the application of these methods because in finance the concept of trend is more important than the concept of frequency. The correlation coefficient does not take into account the time ordering of time series values, so the application of the correlation coefficient to analysis of time series co-movements can be misleading. In the next section, the examples when the correlation coefficient cannot detect co-movement of time series are discussed.

Two time series have positive co-movement in two neighboring time points if both time series increase, negative co-movement if both decrease and contra-movements if one of them increases and another decreases. The methods of analysis of time series co-movements based on the analysis of trends of time series have been considered in several works. In [29, 30] nonparametric tests for co-movements between time series are considered, in these works the co-movement of time series is based on the comparison of the signs of the time series values change in neighboring time points.

A more general approach was considered in [3, 21] where time series values are replaced by series of local trends obtained as slope values of linear regressions of time series in sliding windows of a given size. In [3, 21], the association between two time series is considered as positive if the local trends of these time series in the same windows have the same signs, and as negative if they have opposite signs. Comparing the terminology of the works [3, 21] and [29,30], in two neighboring time points two time series are positively associated if they have positive or negative *co-movement* and they are negatively associated if they have *contra-movement*. Based on local trend association measure (LTAM) [3] the work [21] proposed the method of construction of association patterns in two time series defined as the longest sequence of windows having the same sign of associations for these time series, i.e. patterns of positive associations and patterns of negative associations. This method was used for detecting competitive companies based on the analysis of their stock prices.

In [24] a dynamic correlation model, called DCC-GARCH-JGR, is proposed to investigate how worldwide oil-related events impact the correlation between oil price and the stock market price of oil-importing and exporting countries.

In this chapter it is proposed the method of dynamic local trend associations of time series based on the method of local trend associations considered in [3]. LTAM measures global associations between time series based on comparisons of all local trends of time series [3]. The method proposed in this paper gives possibility to find the time regions (intervals) where two time series are positively or negatively associated.

This chapter has the following structure. Section 4.2 shows that the

| | t=i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (a) | $x_a$ | 1 | 2 | 3 | 2 | 3 | 2 | 3 | 4 | 3 | 4 | 3 | 4 | 5 |
| | $y_a$ | 6 | 7 | 8 | 6 | 7 | 5 | 6 | 7 | 5 | 6 | 4 | 5 | 6 |
| (b) | $x_b$ | 1 | 2 | 3 | 2 | 3 | 2 | 3 | 4 | 3 | 4 | 3 | 4 | 5 |
| | $y_b$ | 6 | 5 | 4 | 6 | 5 | 7 | 6 | 5 | 7 | 6 | 8 | 7 | 6 |

Table 4.1: Synthetic time series

correlation coefficient can be useless or misleading in measuring time series co-movements. Section 4.3 shows the results of applying the new method to financial time series. Last section contains discussion and conclusion.

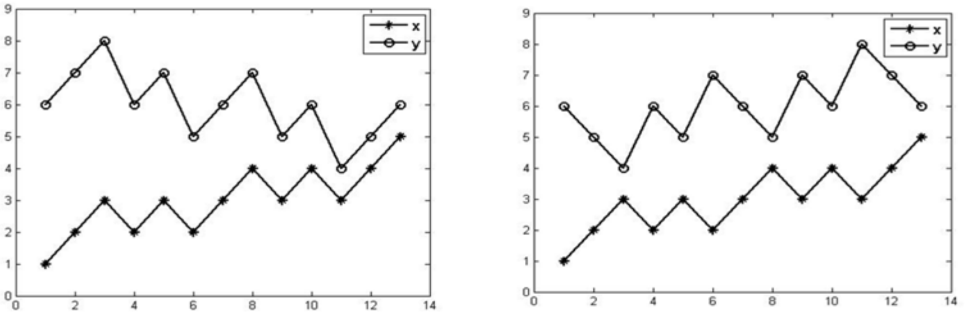
## 4.2 Correlation Coefficient and Time Series Co-movements

Pearson's correlation coefficient

$$corr(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 . \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.1}$$

plays a fundamental role in the analysis of relationships between variables but it can be useless or misleading in measuring of co-movements of time series because it does not take into account the time ordering of time series values.

Consider the example given in Table 4.1 and Figure 4.1. The pairs of time series $x_a$, $y_a$ show excellent co-movement (Figure 4.1a) and the pairs of time series $x_b$, $y_b$ show excellent contramovement (Figure 4.1b). The reasonable measure of time series co-movements should have positive value in the first case and negative value in the second case. Both pairs of time series are composed from the same pairs of points but ordered in different manner. For example, the pair $(2,7)$ located in time series $x_a, y_a$ on the position $i = 2$ is equivalent to the point at the position $i = 6$ in time series $x_b$, $y_b$ . The correlation coefficient does not take into account the time ordering of time series values and gives for both cases $corr(x_a, y_a) = corr(x_b, y_b) = 0$. This example shows that the correlation coefficient and its modifications can be useless or misleading in analysis of time series co-movements.



(a)

(b)

Figure 4.1: Examples of the positively (a) and negatively (b) associated time series $x$ and $y$ using the same points of Table 4.1 where the correlation coefficient is $corr(x, y) = 0$

## 4.3 Results

First, the results of applying the $LTAM_k(x, y)$ and the $DLTAM_{(k,L,s)}(x, y)$ to time series from Table 4.1 are shown. Using $LTAM_k(x, y)$ for sliding window sizes $k = 2$ and $k = 3$ positive associations between $x_a$ and $y_a$ ( $LTAM_2(x_a, y_a) = 0.94$ , $LTAM_3(x_a, y_a) = 0.77$ ) and negative associations between $x_b$ and $y_b$ ($LTAM_2(x_b, y_b) = -0.94$ , $LTAM_3(x_b, y_b) = -0.77$ ) are obtained. These results correspond to human perception about co-movement of time series $x_a$ and $y_a$ and contra-movement of time series $x_b$ and $y_b$. Remember that Pearson's correlation coefficient cannot capture this information about co-movement and contra-movement of these time series giving value $corr(x_a, y_a) = corr(x_b, y_b) = 0$. Figure 4.2 depicts $DLTAM$ values for time series $x_a$ and $y_a$ calculated for parameter values $k = 3$ and $L = 2, 3$. As one can see $DLTAM$ has high positive dynamic local trend association values corresponding to human perceptions about co-movement of these time series. Figure 4.3 depicts $DLTAM$ values for time series $x_b$ and $y_b$ calculated for parameter values $k = 3$ and $L = 2, 3$. $DLTAM$ has high negative local trend association values corresponding to human perceptions about contra-movement of these time series.

Using the dynamic local trend association measure two time series are analyzed: the daily closing prices of the companies BlackBerry Limited (NASDAQ:BBRY) and Apple (NASDAQ:AAPL) between February 19, 2013 and

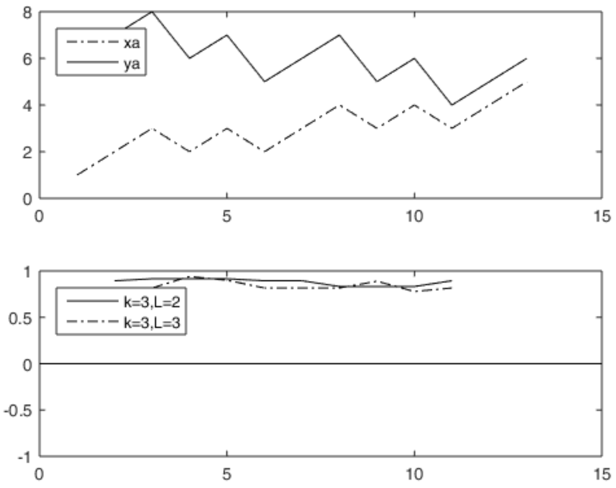


Figure 4.2: Time series $x_a$ and $y_a$ and their $DLTAM$ values

February 14, 2014, see Figure 4.4. The time periods of positive dynamic local trend associations between these time series mainly in the first part of the time series (until point 150 except around point 100) and negative local trend associations on the second part (after point 150) were found. Blackberry and Apple can be considered as competitive companies during the time period after the point 150. As it was mentioned in [21], the negative associations between financial time series can give more adequate information about possible mutual relationship between competitive companies because the positive association can be caused by a general tendency of stock market when many companies have similar co-movements. As it can be seen on Figure 4.4, negative associations correspond to contra-movements (in windows of size $k = 30$), and positive associations correspond to co-movement of time series, that is, the proposed measure of dynamic local trend associations can capture local co-movements and contra-movements of time series.

The following events were retrieved from Google, and correspond to news that are related to the companies during the same period of the time series . The events are: Apple launched new 16 GB iPod Touch (E1), BBRYs Q1 earnings below expectations (E2), BlackBerry Puts Itself Up for Sale (E3), BlackBerry getting bought (E4), Apple shows off iPad Air (E5), BlackBerry is replacing its CEO (E6) and BlackBerry launched Z30 in USA (E7).

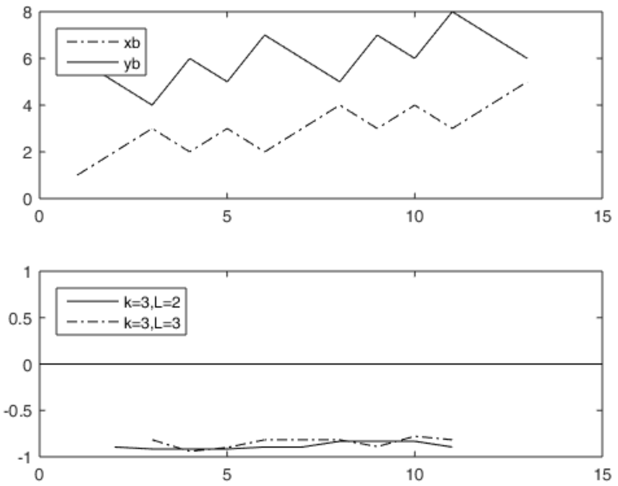


Figure 4.3: Time series $x_b$ and $y_b$ and their $DLTAM$ values

The negative association around point 100 may be explained by the low earnings report of Blackberry, after this report its price plunged, while its competitor's price rises. The beginning of the second negative association also has a possible explanation because at the same time Blackberry is being bought, Apple is releasing the iPhone 4S, 5C and 5S on September 20, 2013. The events that occur later, while the negative association is maintained, are positive for Apple (the announcement of the iPad Air) and negative for Blackberry (the replacement of its CEO). Note that the negative associations in time periods 150–200 and after 210 are caused by different contra-movements of time series: increasing of APPL price in the first period and decreasing in the second one, and, on contrary, decreasing of BBRY price in the first period and increasing in the second one. The last can be caused by launching Z30 in USA by BlackBerry.

## 4.4 Conclusion

It was shown that the correlation coefficient can be misleading in the analysis of time series co-movements. This paper proposes the dynamic local trend association measure that captures the changing positive and negative associations between time series. This method was tested using real time series

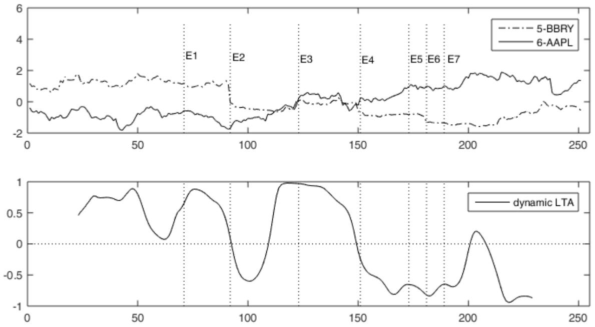


Figure 4.4: $DLTAM$ between daily close price of Blackberry and Apple with the most relevant events using windows $k = 30$ and $L = 16$

from the stock market. A brief set of news related to them was considered. It was also shown, using the $DLTAM$ plot that the dynamic local trend associations of the two rival companies and their news are related.

# Chapter 5

# Analysis of Relationships between Tweets and Stock Market Trends

## 5.1 Introduction

With the advances in NLP, machine learning techniques and the availability of data, the task of forecasting the stock market using textual information has grown in popularity. The first works used news as a source of information with promising results [5, 15, 33]. Recent works use the social media as input [34–36]. Social media includes stock market events, news, opinions and insights from investors.

In this chapter the experiments are conducted using Twitter data. A corpus was created from tweets related to nine IT companies. Twitter was chosen because it is a popular platform and also because the messages or tweets are tagged with the company's stock symbol by the users themselves.

In this chapter the temporal relationship between some stock price indicators and the volume and content of the tweets is analyzed. It is hypothesized that it is possible to identify tweets depending on the positive, negative or neutral market trends. Two textual representations are used, bag of words (BOW) and word embeddings (WE). For the latter *word2vec*, a model trained on news, was used.

Previous works [37,38] have compared both textual representations in the task of sentiment analysis using stock related tweets. They manually tag a set and then split it into training and testing sets where machine learning classifiers are evaluated. However the approach in this chapter is to tag a tweet automatically by the trend in which the tweet appears, i.e. if the tweet is generated while the prices rise or fall, or while the prices show a steady behavior. The results are contrasted against a baseline that considers the distribution of the corpus. Not many related works consider how the distribution of the data may influence the results [39].

The rest of the chapter is structured as follows. In Section 5.2, the related work is presented. In Section 5.3, the corpus and the experiments are described. In Section 5.4 the experiments with tweet volume are described. In Section 5.5 the experiments with the content of the tweets are described. Section 5.6 includes the conclusions.

## 5.2 Related Work

In this section are analyzed some studies that investigate the relationship between financial time series and different sources of text, such as social media, financial reports and the news. Also the approaches that took place in the last 20 years and the principal variables involved in the problem are discussed.

In financial analysis there are two classical forecasting paradigms, Technical Analysis [40] and Fundamental Analysis [41]. In Technical Analysis the historical prices are the principal indicator of future prices. Different methods such as auto-regression, ARIMA, state-space models or neural networks have been incorporated to financial analysis under this paradigm. On the other hand Fundamental Analysis relies on financial indicators such as the reported earnings, expected returns, balance sheets and also other global indicators such as the overall economy, unemployment, welfare, etc.

An approach that has been gaining importance in the last years is the use of textual information as a predictor of the financial market. In 1982 Klingemann [42] (as cited in [43]) studied the relationship between the Germany economy and wealth related words in the speeches of the Emperor between 1871 and 1914. That is the first study that was found that deals with the relationship between textual information and the economy in some form.

In the literature the financial market is addressed at three levels:

- **Company level:** Focus on the price of individual stocks e.g. “Apple Inc.” or “Facebook Inc.”.

- **Industry level:** Includes all companies that belong to the same industry, e.g. “Telecommunications” or “Public health”.

- **Global level:** Considers indicators of the global economy that include the most important companies in a country, e.g. DJIA or S&P 500.

Just a few works address the three market levels, most of them fall on either the company or the global levels.

### 5.2.1 Sources of Information

The principal sources of information for textual analysis are financial reports, news and social media.

#### 5.2.1.1 Financial Reports

The works that use financial reports are scarce, however, this approach allows the integration of financial indicators that are used on the *Fundamental Analysis*.

Lee et al. [16] created a corpus of financial reports that ranges form 2002 to 2012. They tested several machine learning methods, the method reported with the best results was Random Forest. In this article financial and text features are combined, both extracted from the reports. Twenty one financial features are used, the most helpful was the earnings surprise (the difference between the company's expected and reported earnings). It was found that adding text features to financial features improved the accuracy from 50.1% to 55.5%.

In Zhang's dissertation [7] the Lee's corpus was used. They implemented three different neural networks architectures. Although they did not considered the financial features but only language features the accuracy they reported was better than Random Forest and Support Vector Machine classifiers. They selected only 15 out of the 1500 companies in the corpus.

#### 5.2.1.2 News

Using news as source of information is a popular approach. One of the first research works that investigates such strategies is done by Sankaran [14]. In his work, traders are asked to assign weights to different attributes they call "unmeasurable factors", including government policies or political news that may influence the exchange rates. In this approach the knowledge of experts is taken from the news they read and other internal information they may posses, but the news are not automatically processed.

Leung [5] creates a system that requests the user to manually download news before the market opens to predict the Hang Seng Index, generating decision rules from the news articles. Later, Wüthrich [15] extended the Leung's approach by adding other global indexes such as DJIA and FTSE and improving the processing techniques. This last approach resulted in greater performance forecasting the DJIA, it was attributed to the fact that most news sources they used were from the U.S.

Lavrenko et al. [33] collect news automatically, they consider prediction as a classification problem. The text from the news is aligned with the future financial time series trend. He analyzed different companies, one of his interesting findings is that the same word may have different effects for different companies. The set up of prediction as a classification problem has been widely used in the literature [7, 16, 44].

#### 5.2.1.3 Social Media

On recent years it is common to use information from social media, considering that it is a popular channel for communication and that it is changing the ways users interact with the news [45]. Social media includes information from different sources. In [46] it was found that 85% of the topics in social media are related to news.

Twitter is a micro-blogging social media platform launched in 2006. Its mission is *"To give everyone the power to create and share ideas and information instantly, without barriers"*. It has 313 million active users[1] and any topic can be tackled in this platform including discussions related to the stock market. When a company stock is being discussed it is clearly referenced by preceding its stock symbol with the dollar sign e.g. $MSFT stands for the Microsoft's stock symbol.

Some of the advantages of twitter are the inclusion of different sources of information and the summarization of this information, since only messages shorter than 140 characters are permitted. One of the disadvantages is that noisy tweets in the form of advertisements or even false events can be found. Wolfram [47] points out that his results could have been improved handling better the noise. Zhang et al. [48] only consider re-tweets with the rationale that if a message has been re-tweeted then it is more relevant to users.

Social media can be exploded in different manners. De Choudhury et

[1] `https://about.twitter.com`

al. [49] correlate the magnitude of individual stocks's price change with features of user interaction such as the number of comments on a post, the numbers of replies to them, the elapsed time between each comment and other several features. Another interesting approach by Ruiz et al. [50] is the representation of social media as a graph and the analysis of correlations between the features of the graph and the price and volume of company stocks.

### 5.2.2 Textual Representations

There are three main textual representations found in the literature regarding the task of forecasting the stock market from text, bag of bords (BOW), word embeddings (WE) and sentiment analysis (SA).

#### 5.2.2.1 Bag of Words

The comparison of BOW different textual representation has not been addressed in many works. Schumaker et al. [51], at the company stock level, compares three textual representation techniques: all the words in the document, Noun Phrases and Named Entities (real world persons or organizations). The results favor the Named entities approach.

#### 5.2.2.2 Word Embeddings

Dickinson [37] and Pagolu [38] compare BOW and WE representation in the task of sentiment classification of stock related tweets. They manually tagged the tweets with the sentiment, represented them in BOW and WE and implemented a machine learning classifier to compare the results. Dickinson found BOW (accuracy of 68.5%) performing better than WE (accuracy of 63.4%) and Pagolu did not report a significant difference between both approaches (BOW accuracy of 70.5% and WE 70.2%).

#### 5.2.2.3 Sentiment Analysis

Sentiment analysis (SA) is an application that has attracted the interest in the field of opinion mining. For example, companies may be interested in knowing what the people think about a product that has been launched or politicians may be interested in determining the sentiment towards a candidate, etc. In this task its important to identify the transmitter, the sentiment, the object, the attribute of the object and the time. The most common approach toward sentiment analysis is the use of lexicons with positive and negative values for words (e.g. terrible has a more negative connotation than bad). Counting the number of times that such words appear in a text can give a sentiment polarity to the text.

Sentiment analysis has become a popular approach in the task of forecasting the stock market. It is used commonly for forecasting large indicators such as the DJIA or the S&P500 [34, 36, 52, 53], but has also been used for predicting the stock price of individual companies [35, 54, 55].

Tetlock [52] uses a psychosocial dictionary, the General Inquirer's Harvard IV-4, to measure the sentiment of a Wall Street column, he found that pessimism causes downward trends and that either high or low pessimism can predict high trading volume.

Bollen et al. [34] predicted the daily up and down changes in the closing values of the DJIA using different dimensions of moods such as happy, alert or calm. Each mood is aggregated in a time series in a daily manner. Bollen found a high correlation between the mood calmness and the DJIA and even that the former can predict the later six days beforehand.

Nonetheless the results of sentiment analysis at company level are not encouraging, Oliveira et al. [55] did not find predictive power for the returns even after using five different sentiment lexicons, they also did not find a significance difference among the lexicons.

De Fortuny et al. [56] , also at company stock level, compares BOW and sentiment analysis finding that out that the later performed worse than the former and even worse than random. They attribute this behavior to the fact that sentiment lexicons are usually extracted from different contexts, such as book or movie reviews and also in finance sometimes the nouns and verbs are more informative for decision making than the adjectives. A good example is the keywords used by [5] who obtained them by looking at the newspapers such as “shares surged”, “market falls”, “bond lost”.

Lee et al. [16] reached the same conclusion after using generic and specialized sentiment lexicons, none of such approaches boosted the accuracy results predicting individual company stocks.

From the evidence in the literature it seems that sentiment analysis is more useful at the global market level than at the company stocks level.

### 5.2.3 Volume

Apart from stock price forecasting, the transaction volume is another financial variable than has been studied, it can be defined as the number of transactions performed in a given period of time. The volume of transaction is useful to study because in technical analysis it determines the importance of the changes in price and the risk involved in a transaction.

Ruiz et al. [50] tested correlation at different lags between volume and graph features generated from Twitter interaction. The greatest correlation was found at lag zero. Mao et al. [57] found a weak correlation between price and tweets mentioning the stock symbol AAPL, and moderate correlation to the daily volume traded. Sprenger et al. [58] and Oliveira et al. [55] found that the message volume is useful to predict the next-day trading volume using regression analysis.

### 5.2.4 Performance Metrics

In the literature different performance metrics have been used [59]. They depend on the experimental setup.

The most common experimental setup is to transform the forecasting problem to a classification problem [34, 56]. In order to predict, the value is binned into discrete quantities e.g. 1 (positive), -1 (negative) and 0 (neutral). To bin the values a threshold for the price change must be selected beforehand. If the price change is greater or equal than the threshold then the price change is considered positive, if the price change falls below the threshold the tweet is considered negative, otherwise the tweet is considered neutral. To evaluate the first experimental setup, measures of accuracy, precision and recall are used. This experimental setup must consider at least a majority class baseline, if this is not considered, then the accuracy measure can be misleading, this is particularly true in cases where the classes are not balanced, which is usually the case in the stock market where some companies show an overall downward trend and others upward trend.

Another typical experimental setup uses time series forecasting performance metrics [48, 58]. In this type of experiment the real values are predicted, and not only one value, but a horizon $h$ of prediction is defined, i.e. to forecast the next $h$ values in the time series . The evaluation in this type of setup are: The coefficient of determination ($r^2$), the Root Mean Square Error (RMSE) or the Mean Absolute Percentage Error (MAPE).

Finally an evaluation method found in the literature is to measure directly the profits that the proposed method is able to make [33, 51]. In this type of evaluation a simulation is prepared and the algorithm buys when it predicts that the price will rise. Some simulations allow profiting from short selling i.e. to sell a stock not owned and buy it later to profit from the difference when the algorithm predicts that the stock price will fall.

## 5.3 Experimental Setup

The tweets were collected from March 23, 2017 to July 3, 2017 using the Twitter API *tweepy*[2]. Only the tweets that mention a single company were kept, this operation reduced the corpus to approximately 25% of its original size. The rationale for this filter is that most of the tweets with more than one stock symbol in the content are either advertisements or contain a relationship between such the companies that is outside of the scope of this work. By filtering out such tweets they can not be used for different forecasts. After the filter 141'007 tweets remained.

The stock symbols were chosen based on their popularity and are shown on Table 5.1[3]. Other stock symbols were considered but not enough tweets were collected (less than 100). The distribution of the tweets[4] in the corpus is shown on Figure 5.1. The average of daily tweets is shown in Table 5.2. Below a sample of the tweets is shown.

> *"$AMZN Amazon launches store-pick grocery service in Seattle <url>"*
>
> *"Signal is Positive upward for Apple! $AAPL #AAPL #stocks #DayTrade #AI"*
>
> *"#Facebook Messenger Reaches 1.2B Monthly Active Users Milestone. Read more: <url>$FB"*
>
> *"Insider Trading Activity Alphabet Inc (NASDAQ :GOOG) Director Sold 24 shares of Stock <url>$GOOG"*

[2] `http://docs.tweepy.org/en/v3.5.0/`

[3] On June 19, 2017, Yahoo changed its stock symbol to AABA and although some tweets started to adopt the new symbol, YHOO was still used in Twitter.

[4] There are less tweets at the beginning of the figure because at first only the trading hours were considered. This fact does not affect the experiments since only the days where there are tweets available were taken into account.

| Stock symbol | Company |
|---|---|
| NASDAQ:AMZN | Amazon.com, Inc. |
| NASDAQ:AAPL | Apple Inc. |
| NASDAQ:FB | Facebook Inc. |
| NASDAQ:GOOG | Alphabet Inc. |
| NASDAQ:MSFT | Microsoft Corporation |
| NYSE:SNAP | Snap Inc. |
| NYSE:TWTR | Twitter Inc. |
| NASDAQ:AABA | Yahoo Inc. |
| NASDAQ:ZNGA | Zynga Inc. |

Table 5.1: IT companies and their stock symbol

The daily stock prices at open and close market times and the volume were collected from *Google Finance*.[5]. Although some stocks are from the NYSE and other from the NASDAQ stock exchanges markets, for the purposes in this work there is no difference in the companies they handle.

## 5.4 Correlation and Association Measures between Tweet Volume and Financial Indicators

It was tested a lagged correlation between the daily number of tweets per company against the company's open price, close price, the difference between close and open prices, the absolute value of such difference, transaction volume, returns

$$Ret_t = \left( \frac{p_t - p_{t-1}}{p_{t-1}} \right), \tag{5.1}$$

and logarithmic returns

$$LogRet_t = \log \left( \frac{p_t}{p_{t-1}} \right), \tag{5.2}$$

where $p_t$ is the price at close market time $t$ and $p_{t-1}$ is the price at open market time $t-1$.

[5] `http://www.google.com/finance`

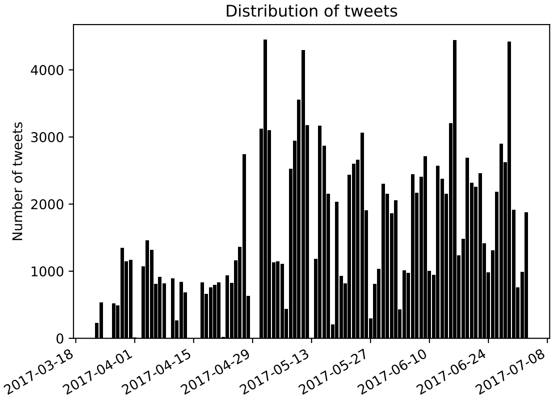


Figure 5.1: Daily amount of tweets in the corpus

The time series of the tweets is created using the days in which at least one tweet was collected. The days with zero tweets are not considered as point in the correlation even if it is a working day, this is because the download process for each company was started at different times.

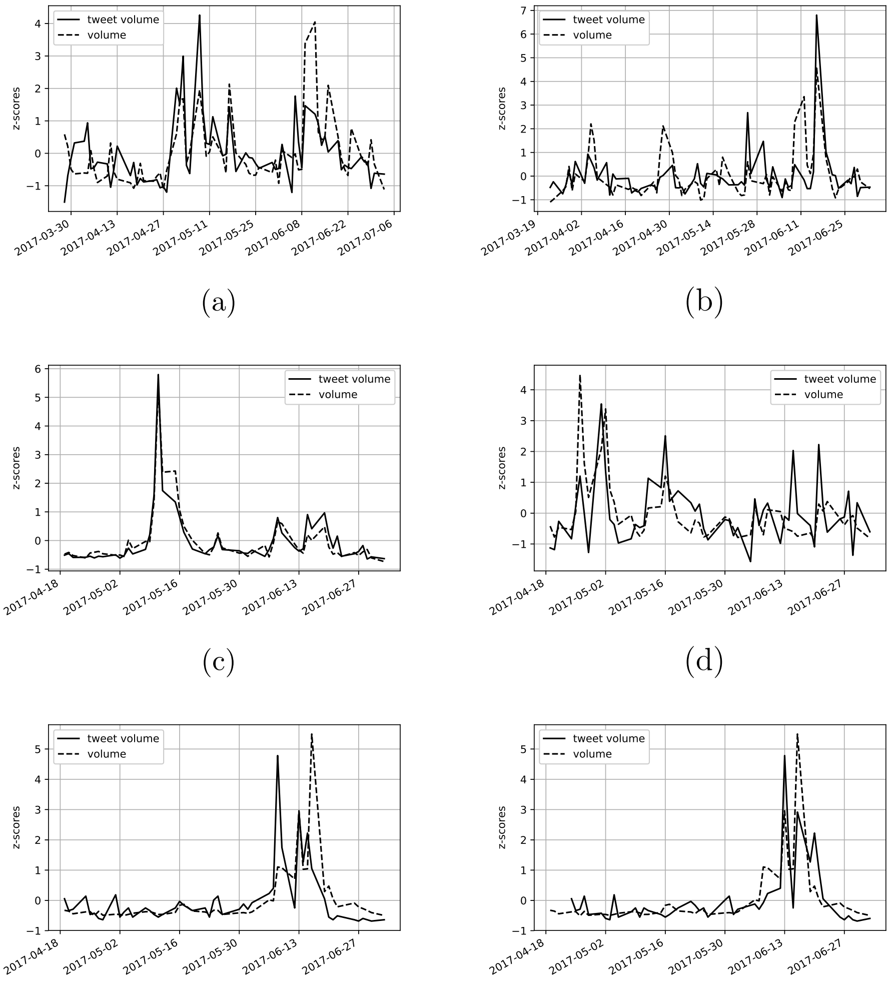
In Table 5.3 the results for $|\rho| > 0.5$ are shown. The Pearson's correlation coefficient is denoted as $\rho$. Although all the indicators mentioned above were compared, volume was the indicator that appeared the most. The column *Days* indicates the total number of points of the time series, for each lag a point is lost. High correlation with a lag 1 indicates that the number of tweets may cause the price or transaction volume fluctuation of the next day. High correlation with lag $-1$ indicates that the price or transaction volume may cause the next day's tweet volume. On Table 5.4 it is shown that the average of the correlation with different variables over all companies.

The time series shown in Figure 5.2 and Figure 5.3 are z-normalized using

$$z(x_i) = \frac{x_i - \mu(X)}{\sigma(X)} \tag{5.3}$$

where $x_i \in X$ are the points in the time series $X$, $\mu(X)$ and $\sigma(X)$ are the mean and standard deviation of $X$ respectively.

In Table 5.3 and Table 5.4 it can be noted that tweet volume and trading volume showed more correlation at lag 0. This is consistent with Ruiz et al., the greatest correlation on the table on average was $\rho = 0.4728$ at lag 0 between tweets and trading volume. In the work of Ruiz they considered the



(a) (b)

(c) (d)

(e) (f)

Figure 5.2: Correlation between **(a)** Apple's tweet volume and transaction volume $lag = 0$, $\rho = 0.61$; **(b)** Amazon's tweet volume and transaction volume $lag = 0$, $\rho = 0.64$; **(c)** Snapchat's tweet volume and transaction volume $lag = 0$, $\rho = 0.96$; **(d)** Twitter's tweet volume and transaction volume $lag = 0$, $\rho = 0.51$; **(e)** Yahoo's tweet volume and transaction volume $lag = 0$, $\rho = 0.60$; **(f)** Yahoo's tweet volume and transaction volume $lag = 3$, $\rho = 0.77$. For all figures **(a)**–**(f)** the volume is measured during trading hours

| Symbol | During<br>trading hours | Outside<br>trading hours | Weekends<br>and holidays |
|---|---|---|---|
| amzn | 151.72 | 125.55 | 87.27 |
| aapl | 163.52 | 167.0 | 103.17 |
| fb | 161.31 | 220.97 | 238.48 |
| goog | 56.97 | 72.39 | 60.36 |
| msft | 110.7 | 130.77 | 118.9 |
| snap | 110.49 | 89.21 | 32.05 |
| twtr | 143.47 | 134.79 | 104.28 |
| yhoo | 15.79 | 14.05 | 11.47 |
| znga | 12.12 | 10.65 | 9.65 |

Table 5.2: Average of daily tweets during and outside trading hours

total number of tweets during the day, the daily change in close prices and daily traded volume. With the data collected in this chapter similar results were achieved using the total number of tweets during the day, this can be seen in Table 5.6 and Table 5.7.

Another interesting observation is the case of the company Snapchat, the result indicates that the volume of tweets the night before and the morning before are highly correlated with the transaction volume during the following trading hours.

The higher results of LTAM are shown on Table 5.5 and Table 5.7. In Figure 5.3 are depicted some of the associations found.

Using the LTAM negative associations, not shown by the correlation analysis, were found.

## 5.5 Relationship between the Content of Tweets and Price Trends

The task of tweet classification is done by automatically tagging the tweets according to the price trend in which they were generated, upward, downward or neutral. A machine learning classifier is trained on some portion of the tweets randomly selected to predict the tag of the remaining tweets.

Tweet and stock time series are labeled with different time zones. It was

| Trading time | Symbol | Indicator | Correlation | Lag | Days |
|---|---|---|---|---|---|
| Before | amzn | Volume | 0.6526 | 0 | 45 |
| | fb | Open | 0.5170 | -1 | 44 |
| | fb | Open | 0.5509 | 0 | 45 |
| | fb | Close | 0.5140 | -1 | 44 |
| | fb | Close | 0.5210 | 0 | 45 |
| | snap | Volume | 0.9141 | 0 | 40 |
| | snap | Volume | 0.5546 | 1 | 39 |
| During | amzn | Volume | 0.6375 | 0 | 70 |
| | aapl | Volume | 0.6056 | 0 | 67 |
| | snap | Volume | 0.5408 | -1 | 51 |
| | snap | Volume | 0.9567 | 0 | 52 |
| | snap | Volume | 0.6765 | 1 | 51 |
| | twtr | Volume | 0.5147 | 0 | 52 |
| | yhoo | Volume | 0.5955 | 0 | 50 |
| | yhoo | Volume | 0.5603 | 1 | 49 |
| | yhoo | Volume | 0.7739 | 3 | 47 |
| After | snap | Volume | 0.6846 | 0 | 52 |
| | snap | Volume | 0.9175 | 1 | 51 |
| | snap | Volume | 0.5748 | 2 | 50 |
| | znga | Volume | 0.5229 | 0 | 54 |

Table 5.3: Higher results for lagged correlation between tweet volume and financial indicators using a lag from $-3$ to 3

| Lag | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| $Open$ | 0.1384 | 0.1459 | 0.1602 | 0.1711 | 0.1860 | 0.1871 | 0.2069 |
| $Close$ | 0.1433 | 0.1515 | 0.1561 | 0.1947 | 0.1937 | 0.2035 | 0.1927 |
| $Close - Open$ | 0.0228 | 0.0256 | -0.0286 | 0.0800 | 0.0033 | 0.0339 | -0.0716 |
| $\|Close - Open\|$ | 0.0230 | 0.1360 | 0.0642 | 0.2332 | 0.1344 | 0.0799 | 0.0818 |
| $Volume$ | 0.1206 | 0.1175 | 0.2064 | 0.4728 | 0.3142 | 0.1653 | 0.1503 |
| $Ret_t$ | 0.0230 | 0.0307 | -0.0286 | 0.0829 | 0.0076 | 0.0336 | -0.0749 |
| $LogRet_t$ | 0.0225 | 0.0291 | -0.0289 | 0.0810 | 0.0065 | 0.0334 | -0.0752 |

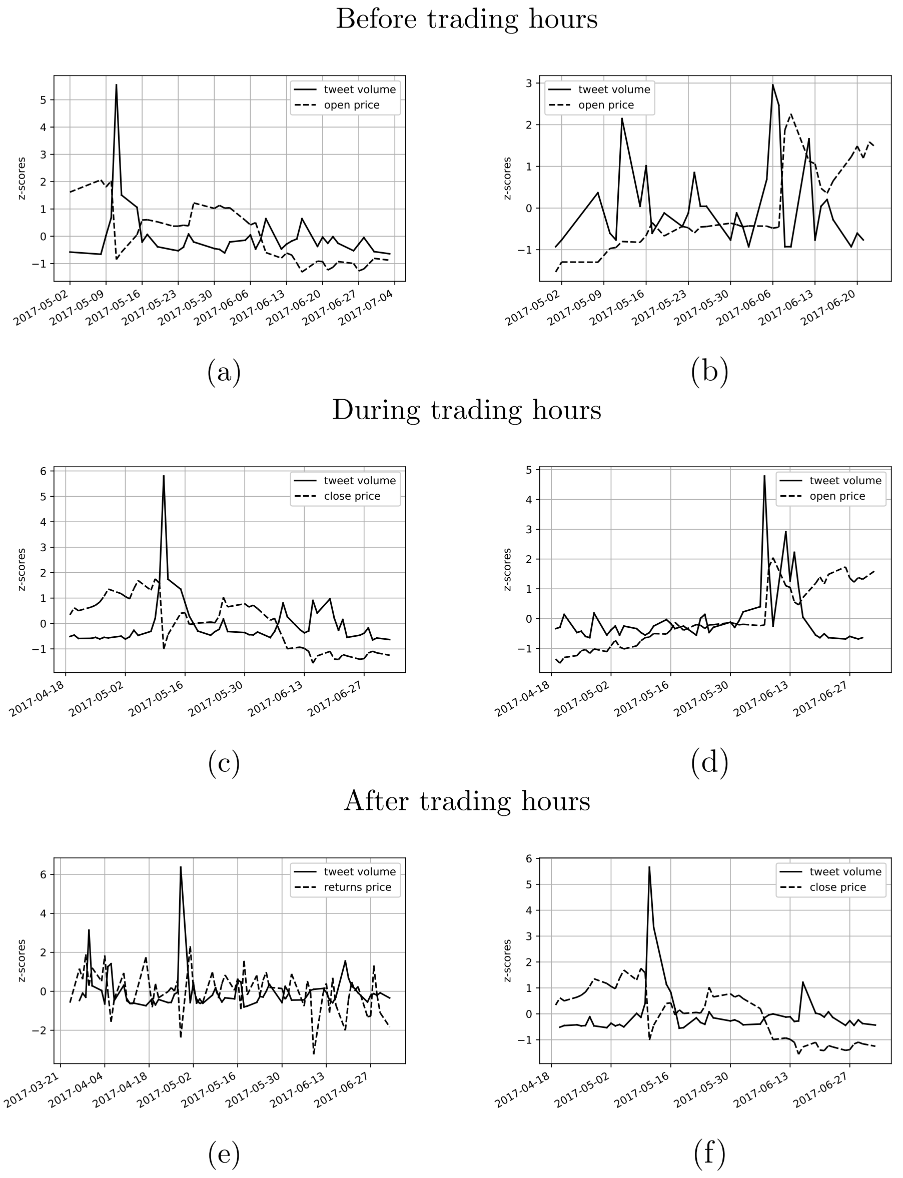
Table 5.4: Average correlation of tweet volume with different indicators during trading hours

| Trading time | Symbol | Indicator | LTAM | Lag | Days |
|---|---|---|---|---|---|
| Before | amzn | Volume | 0.6856 | 0 | 45 |
| | snap | Open | 0.5056 | -1 | 39 |
| | snap | Open | -0.7361 | 0 | 40 |
| | snap | Close | -0.6812 | 0 | 40 |
| | snap | Volume | 0.8669 | 0 | 40 |
| | yhoo | Open | -0.5773 | -2 | 32 |
| During | amzn | Volume | 0.6741 | 0 | 70 |
| | snap | Open | -0.6019 | 0 | 52 |
| | snap | Close | -0.6086 | 0 | 52 |
| | snap | Volume | 0.9402 | 0 | 52 |
| | yhoo | Open | -0.5669 | -1 | 49 |
| | yhoo | Volume | 0.7210 | 3 | 47 |
| After | amzn | Close-Open | -0.5344 | 1 | 67 |
| | amzn | Returns | -0.5468 | 1 | 67 |
| | amzn | Log Returns | -0.5465 | 1 | 67 |
| | snap | Open | -0.7599 | 1 | 51 |
| | snap | Close | -0.7487 | 1 | 51 |
| | snap | Volume | 0.7715 | 1 | 51 |

Table 5.5: Higher results for lagged LTAM between tweet volume and financial indicators using a lag from $-3$ to 3

| Symbol | Indicator | Correlation | Lag | Days |
|---|---|---|---|---|
| amzn | volume | 0.6554 | 0 | 71 |
| aapl | volume | 0.5239 | 0 | 69 |
| fb | open | 0.5485 | -3 | 68 |
| fb | open | 0.5962 | -2 | 69 |
| fb | open | 0.6526 | -1 | 70 |
| fb | open | 0.6609 | 0 | 71 |
| fb | open | 0.5744 | 1 | 70 |
| fb | open | 0.5787 | 2 | 69 |
| fb | open | 0.5568 | 3 | 68 |
| fb | close | 0.5755 | -3 | 68 |
| fb | close | 0.6292 | -2 | 69 |
| fb | close | 0.6544 | -1 | 70 |
| fb | close | 0.6269 | 0 | 71 |
| fb | close | 0.5945 | 1 | 70 |
| fb | close | 0.5749 | 2 | 69 |
| fb | close | 0.5394 | 3 | 68 |
| snap | volume | 0.9281 | 0 | 53 |
| snap | volume | 0.7883 | 1 | 52 |
| snap | volume | 0.5511 | 2 | 51 |
| yhoo | volume | 0.5365 | 0 | 53 |
| yhoo | volume | 0.5064 | 1 | 52 |
| yhoo | volume | 0.5466 | 2 | 51 |
| yhoo | volume | 0.6645 | 3 | 50 |
| znga | volume | 0.6382 | 0 | 64 |

Table 5.6: Higher results for lagged correlation between tweet volume and financial indicators using a lag from $-3$ to 3 (All daily tweets)


Before trading hours

(a)

(b)

During trading hours

(c)

(d)

After trading hours

(e)

(f)

Figure 5.3: Associations between **(a)** Snapchat's tweet volume and open price $lag = 0$, $\rho = -0.17$, $LTAM = -0.74$; **(b)** Yahoo's tweet volume and open price $lag = -2$, $\rho = -0.22$, $LTAM = -0.58$; **(c)** Snapchat's tweet volume and close price $lag = 0$, $\rho = -0.19$, $LTAM = -0.61$; **(d)** Yahoo's tweet volume and open price $lag = -1$, $\rho = 0.14$, $LTAM = -0.57$; **(e)** Amazon's tweet volume and returns $lag = 1$, $\rho = -0.34$, $LTAM = -0.55$; **(f)** Snapchat's tweet volume and close price $lag = 1$, $\rho = -0.19$, $LTAM = -0.75$

| Symbol | Indicator | LTAM | Lag | Days |
|---|---|---|---|---|
| amzn | volume | 0.6595 | 0 | 71 |
| snap | close | 0.5022 | -1 | 52 |
| snap | volume | 0.8791 | 0 | 53 |
| yhoo | open | -0.5240 | -1 | 52 |
| yhoo | close | 0.5460 | 0 | 53 |
| yhoo | volume | 0.5212 | 3 | 50 |
| znga | volume | 0.5064 | 0 | 64 |

Table 5.7: Higher results for lagged LTAM between tweet volume and financial indicators using a lag from $-3$ to 3 (All daily tweets)

necessary to adjust the tweets UTC time-stamp to match financial time series ET time zone. NYSE and NASDAQ markets open at 9:30 ET and close at 16:00 ET.

Two thresholds were used, 0.3 (Leung's tagging) and 0.5 (Wütrich's tagging), for an automated binning procedure to tag the tweets in the following manner: If the price change is greater or equal than the threshold then the tweet is considered positive, if the price change falls beyond the threshold the tweet is considered negative, otherwise the tweet is considered neutral. This type of tagging for the prices was used in [5] and [15]. The distribution of the different tagging schemes is depicted in Table 5.8.

It was measured the price change with Eq. 5.1. For tweets generated within trading hours $p_t$ is the close price and $p_{t-1}$ is the open price of the tweet date. For tweets generated outside trading hours $p_t$ is the next open price and $p_{t-1}$ is the previous close price.

The preprocessing steps were:

1. Lowercase the tweet.

2. Replace links by the word *url*.

3. Replace usernames by the word *username*.

4. Remove the stock symbol.

5. Replace some non-ambiguous english contractions e.g. "'m" with "am".

6. Tokenization.

| | | Wütrich's threshold = 0.5 | | | Leung's threshold = 0.3 | | |
|---|---|---|---|---|---|---|---|
| company | tweets | positive(%) | neutral(%) | negative(%) | positive(%) | neutral(%) | negative(%) |
| amzn | 21606 | 30.38 | 46.24 | 23.36 | 46.58 | 27.29 | 26.11 |
| aapl | 25179 | 27.78 | 49.17 | 23.04 | 33.44 | 37.53 | 29.01 |
| fb | 32627 | 20.82 | 61.31 | 17.85 | 32.16 | 45.52 | 22.30 |
| goog | 10513 | 23.36 | 57.41 | 19.22 | 33.97 | 41.08 | 24.94 |
| msft | 19523 | 19.24 | 63.61 | 17.14 | 32.14 | 43.41 | 24.44 |
| snap | 11193 | 48.88 | 13.71 | 37.39 | 50.46 | 10.13 | 39.40 |
| twtr | 16938 | 34.10 | 43.48 | 22.41 | 45.74 | 27.44 | 26.80 |
| yhoo | 1777 | 35.05 | 49.57 | 15.36 | 46.82 | 29.31 | 23.86 |
| znga | 1651 | 50.33 | 32.76 | 16.89 | 52.39 | 28.64 | 18.95 |

Table 5.8: Tag balance using different offsets

7. Remove stopwords.

8. Remove tokens that are a number.

9. Remove tokens that do not contain letters.

10. Remove continuous repeated letters in each token if there are more than three, e.g. “haaaaaappy” is replaced by “happy”.

In the BOW approach different vocabulary sizes were attempted, a size of 1000, with the most common words gave the best results.

In the WE approach each word was transformed to their *word2vec* representation, i.e. a 300-dimensional vector. Tweets in which none of its words were found on the *word2vec* model are discarded. Then all retrieved vectors are averaged to obtain a single 300-dimensional vector that represents the tweet. The process of averaging all words returns a vector that is semantically similar to the tweet vectors [11].

Tweets are split randomly using 70% for train and the rest for test. Different classifiers were tested, however the ones with greater accuracy are shown on Tables 5.9 to 5.12.

Majority vote (MV) was set as the baseline, which consists on predicting always the class with more examples on the training set. Also implemented a voting classifier (VC) was implemented, which votes among the classifier predictions.

Dickinson [37] and Pagolu [38] compare both word representations on the sentiment classification task of stock tweets, they manually tagged tweets from the time series, 1000 and 3216 tweets respectively. Dickinson obtained an accuracy of 68.5% for n-gram model and 63.4% for WE model, Pagolu obtained 70.5% and 70.2%. Comparing with these works, the proposed task in this work is to classify the tweet according to the trend in which they were generated. For this case, the WE approach gave better results than BOW.

Other classifiers such as Naive Bayes, k-nearest-neighbors, logistic regression or multilayer perceptron were tested however the top three classifiers were selected based on the accuracy results. However the best classifiers were: Support Vector Classifier (SVC), Decision Tree (DT), Random Forest (RF) and Stochastic Gradient Descent Classifier (SGDC).

Once the best classifiers were selected, their parameters were optimized. The best parameters are shown next.

The parameters of the classifiers for the BOW representation were:

| | MV | SVC | DT | RF | VC |
|---|---|---|---|---|---|
| amzn | 46.3% | 46.3 | 45.9%** | 45.9%** | 46.3% |
| aapl | 37.9% | **38.3%**** | 37.8 | 36.6%** | **38.2%** |
| fb | 45.9% | 45.9 | **46.3%**** | 45.6%* | 45.8% |
| goog | 40.4% | **41.0%**** | 33.8%** | 39.1% | **40.7%** |
| msft | 43.4% | **43.5%** | 43.3% | 43.4% | 43.3% |
| snap | 50.5% | 50.5% | 50.1% | 50.0% | **50.6%** |
| twtr | 45.5% | 45.4% | 45.3% | 44.8%** | 44.8%** |
| yhoo | 44.8% | 44.8% | 44.8% | 43.9% | 44.8% |
| znga | 54.3% | 54.3% | 52.5% | 52.5% | **54.5%** |
| average | 45.4% | **45.5%** | 44.4% | 44.6% | 45.4% |

Table 5.9: Results of Leung's tagging and BOW representation (threshold = 0.3)

| | MV | SVC | DT | RF | VC |
|---|---|---|---|---|---|
| amzn | 46.2% | **46.5%*** | 45.9% | 46.0% | 46.2% |
| aapl | 49.6% | **49.7%** | 48.9%** | 49.4%* | 49.6% |
| fb | 60.8% | 60.8% | 60.6% | 60.8% | 60.8% |
| goog | 56.9% | **57.1%** | 55.3%** | 56.5% | 56.9% |
| msft | 63.9% | 63.9% | 61.6%** | 63.8% | 63.9% |
| snap | 48.9% | 48.9% | 48.5% | 48.6% | 48.8% |
| twtr | 44.6% | **45.0%**** | 33.2%** | 40.5%* | 43.2%** |
| yhoo | 52.2% | 52.2% | 50.5% | 50.5% | 50.8%* |
| znga | 52.3% | 52.3% | 52.3% | 52.1% | 52.3% |
| average | 52.8% | **52.9%** | 50.7% | 52.0% | 52.5% |

Table 5.10: Results of Wütrich's tagging BOW representation (threshold = 0.5)

| | MV | SVC | SGDC | RF | VC |
|---|---|---|---|---|---|
| amzn | 46.3% | **59.8%**** | **46.7%** | **52.8%**** | **56.9%**** |
| aapl | 37.9% | **55.9%**** | **41.3%**** | **50.8%**** | **53.4%**** |
| fb | 45.9% | **62.9%**** | **49.3%**** | **56.9%**** | **60.3%**** |
| goog | 40.4% | **60.3%**** | **42.9%*** | **57.2%**** | **59.1%**** |
| msft | 43.4% | **56.6%**** | **44.3%** | **51.5%**** | **54.2%**** |
| snap | 50.5% | **60.2%**** | **52.9%**** | **55.8%**** | **59.5%**** |
| twtr | 45.5% | **56.4%**** | **45.6%** | **52.4%**** | **55.6%**** |
| yhoo | 44.8% | **61.2%**** | **47.8%** | **57.2%**** | **59.1%**** |
| znga | 54.3% | **65.5%**** | **58.8%** | **64.2%**** | **64.4%**** |
| average | 45.4% | **59.8%** | **47.7%** | **55.4%** | **58.0%** |

Table 5.11: Results of Leung's tagging and WE representation (threshold = 0.3)

| | MV | SVC | SGDC | RF | VC |
|---|---|---|---|---|---|
| amzn | 46.2% | **59.7%**** | 29.9%** | **53.9%**** | **54.6%**** |
| aapl | 49.6% | **60.1%**** | 49.5%** | **56.3%**** | **56.3%**** |
| fb | 60.8% | **69.6%**** | 22.6%** | **65.4%**** | **66.9%**** |
| goog | 56.9% | **69.2%**** | **57.4%**** | **65.4%**** | **67.3%**** |
| msft | 63.9% | **67.4%**** | 63.3%* | **65.2%**** | **64.9%**** |
| snap | 48.9% | **59.6%**** | 48.9%** | **55.7%**** | **56.3%**** |
| twtr | 44.6% | **58.2%**** | 44.6%** | **53.0%**** | **55.8%**** |
| yhoo | 52.2% | **64.2%**** | 52.0% | **63.2%**** | **65.1%**** |
| znga | 52.3% | **65.3%**** | 40.4%** | **64.6%**** | **61.4%**** |
| average | 52.8% | **63.7%** | 45.4% | **60.3%** | **60.9%** |

Table 5.12: Results of Wütrich tagging and WE representation (threshold = 0.5)

- **SVC:** RBF kernel, $\gamma = 10$ and $C = 10$.
- **DT:** maximum depth of tree = 10.
- **RF:** Trees = 10, maximum depth of tree = 10.

The parameters of the classifiers for the WE representation were:

- **SVC:** RBF kernel, $\gamma = 2$ and $C = 1$.
- **SGDC:** Hinge as loss function, L2 penalty, 5 epochs and optimal learning rate.
- **RF:** Trees = 10, maximum depth of tree = 5.

Comparing the results from Table 5.11 and Table 5.12 it can be seen that the threshold of 0.5 gave better results, this may be because it gives a less balanced set. However using the threshold of 0.3 two out of the three classifiers are also able to learn and surpass the threshold. The amount of data did not play a significant role. The companies YHOO and ZNGA with less tweets performed better than AMZN or AAPL with more tweets. Also in Table 5.11 the difference between baseline and SVC with companies FB (32627 tweets) and YHOO (1777 tweets) is similar.

In Tables 5.9 and 5.10 the accuracy results using the BOW representation mostly fall under the baseline. However in Tables 5.11 and 5.12, using the WE representation, the classifiers are able to surpass the threshold and therefore to identify the tweets during different trends, positive, negative and neutral. For the four tables two different tagging schemes were used.

To test the statistical significance of the accuracy results with respect to baseline the McNemar's test described on Section 2.5 was used.

On Table 5.12 considering Leung's tagging for the stock $twtr it is observed that SGD and the baseline have the same accuracy but the difference is statistically significant. This can be explained because although the same number of sample were classified correctly, the samples were not the same ones. As expected, companies with more tweets showed the higher statistical significance. In Tables 5.9 to 5.12 the p-values smaller than 0.01 are marked with ** and p-values smaller than 0.05 are marked with *. The accuracy results greater than the baseline are in bold.

## 5.6 Conclusion

The works in the literature were briefly discussed. and also the principal variables to consider in stock market forecasting, the common textual representations, the market level, the sources of information and the performance metrics.

It was shown that the tweet messages generated during a positive, negative and neutral trends of stock prices can be distinguished by state of the art classifiers. This means that the topics on social media depend on the price change of a company's stock. The results were measured considering how the balance of the dataset may influence the results.

The results obtained using the word embeddings representation of tweets significantly outperform the results obtained by bag of words. This may be caused by the significant reduction in the dimensionality.

For more than half of the considered companies, moderate correlation was found between tweet volume and trading volume, which is consistent with the findings in [50] and with the Efficient Market Hypothesis [1]. It was shown that for two of the companies tweet volume can be used for prediction of transaction volume with lag of one or three days. Regarding the consideration of the trading hours it was found for one company that the volume of tweets generated the day before after the close time is correlated with the volume during trading hours the next day.

The MAT and the LTAM were used to find associations between the tweet volume different indicators. Inverse relationships that were not discovered by correlation analysis were found. For this reason, LTAM can be used as complementary measure to the correlation coefficient.

# Chapter 6

# Conclusions

In this work the main object of analysis was the financial time series. In Chapters 3 and 4 it was explored the relationship between financial time series using new methods. In Chapter 5 it was studied the relationship between the financial indicators and social media. Specifically the conclusions are:

- New association measures were proposed that consider the time series trends.
- An association measure that considers the dynamics of financial time series was also proposed.
- There is a relationship (association and correlation) between volume of tweets and transaction volume
- The tweets created during different trends can be distinguished by automatic classifiers
- In this task WE representation was better than BOW representation

For all the experiments and the data collection mentioned above software modules were developed.

## 6.1 Contributions

- **Window selection.** It was proposed a method to select the window size, highlighting the most notable time series associations. The problem of selecting an appropriate window has not been explored before.
- **Pattern similarity.** A method to measure similarity based on the size of association of time series patterns, this method considers time series

to be more similar when they hold an association together continuously. The patterns of the time series were displayed graphically and based on those patterns the companies examined using the time series of the stock price show similarity that is congruent geographically for the case of Oil&Gas companies and by their specialization (hardware or software) for the case of IT companies.

- **Dynamic visualization.** In Chapter 4 it was shown why the correlation coefficient may be misleading in some cases and it was also presented a method to dynamically visualize how two time series associate in time.

- **Relation between dynamic association and financial events.** It was also explored the relationship of these associations with events that may be related to the behavior of the stock price of the companies.

- **Corpus.** A corpus of stock related tweets of IT companies for a period of three months was collected for a period of over three months.

- **Correlation between financial indicator and tweet volume.** From all the financial indicators tested against the tweet volume, the transaction volume showed greater correlation.

- **Association between financial indicator and tweet volume.** Using a measure of association from previous chapters it was also found a relationship between tweets volume and other indicators, especially negative relationship was shown using this measure.

- **Comparison between textual representations.** For this task the results using the WE representation was higher than using the BOW representation.

- **Comparison between machine learning classifiers with respect to a baseline.** The classifiers that performed better in this task were Support Vector Classifier, Stochastic Gradient Descent Classifier, and Random Forests.

One of the research questions was regarding the predictability power of social media on the stock market. In the lasts experiment performed, the textual data was aligned with the next day's stock price trends without good results so far. This experiments will be left for future work as it is described in Section 6.3.

## 6.2 Publications

### Published

1. Francisco Javier Garcia-Lopez, Ildar Batyrshin, Alexander Gelbukh, "Similarity of time series based on the length of the patterns of the moving approximation transform", Research in Computing Science, vol. 115, pp. 79–92, 2016.

2. Francisco Javier Garcia-Lopez, Ildar Batyrshin, Alexander Gelbukh, "Dynamic Local Trend Associations in Analysis of Comovements of Financial Time Series", North American Fuzzy Information Processing Society Annual Conference, Springer, pp. 181–188, 2017.

### Accepted

1. Francisco Javier Garcia-Lopez, Ildar Batyrshin, Alexander Gelbukh, "Analysis of relationships between tweets and stock market trends", Journal of Intelligent & Fuzzy Systems. Special issue on LKE 2017: 5th International Symposium on Language & Knowledge Engineering, Puebla, Mexico. November 22–24, 2017.

## 6.3 Future Work

As for the future work, different techniques on the textual representations will be used, such as dimensionality reduction for the BOW model and other types of aggregation will be attempted for the WE vectors. Also, the textual approaches reviewed in Chapter 5 will be combined with time series forecasting techniques such as regression or ARIMA, this combination has shown good results in the literature [55, 60]. The corpus size will be increased and other source of data will be added, the news, this is because most of the work uses only one source of data. Other methods to aggregate information from different sources, such as artificial neural networks will also be explored.

# Glossary

**Balanced dataset:** In machine learning a dataset is balanced if the number of samples on each class is approximately the same.

**Forecast:**[1] Predict or estimate a future event or trend.

**Over fit:** In machine learning over fitting refers to the situation when the algorithm learns to classify the training set too well (even the noise and outliers) rather than learning a hypothesis that generalizes to samples beyond the training set. The characteristic of over fitting is that the classifier performs too well in the training set but badly on the test set [61].

**Predict:**[1] Say or estimate that (a specified thing) will happen in the future or will be a consequence of something.

**Returns:**[2] "A return is the gain or loss of a security in a particular period. The return consists of the income and the capital gains relative on an investment, and it is usually quoted as a percentage"

**Trading volume:**[2] "Volume is the number of shares or contracts traded in a security or an entire market during a given period of time."

[1] `https://en.oxforddictionaries.com`

[2] `http://www.investopedia.com/terms/`

# Appendix A

# Stock Symbols and Time Series

| Stock symbol | Oil and gas company |
|---|---|
| NYSE:BP | BP plc |
| NYSE:CVX | Chevron Corporation |
| OTCMKTS:OGZPY | Gazprom PAO |
| NYSE:PTR | PetroChina Company Limited |
| NYSE:RDS.A | Royal Dutch Shell plc |
| NYSE:XOM | Exxon Mobil Corp. |
| NYSE:PBR | Petroleo Brasileiro SA |
| OTCMKTS:LUKOY | NK LUKOIL PAO |
| Stock symbol | IT company |
| NASDAQ:AAPL | Apple Inc. |
| NASDAQ:AMZN | Amazon.com, Inc. |
| NYSE:IBM | International Business Machines Corp. |
| NASDAQ:INTC | Intel Corporation |
| OTCMKTS:LNVGY | Lenovo Group Ltd. |
| NASDAQ:MSFT | Microsoft Corp. |
| OTCMKTS:PCRFY | Panasonic Corp. |
| NYSE:PHG | Koninklijke Philips NV |
| NYSE:SNE | Sony Corp. |
| NASDAQ:FB | Facebook Inc. |
| NASDAQ:GOOGL | Alphabet Inc. |

Table A.1: Companies and their stock symbol

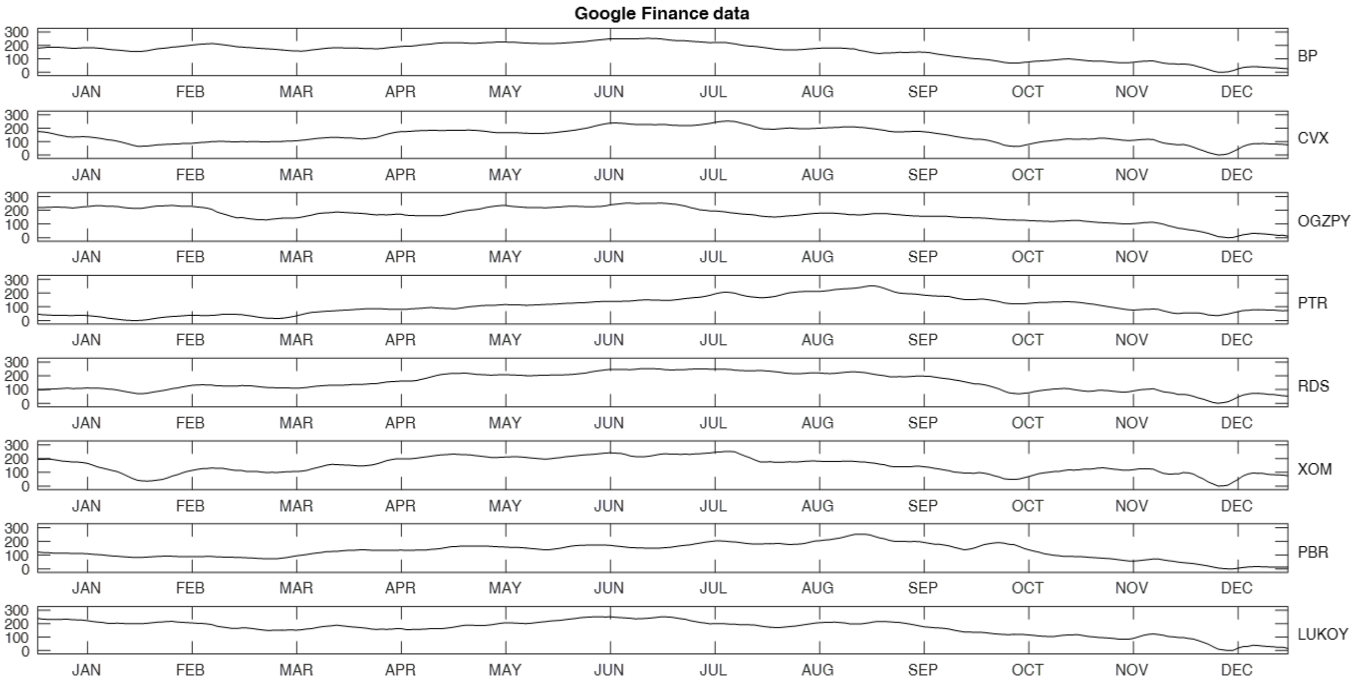


Figure A.1: Time series of daily close prices of oil companies in 2014

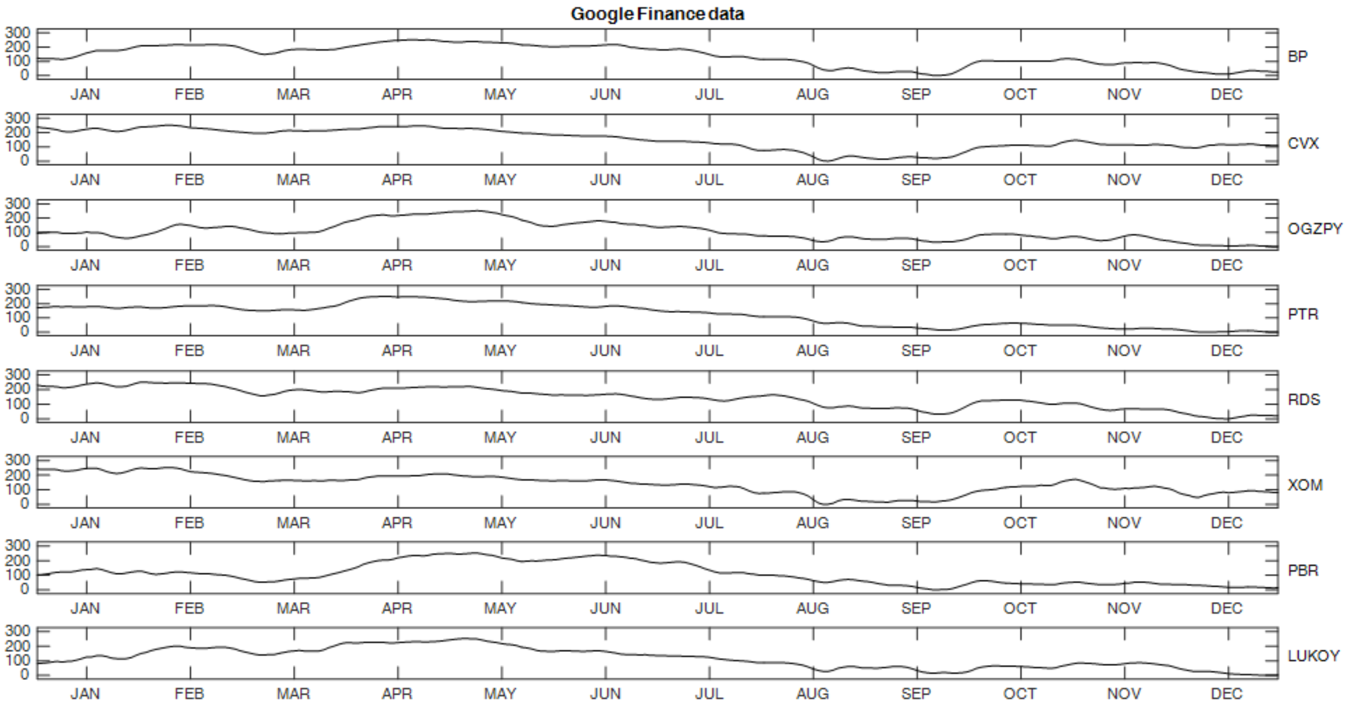


Figure A.2: Time series of daily close prices of oil companies in 2015

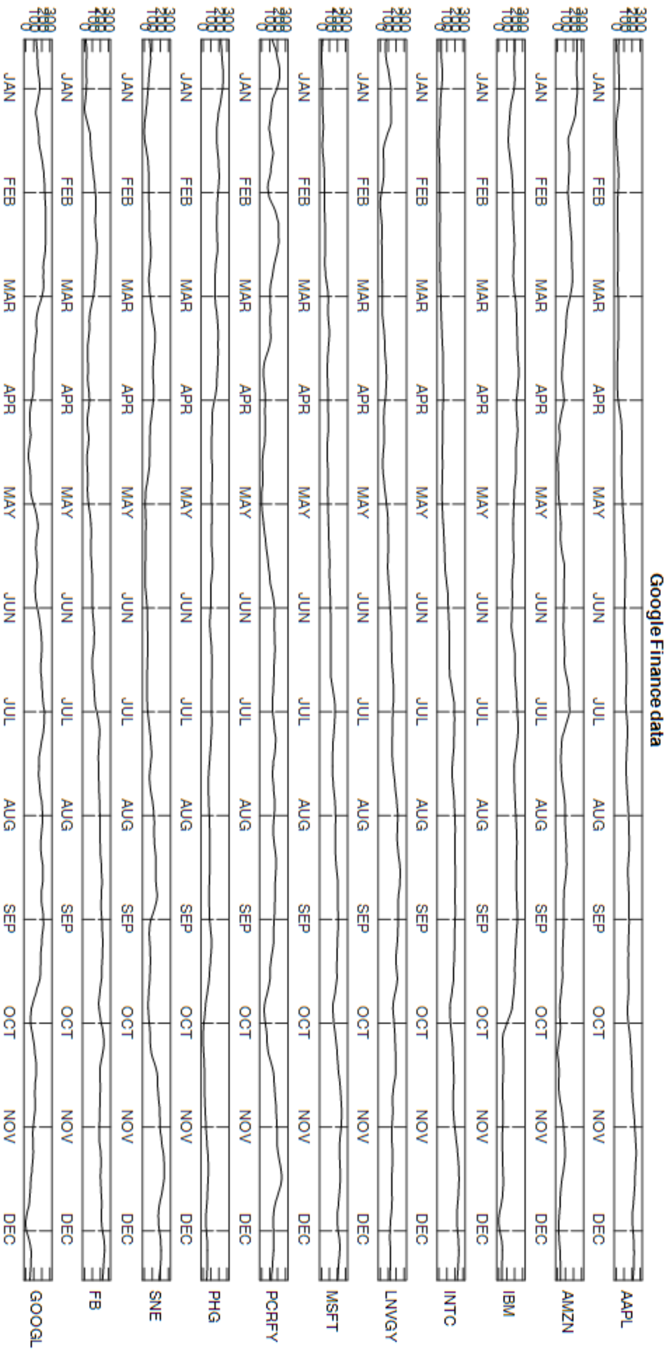


Figure A.3: Time series of daily close prices of IT companies in 2014

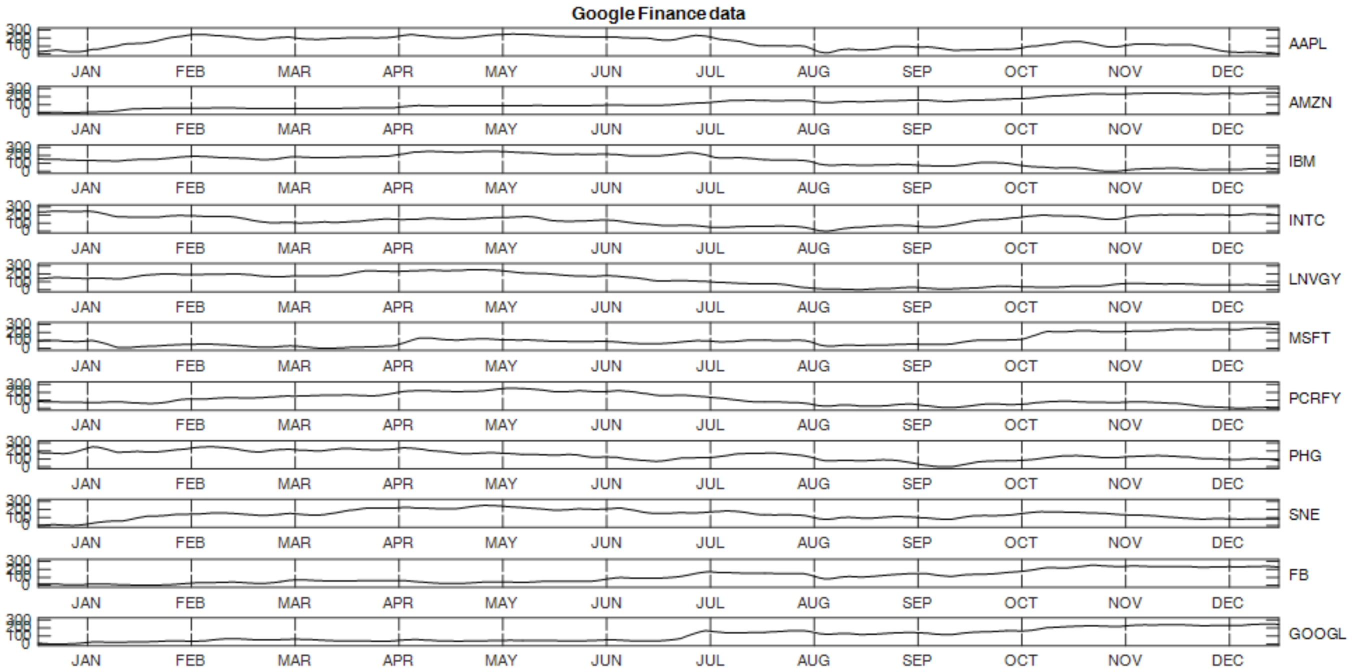


Figure A.4: Time series of daily close prices of IT companies in 2015

# Appendix B

# Tweet Classification

Tweet experiments using all classifiers. With these preliminary experiments the best classifiers were selected and their parameters were optimized. The dataset for these experiments was smaller, only the tweets from March 23, 2017 to May 08 2017.

| Method | Code |
|---|---|
| Baseline (Majority Vote) | MV |
| SVM (linear kernel) | SVl |
| SVM (RFB kernel) | SVr |
| SVM (RBF kernel, $\gamma = 2$) | SVg |
| Naive Bayes (Bernoulli Distribution) | NB_B |
| Logistic Regression | LR |
| Stochastic Gradient Descent | SGD |
| K Nearest Neighbors | KNN |
| Decision Tree | DT |
| Random Forest | RF |
| Multilayer Perceptron | MLP |
| Naive Bayes (Gaussian Distribution) | NB_G |
| Vote Classifier | VC |

| | MV | SV_lin | SV_rbf | SV_g | NB_B | LR | SGD | KNN | DT | RF | MLP | NB | VC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| amzn | 0.43 | 0.38 | 0.403 | 0.404 | 0.295 | 0.424 | 0.352 | 0.378 | 0.411 | 0.405 | 0.425 | 0.404 | 0.428 |
| aapl | 0.567 | 0.324 | 0.567 | **0.568** | 0.237 | 0.347 | 0.368 | 0.372 | 0.202 | 0.567 | 0.326 | 0.359 | 0.387 |
| fb | 0.586 | 0.444 | 0.586 | 0.585 | 0.365 | 0.521 | 0.516 | 0.3 | 0.584 | 0.586 | 0.531 | 0.538 | 0.554 |
| goog | 0.584 | 0.389 | 0.584 | 0.584 | 0.322 | 0.442 | 0.391 | 0.142 | 0.349 | **0.585** | 0.484 | 0.435 | 0.44 |
| msft | 0.646 | 0.456 | 0.646 | 0.646 | 0.424 | 0.566 | 0.443 | 0.625 | 0.638 | 0.646 | 0.533 | 0.624 | 0.628 |
| snap | 0.565 | 0.43 | 0.565 | 0.56 | 0.4 | 0.425 | 0.395 | 0.34 | 0.515 | 0.535 | 0.425 | 0.44 | 0.435 |
| twtr | 0.469 | 0.414 | 0.469 | 0.465 | 0.4 | **0.486** | 0.442 | 0.39 | 0.392 | **0.47** | **0.496** | 0.455 | **0.486** |
| yhoo | 0.571 | 0.403 | 0.571 | 0.571 | 0.494 | 0.455 | 0.442 | 0.104 | 0.494 | 0.558 | 0.481 | 0.545 | 0.545 |
| znga | 0.569 | 0.367 | 0.569 | 0.569 | 0.431 | 0.514 | 0.477 | 0.339 | 0.55 | 0.569 | 0.495 | 0.495 | 0.532 |

Table B.1: Wütrich tagging BOW

| | MV | SVl | SVr | SVg | NB_B | LR | SGD | KNN | DT | RF | MLP | NB | VC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| amzn | 0.43 | **0.509** | 0.43 | **0.556** | 0.414 | **0.503** | **0.512** | **0.545** | **0.504** | **0.489** | **0.47** | **0.438** | **0.526** |
| aapl | 0.567 | **0.612** | 0.567 | **0.645** | 0.551 | **0.618** | **0.607** | **0.645** | **0.601** | **0.585** | **0.594** | 0.557 | **0.618** |
| fb | 0.586 | **0.62** | 0.586 | **0.64** | 0.495 | **0.62** | **0.635** | **0.653** | **0.611** | **0.588** | **0.607** | 0.474 | **0.622** |
| goog | 0.584 | **0.631** | 0.584 | **0.645** | 0.576 | **0.628** | **0.635** | **0.672** | **0.634** | **0.622** | 0.584 | 0.542 | **0.629** |
| msft | 0.646 | **0.659** | 0.646 | **0.672** | 0.507 | **0.667** | 0.636 | **0.648** | **0.666** | **0.65** | **0.649** | 0.386 | **0.668** |
| snap | 0.565 | **0.585** | 0.565 | 0.555 | 0.505 | 0.565 | **0.6** | 0.46 | 0.525 | 0.56 | 0.54 | 0.525 | 0.555 |
| twtr | 0.469 | **0.519** | 0.469 | **0.545** | **0.498** | **0.515** | **0.498** | **0.471** | **0.491** | **0.518** | **0.488** | 0.395 | **0.521** |
| yhoo | 0.571 | 0.571 | 0.571 | **0.636** | **0.597** | 0.571 | **0.662** | **0.714** | **0.714** | **0.662** | **0.675** | **0.649** | **0.636** |
| znga | 0.569 | 0.569 | 0.569 | **0.596** | 0.532 | 0.569 | 0.477 | **0.587** | **0.606** | 0.532 | 0.56 | 0.56 | **0.587** |

Table B.2: Wütrich tagging WE

# Bibliography

[1] Burton G Malkiel. Reflections on the efficient market hypothesis: 30 years later. *Financial Review*, 40(1):1–9, 2005.

[2] Samuel Dupernex. Why might share prices follow a random walk. *Student Economic Review*, 21(1):167–179, 2007.

[3] Ildar Batyrshin, Raul Herrera-Avelar, Leonid Sheremetov, and Aleksandra Panova. Moving approximation transform and local trend associations in time series data bases. In *Perception-based Data Mining and Decision Making in Economics and Finance*, pages 55–83. Springer, 2007.

[4] Miguel A Sanchez-Perez, Ilia Markov, Helena Gómez-Adorno, and Grigori Sidorov. Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 145–151. Springer, 2017.

[5] Steven Kung Fan Leung. Automatic stock market: predictions from world wide web data. Master's thesis, The Hong Kong University of Science and Technology, 1997.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[7] Yao Zhang. *Using Financial Reports to Predict Stock Market Trends with Machine Learning Techniques*. PhD thesis, University of Oxford, 2015.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[9] XingYi Xu, HuiZhi Liang, and Timothy Baldwin. Unimelb at semeval-2016 tasks 4a and 4b: An ensemble of neural networks and a word2vec based model for sentiment classification. *Proceedings of SemEval*, pages 183–189, 2016.

[10] Andi Rexha, Mark Kröll, Mauro Dragoni, and Roman Kern. Polarity classification for target phrases in tweets: A word2vec approach. In *International Semantic Web Conference*, pages 217–223. Springer, 2016.

[11] Abhineshwar Tomar, Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. Towards twitter hashtag recommendation using distributed word representations and a deep feed forward neural network. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI 2014)*, pages 362–368. IEEE, 2014.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[14] Krishnan P Sankaran. *A system to forecast currency exchange rates.* PhD thesis, Hong Kong University of Science and Technology, 1996.

[15] Beat Wüthrich, D Permunetilleke, Steven Leung, W Lam, Vincent Cho, and J Zhang. Daily prediction of major stock indices from textual www data. *HKIE Transactions*, 5(3):151–156, 1998.

[16] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. On the importance of text analysis for stock price prediction. In *LREC*, pages 1170–1175, 2014.

[17] Philippe Esling and Carlos Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):12, 2012.

[18] Béla Bollobás, Gautam Das, Dimitrios Gunopulos, and Heikki Mannila. Time-series similarity problems and well-separated geometric sets. In *Proceedings of the thirteenth annual symposium on Computational geometry*, pages 454–456. ACM, 1997.

[19] Robert J Alcock, Yannis Manolopoulos, et al. Time-series similarity queries employing a feature-based approach. In *7th Hellenic conference on informatics*, pages 27–29, 1999.

[20] Jessica Lin, Rohan Khade, and Yuan Li. Rotation-invariant similarity in time series using bag-of-patterns representation. *Journal of Intelligent Information Systems*, 39(2):287–315, 2012.

[21] Ildar Batyrshin, Valery Solovyev, and Vladimir Ivanov. Time series shape association measures and local trend association patterns. *Neurocomputing*, 175:924–934, 2016.

[22] Jiaqi Ye, Chengwei Xiao, Rui Máximo Esteves, and Chunming Rong. Time series similarity evaluation based on spearman's correlation coefficients and distance measures. In *International Conference on Cloud Computing and Big Data in Asia*, pages 319–331. Springer, 2015.

[23] Rudi Schaefer, Nils Fredrik Nilsson, and Thomas Guhr. Power mapping with dynamical adjustment for improved portfolio optimization. *Quantitative Finance*, 10(1):107–119, 2010.

[24] George Filis, Stavros Degiannakis, and Christos Floros. Dynamic correlation between stock market and oil prices: The case of oil-importing and oil-exporting countries. *International Review of Financial Analysis*, 20(3):152–164, 2011.

[25] John Paparrizos and Luis Gravano. Fast and accurate time-series clustering. *ACM Transactions on Database Systems (TODS)*, 42(2):8, 2017.

[26] Yangtuo Peng and Hui Jiang. Leverage financial news to predict stock price movements using word embeddings and deep neural networks. *arXiv preprint arXiv:1506.07220*, 2015.

[27] Nicholas Barberis, Andrei Shleifer, and Jeffrey Wurgler. Comovement. *Journal of Financial Economics*, 75(2):283–317, 2005.

[28] Christophe Croux, Mario Forni, and Lucrezia Reichlin. A measure of comovement for economic variables: Theory and empirics. *Review of Economics and Statistics*, 83(2):232–241, 2001.

[29] Leo A Goodman. Tests based on the movements in and the comovements between m-dependent time series. Technical report, COLUMBIA UNIV NEW YORK, 1961.

[30] Leo A Goodman and Yehuda Grunfeld. Some nonparametric tests for comovements between time series. *Journal of the American Statistical Association*, 56(293):11–26, 1961.

[31] Spiros Papadimitriou, Jimeng Sun, and S Yu Philip. Local correlation tracking in time series. In *Sixth International Conference on Data Mining, 2006. ICDM'06.*, pages 456–465. IEEE, 2006.

[32] Robert S Pindyck and Julio J Rotemberg. The excess co-movement of commodity prices, 1988.

[33] Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen, and James Allan. Mining of concurrent text and time series. In *KDD-2000 Workshop on Text Mining*, pages 37–44, 2000.

[34] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[35] Ray Chen and Marius Lazer. Sentiment analysis of twitter feeds for the prediction of stock market movement. 2013. [Online]. Available: http://cs229.stanford.edu/proj2011/ChenLazer-SentimentAnalysisOf TwitterFeedsForThePredictionOfStockMarketMovement.pdf.

[36] Xue Zhang, Hauke Fuehres, and Peter A Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia-Social and Behavioral Sciences*, 26:55–62, 2011.

[37] Brian Dickinson and Wei Hu. Sentiment analysis of investor opinions on twitter. *Social Networking*, 4(03):62, 2015.

[38] Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, and Babita Majhi. Sentiment analysis of twitter data for predicting stock market movements. *arXiv preprint arXiv:1610.09225*, 2016.

[39] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670, 2014.

[40] Andrew W Lo, Harry Mamaysky, and Jiang Wang. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The journal of finance*, 55(4):1705–1765, 2000.

[41] Jeffrey S Abarbanell and Brian J Bushee. Fundamental analysis, future earnings, and stock prices. *Journal of Accounting Research*, 35(1):1–24, 1997.

[42] Hans-Dieter Klingemann, PP Mohler, and Robert Philip Weber. Das reichtumsthema in den thronreden des kaisers und die okonomische entwicklung in deutschland 1871-1914. *Computerunterstutzte Inhaltsanalyse in der empirischen Sozialforschung*, 1982.

[43] Viorel Milea. *News analytics for financial decision support*. Number EPS-2013-275-LIS in ERIM Ph.D. Series Research in Management. Erasmus University Rotterdam , Erasmus Research Institute of Management, 2013.

[44] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Using structured events to predict stock price movement: An empirical investigation. In *EMNLP*, pages 1415–1425, 2014.

[45] Chei Sian Lee and Long Ma. News sharing in social media: The effect of gratifications and prior experience. *Computers in Human Behavior*, 28(2):331–339, 2012.

[46] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[47] M Sebastian A Wolfram. Modelling the stock market using twitter. Master's thesis, University of Edinburgh, 2010.

[48] Xue Zhang, Hauke Fuehres, and Peter A Gloor. Predicting asset value through twitter buzz. In *Advances in Collective Intelligence 2011*, pages 23–34. Springer, 2012.

[49] Munmun De Choudhury, Hari Sundaram, Ajita John, and Dorée Duncan Seligmann. Can blog communication dynamics be correlated with stock market activity? In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 55–60. ACM, 2008.

[50] Eduardo J Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 513–522. ACM, 2012.

[51] Robert P Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):12, 2009.

[52] Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[53] Eric Gilbert and Karrie Karahalios. Widespread worry and the stock market. In *ICWSM*, pages 59–65, 2010.

[54] Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. Predictive sentiment analysis of tweets: A stock market application. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pages 77–88. Springer, 2013.

[55] Nuno Oliveira, Paulo Cortez, and Nelson Areal. Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from twitter. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, page 31. ACM, 2013.

[56] Enric Junqué De Fortuny, Tom De Smedt, David Martens, and Walter Daelemans. Evaluating and understanding text-based stock price prediction models. *Information Processing & Management*, 50(2):426–441, 2014.

[57] Yuexin Mao, Wei Wei, Bing Wang, and Benyuan Liu. Correlating s&p 500 stocks with twitter data. In *Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research*, pages 69–72. ACM, 2012.

[58] Timm O Sprenger, Andranik Tumasjan, Philipp G Sandner, and Isabell M Welpe. Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5):926–957, 2014.

[59] B Shravan Kumar and Vadlamani Ravi. A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114:128–147, 2016.

[60] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction. *ACL (2)*, 2013:24–29, 2013.

[61] Jyothi Subramanian and Richard Simon. Overfitting in prediction models–is it a problem only in high dimensions? *Contemporary clinical trials*, 36(2):636–641, 2013.