Università degli Studi di Cagliari

DOTTORATO DI RICERCA

Ingegneria Elettronica ed Informatica

Ciclo XXVIII

TITOLO TESI

The Predictor Impact of Web Search and Social Media

Settore/i scientifico disciplinari di afferenza

ING-INF/05

Presentata da:	Martina Matta
Coordinatore Dottorato	Fabio Roli
Tutor	Michele Marchesi

Esame finale anno accademico 2014 – 2015

UNIVERSITY OF CAGLIARI

DOCTORAL THESIS

The Predictor Impact of Web Search and Social Media

Author:
Martina MATTA

Supervisor:
Prof. Michele MARCHESI

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Computer Science and Electronic Engineering

in the

Agile Group
DIEE

January 31, 2016

Contents

List of Figures

List of Tables

Dedicated To my Family

Chapter 1

Introduction

The advent of the Internet has completely changed the way real life works. The users can interact at once and exchange information in a very simple way. By enabling practically all Internet users to interact at once and to exchange and share information almost cost-free, more efficient decisions on several fields are possible.
The majority of daily activities radically changed, moving towards a “virtual sector”, such as web actions, credit card transactions, electronic currencies, navigators, games, and so on. In recent years, web search and social media have emerged online.
Search engine technology has had to speed up to keep up with the growth of the World Wide Web, that has turned the Internet into a wide information space with different and badly managed content. Millions of people all over the world search online several information each day, which makes Web search queries a valuable source of information.
Due to the huge amount of available information, searching has become dominant in the use of Internet. Users that daily interact with search engines, produce valuable sources of interesting data regarding several aspects of the world. The amount of data on the web is growing more and more, as well as the number of new users that explore the web research.
The rise pace of Internet led companies of different businesses to think new way of communication with users. Among the fastest growing online tools for reaching the consumers is the so called "social media". It has exploded as a category of online discourse where people create content, share it, bookmark it and network at a prodigious rate.
Social media increasingly pervades life in several fields of the world, enabling communication among users and collecting massive amount of information for social media companies that want to refine their products.
Social media represents the online content publicly available to end users. Blogs, micro blogging services, forums, networking sites stands for groups of users that share content, worldwide available. Social media have become popular communication platforms in order to have the possibility to share ideas, opinions and sentiments about several aspects of the world. A big amount of studies have been applied to monitor the trending topics or events like political events, stock marketing fluctuations, epidemics and so on.
Popular services like Twitter and Facebook attract a lot of users who share facts of their daily life. This kind of content has become more present on the web and, due to its public nature, even appears in search results from search engines, like Google and Bing. With the explosion of user generated content, came the need by politicians, analysts, researcher to monitor the

content of different users.
Sentiment analysis or opinion mining, is one of the areas of computational studies. The research field of sentiment analysis has developed many algorithms to verify whether an online text is subjective or objective, and whether the expressed opinion is positive or negative [64].
During my PhD, I decided to investigate whether social media activity or information collected by web search media could be profitable and used for predictive purposes. I studied whether some relationship exists between particular phenomena and volume of search data, considering the examined topic on web engines. Then, I analyzed the related social volume in order to discover whether the chatter of the community can be used to make qualitative predictions about the considered phenomena, attempting to establish whether there is any correlation.
The frequency of searches of terms could have a good explanatory power, so I decided to examine Google, one of the most important search engine. I studied whether web search media activity could be helpful and used by investment professionals, analyzing the search volumes power of anticipate particular volumes related to the topic monitored.
Simultaneously, I decided to apply automated Sentiment Analysis on shared short messages of users on Twitter in order to automatically analyze people opinions, sentiments, evaluations and attitudes. We wondered whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can be used to predict something in particular.
The analyzed phenomena are listed below.

- Trading Volume of Bitcoin
- Bitcoin price pace, expressed in dollars
- Agile Project Management tools
- Belgian Government Bond

1.1 Thesis Overview

This thesis is organized as follows:

- Chapter 2 illustrates the related works that have motivated this kind of research.
- Chapter 3 illustrates the background of this thesis. The web search media is explained more in detail, with a particular focus toward Google Trends, based on Google as search engine. This chapter presents, also, the social media, with an analysis of the most used systems: Facebook and Twitter.
  Finally, I describe the sentiment analysis, a particular technique applied to written texts in order to identify automatically the sentiment (or opinion) transmitted by users. The chapter illustrates the most common tools used to verify it.
- Chapter 4 describes the prediction study pointed towards the Bitcoin, a digital currency created in 2008 by Satoshi Nakamoto with the purpose to replace cash, credit cards and bank wire transactions. I compared the web search and social volume with two different aspects of

Bitcoin's market: the price, expressed in dollars, and the trading volume. I found positive outcomes from these analysis, demonstrating our objectives.

- Chapter 5 presents an analysis of prediction related to the Agile methodologies. From a multitude of available tools that support Agile methods, I investigated about which tools are the most looked for using web search media, most mentioned and most appreciated through social media. Furthermore, I applied automated Sentiment Analysis on shared short messages of users on Twitter in order to analyze automatically user's opinions, sentiments, evaluations and attitudes.

- Chapter 6 shows a predictive study, oriented toward the Belgian Government Bond and the economic policy uncertainty effect. During my abroad staying in the University of Antwerp, I decided to investigate in more detail about the predictive power of social volume, transmitted by the main economists of the country, about the Belgian economy. I found a striking similarity between these data, demonstrating the predictive capability of social media.

- Chapter 7 presents the conclusions and plans for future work.

Chapter 2

Related Work

Internet has been one of the most revolutionary technologies in the last decades. The majority of daily activities radically changed, moving towards a “virtual sector", such as Web actions, credit card transactions, electronic currencies, navigators, games, etc. In recent years, web search and social media have emerged online.
On one hand, services such as blogs, tweets, forums, chats, email have gained wide popularity. Social media data represent a collective indicator of thoughts and ideas regarding every aspect of the world. It has been possible to assist to deep changes in habits of people in the use of social media and social networks [45].
In these decades, social web has been commercially exploited for goals such as automatically extracting customer opinions about products or brands, to find which aspects are liked and which are disliked [79]. In their work, Ye and Wu demonstrate how particularly interesting is the influence of Twitter users and the propagation of the information related to their tweets[85].
According to Alexa [1], Twitter had become the world’s seventh most popular website by March 2015. Java et al. affirmed that it seems to be used to share information and to describe minor daily activities [43]. The short format of tweet is a defined characteristic of the service, allowing informal collaboration and quick information sharing.
Black et al. presented a survey conducted to collect information on social media use in global software systems development. Twitter was found to be the most popular media and respondents affirmed that specification, source codes and design information were shared over social media [10].
Singer et al. showed that Twitter helps developers keep up with the fast-paced development landscape. They use it to stay aware of industry changes, for learning and for building work relationships [76].
Romero et al. showed that the aspect responsible for the popularity of certain topics is the influence of users of the network on the spread of content. Some members produce content that resonates very strongly with their followers thus causing the content to propagate and gain popularity [70].
Twitter is a rich source of real-time information regarding current societal trends and opinions. There are also studies that report another use of Twitter, namely as a possible predictor of market trends. Indeed, in 2010, a publication of the professor Johan Bollen showed that combining information on Wall Street with the millions of Tweets and posts, makes possible to anticipate financial performance [12]. The analysis of tweets made by Bollen would have had 87% of chance to successfully predict prices of the stock, 3 or 4 days in advance. This study and analysis of millions of posts

[1] http://www.alexa.com/

on Twitter represents a thermometer of emotions, on a large scale, which reflects the whole of society.
Earlier studies had found that blogs can be used to evaluate public mood, and that tweets about movies can predict box office sales. Investigating the literature related to different uses of social media, and Twitter in particular, we collected information about the use of Twitter for seeking real world emotions that could predict real financial markets trend [44].
In their paper, Rao and Srivastava investigated the complex relationship between tweet board literature (like bullishness, volume, agreement etc) with the financial market instruments (like volatility, trading volume and stock price) [68].
Twitter and other social media offer a plethora of opportunities to reveal business intuitions, where it remains a challenge to identify the potential social audience. In their work Ling et al. [53] analyzed the Twitter content of an account owner and its list of followers through various text mining methods and machine learning approaches, in order to identify a set of users with high-value social audience members.
In their paper, Ciulla et al. [21] assessed the usefulness of open source data that come from Twitter for prediction of societal events by analysing in depth the microblogging activity surrounding the voting behaviour on a specific event.
Mocanu et al. performed a comprehensive survey of the worldwide linguistic landscape emerging from mining the Twitter microblogging platform [59]. Hick et al. explored the opportunities and challenges in the use of Twitter as platform for playing games, through crawling game that uses Twitter for collaborative creation of game content[42].
Social media technologies have produced completely new ways of interacting [40], bringing the creation of hundreds of different social media platforms (e.g., social networking, shared photos, podcasts, streaming videos, wikis, blogs).
On the other hand, due to the huge amount of available information, searching has become dominant in the use of Internet. Millions of users daily interact with search engines, producing valuable sources of interesting data regarding several aspects of the world.
Recent studies demonstrated that web search streams could be used to analyze trends about several phenomena [20] [72] [13]. In one of the most interesting works, Ginsberg et al. proved that search query volume is a sophisticated way to detect regional outbreaks of influenza in USA almost 7 days before CDC surveillance [35].
There are also studies that report another use in a search engine, namely as a possible predictor of market trends. Bollen et al. showed that search volumes on financial search queries have a predictive power. They compared these volumes with market indexes such as Dow Jones Industrial Average, trading volumes and market volatility, demonstrating the possibility to anticipate financial performances [12].
In this work, Granger causality analysis and a Self-Organizing Fuzzy Neural Network are used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. Bordino et al. proved that search

volumes of stocks highly correlate with trading volumes of the corresponding stocks, with peaks of search volume anticipating peaks of trading volume by one day or more [13].
Search queries prove to be a useful source of information in financial applications, where the frequency of searches of terms related to the digital currency can be a good measure of interest in the currency and it has a good explanatory power [47].
Mondria et al. proved that the number of clicks on search results stemming from a given country correlates with the amount of investment in that country [60]. Further studies showed that changes in query volumes for selected search terms mirror changes in current volumes of stock market transactions [67].
Kristoufek [48] studied the popularity of the Dow Jones stocks, measured by Google search queries for portfolio diversification. Curme et al. [24] clustered the online searches into groups and showed that mainly politics and business oriented searches are connected to the stock market movements.
Preis et al. [66] demonstrated that Google searches, for financial terms, can support profitable trading strategies. Dimpfl et al. found a strong relationship between internet search queries and the leading stock market index. In addition they found a strictly correlation between the Dow Jones' realised volatility and the volume of search queries [30].
Kristoufek proposed the study of Power-law correlations for Google searches queries for Dow Jones Industrial Average (DJIA) component stocks, and their cross-correlations with volatility and traded volume [49]. Bordino et al. proved that search volumes of stocks highly correlate with trading volumes of the corresponding stocks, with peaks of search volume anticipating peaks of trading volume by one day or more [13].
In his work [15], Bulut described that internet search data, via Google Trends, is utilized to nowcast the known variates of two structural exchange rate determinations models. By using internet search data, the author aims to get a timely description of the state of the economy way before the official data are released to the market participants.
Kim et al. [46] introduced an analysis system to predict the value fluctuations of virtual currencies used in virtual worlds, and based on user opinion data in selected online communities. In their proposed method, data of user opinions on a predominant community are collected by employing a simple algorithm and guaranteeing a stable prediction of value fluctuations of more than one virtual currency.
Search queries prove to be a useful source of information in financial applications, where the frequency of searches of terms related to the digital currency can be a good measure of interest in the currency [47].
Mondria et al. proved that the number of clicks on search results stemming from a given country correlates with the amount of investment in that country [60]. Further studies showed that changes in query volumes for selected search terms mirror changes in current volumes of stock market transactions [67].

Chapter 3

Background

3.1 Web Search Media

A web search media is an online system, built to look for several information regarding different phenomena on the world wide web. The search results are generally exhibited in a line of results, often referred to as search engine results pages. The information may include a mix of web pages, like images, video, music and other kind of files distributed over the Internet, which are either non-copyrighted or copyrighted materials, provided either freely or for a fee.
Some search engine are also able to mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.
Search engine technology has had to scale dramatically to equalize itself to the growth of the web. In 1994, one of the first search engines, the World Wide Web Worm had an index of 110000 web pages and web accessible files. In April 1994, the world wide web worm received almost 1500 queries per day. Starting from November 1997, the top search engines claim to index from 2 million to 100 million web documents.
In November 1997, Altavista claimed it reached 20 million queries per day. After the year 2000, a comprehensive index of the web included over a billion of documents. At the same time, the number of queries from search engines has grown substantially too. With the increasing number of users on the web, and automated systems which interrogate search engines, it is well evident that top search engines started to manage millions of queries per day by the year 2000 [1].
Internet has been one of the most revolutionary technologies in the last decades. Due to the huge amount of available information, searching has become dominant in the use of Internet. Millions of users daily interact with search engines, producing valuable sources of interesting data regarding several aspects of the world.
The available information and data on the web is swiftly increasing, as well as the number of new users unpractised in the skill of web research. Looking for considerable information on the world wide web is often an arduous and disheartening task for casual and experienced users.
The exponential growth of the world wide web has turned Internet into a limitless information space with several and rough content. The new world

[1] https://en.wikipedia.org/wiki/Websearchengine

of the web found new challenges in several fields, like the information retrieval. Nowadays, different sources of data regarding the economic activity is available from sector companies like Google, MasterCard, Federal Express.
Nowadays, Google is the world's most popular search engine, with a desktop market share of 69,24%, referred to September 2015. Bing comes in at second place (12,26%), followed by Yahoo and Baidu (the search engine most used in China). Taking into account the mobile/tablet search engine market share, Google achieves a share of 92,83%, followed by Yahoo with only 4,67% of share [2].
For this reason, I decided to analyze Google as web search engine, given its growing fame. A brief introduction to Google Trends is necessary to provide a context for that study.

3.1.1 Google Trends

The web is becoming always more complex and various, offering several kind of information that could be used for research aims. Every day, an increasing number of people is continuously looking for information about issues of interest or concern. Search topics reflect the real interests of the society and for this reason it's truly interesting analyze economic, social, political phenomena such as political preferences, disasters opinions, inflations or consumer behaviours.
Google Trends at *http://www.google.com/trends* is a service that supplies several data about particular researches from users on Google as web search engine. Google Trends service has been developed by Google Labs to give the possibility to analyze queries and news articles from Google's databases of searches and new articles and to verify the popularity of a fixed topic. It's the only service that provides a real-time index, allowing the users to insert one or more words and to monitor the query volume for those keywords.
The user has the possibility to insert additional parameters, such as the state, time period and category, in order to analyze specific search volumes. Google Trends provides a time series index of the volume of queries inserted from users into google as search engine. The query index is built by calculating a section of all web searches performed worldwide for fixed terms relative to the total amount of searches done over time. Only words with a substantial query volume are available in order to avoid recognizable data. The maximum query index is normalized to be 100 and the minimum is normalized to be zero.
The searches on Google Trends system exploits a query language different from the classic google search technique. The user has the possibility to insert a single word or more than one and to combine them using a simple vertical line "|" or remove words using the minus character "-". Up to five topics can be compared at one time (separated by a comma) to view their relative popularity during the time.
The results are illustrated using two different graphs:

- *search volume graph*, calculated by the number of web searches for a fixed word under examination relative to all the web searches.

[2] https://www.netmarketshare.com/

- *News volume graph*, calculated by the number of times that a topic appears in a news article indexed by google news service.

Many studies have examined the relationship existing between web search volume and several phenomena regarding different aspects of the world. Using Google Trends, all these aspects can be monitored. To date, the most common fields of Google search data in social science have been economics, finance, technology, industry and policy. Several researches demonstrated that the query indexes are strongly correlated with economic indicators and it could be useful for economic predictions [20] [47].
Google can be used to examine several fields and some examples are shown below.

- Support acquisition decision
- Support business choice
- Identification of new research activities
- Investigation markets trends
- Interest in the new technologies

After the comparison with the search volume, it's possible to identify what factors attract more attention in the public or in the media. Peaks in the news could cause a peak in the searches and a more intensive analysis can be useful to determine the causes of the increased interest. So, Google Trends can be seen as an interesting new chance that offers a lot of applications in several fields.

3.2 Social Media

The wide spread of Internet and the consequent exponential growth of users in the use of digital media obliged several companies to consider a new way of communication with customers [19]. This quick growth of online tools for connect companies and consumers is called "social media".
With the beginning of Internet-based smartphones, more and more people are connected on the web. These services have changed the ways to communicate and share about his own life with families, friends or colleagues. Social media can be seen as a collection of online communications channels where users create online communities to share information, ideas and other contents regarding several aspects of the world. So, it represents an online collection, created and available to end users.
The impact of these services on our society is huge from the economical and sociological point of view, reflecting good and bad aspects. External media emphasize more the negative effects than the positively changes of the social life. More than 110 billion minutes are spent on social media networks and blog sites [3].
Kaplan et al. showed in detail the meaning of the social media, its use and origin, affirming that web 2.0 is the main platform for the evolution of social media [45]. Web 2.0 appeared, for the first time, in 2004 with the aim

[3] http://www.nielsen.com/us/en/insights/news/2010/social-media-accounts-for-22-percent-of-time-online.html

to illustrate how software developers and users operated with the world wide web. The web 2.0 was built to express the new evolutionary trends of the web with a collection of new technologies, like applications, ideas, strategies and social fame in the web [62].
The first social network was UseNet Newsgrous (www.usenet.com), created and implemented by Duke University students in 1979. From that year, the online social networks have been a continuous growth in size and numbers. The User Generated Content, shows all the possible ways to use the social media.
With the proliferation of web 2.0, websites and applications dedicated to social networking, microblogging, forums and wikis are among different kind of social media. Content communities give the possibility to share several kind of media content between users, like the exchange of videos (youTube), photos (Flicker or Instagram) or PowerPoint presentations (SlideShare). By means of these platforms, users can share personal information, videos and instant messages with the world wide people.
The most important social media are briefly described below.

- *Facebook*: popular free social networking platform that allows users to create profiles, upload several kind of files, keep in touch with other users. According to statistics from the NIELSEN group, United States citizens spend more time on Facebook than any other website. Facebook started as a private network for Harvard University students.

- *Twitter*: free social network and microblogging service that allows users to post and read short messages, that are up to 140 characters, called tweets. Twitter members can share tweets and follow other users using several platforms and devices.

- *Google+*: social network platform designed to replicate the way people interact offline more closely than is in the case of other social services.

- *LinkedIn*: social network designed for the business community. The aim is to create a business connection of users they know professionally.

- *Wikipedia*: a free, web-based, online encyclopedia written by volunteer users and available by anyone who has an internet connection.

Nowadays, social media is becoming an important part of life online. In business, social media is used to market products, promote brands and to create new business. Thanks to the grown social media, mobile and web-based technologies are used to create interactive platforms where users share, create and discuss user-generated content.
People use social media for a number of reasons, like staying in touch with friends, making new friends, monitoring them and so on. The main reason is communication and preserving relationships. Popular activities include the update of statuses, sharing pictures and events, sending private messages and posting public testimonials and feedbacks. This overview of social media started a dozen years ago. With the explosion of user generated content, companies, researchers, social psychologists and analysts started to monitor the content for several aims.

3.2.1 Twitter and its API

Recently, the phenomenon of micro-blogging services is based on real-time updates and Twitter has become one of the fastest growing trends on the Internet. Twitter is a very popular microblogging service, where users can publish their statuses in short messages, follow updates of the people they are interested, resend directly other posts and so on.
Twitter, launched in 2006, is growing exponentially and now it has more than 145 million of users that send on average 90 million of tweets per day, equally 750 tweets at second. Each tweets contains a maximum of 140 characters and represents short status updates of what users are doing, thinking or how they are feeling in that moment. The messages are visible publicly or semi-publicly on a notice board of its platform or using third-party applications.
Due to its growing popularity, tweets are about different topics, ranging from economic news to brand information. Twitter's fame is attracting a lot of attention from researchers of several disciplines. There are several fields of research, one of them focuses on understanding its usage and community structure. Another stream of research is about the influence of users and the propagation of information. Finally, there is a field of research focused on the predictive power and the potential application to different areas [73]. There are several studies that investigate the Twitter sentiment during elections or economical/political events [81], analyze discussions performances [18], measure user influence or study the predictive power of Twitter in several fields, like financial markets.
Following someone on Twitter allows to receive all his tweets updates, and every time that he posts something online, it will appear on the personal main webpage. Each user preserves a user page where the posts of the followed people are showed. Since Twitter doesn't give the possibility to group tweets, the *Hashtag* has been created. It is a convention for the Twitter users, just to follow a thread of discussion by prefixing a word with # as character. In that way, popular hashtags are captured in trending topics.
Twitter gives third party applications developers access to its data store. The Twitter API [82] is based on three parts, two Rest API and a streaming API.

- *The Original REST API* allows developers to visit core Twitter data in order to update tweets or profile information.
- *Search API* supports short-lived connection, rate-limited and permits to interact with Twitter Search and its trends. This is read-only in the search database and the search queries return up to 1500 tweets from up to seven days before. Unfortunately, this time frame is getting smaller as traffic online to Twitter increases and it has a strong impact on the performance of the service.
- *Streaming API* provides real time data access to tweets in filtered forms and supports long-lived connection. The filter method returns public statuses that match with one or more filters. It allows up to 400 track keywords, 5000 follow user id's and 25 location coordinates. The stream could be filtered by one or more keywords or by a list of users that post the messages. It gives back the most recent tweets that match with the initial parameters. By default, this kind of API

doesn't allow developers to find all the historical tweets but enables the access to data as it is being tweeted.

3.2.2 Facebook and its API

Facebook is a social networking system, built in the first months of 2004. According to statista.com [4], as of the third quarter of 2015, Facebook had 1.55 billion monthly active users. In the third quarter of 2012, the number of active Facebook users had surpassed 1 billion.
It has been created by Mark Zuckerberg, a Harvard student that opened his website to the other students and it grew to encompass other people and finally the general world community. Facebook original aim was to connect students together through interests, residence, classes and so on.
This social network is free to its users and its profit comes from advertisements and sponsor groups. People create personal profiles, share any kind of information about themselves with other people that they label as "friends". This friendship gives the opportunity to monitor the actions of the other side.
Thanks to the socio-technological features, Facebook achieved million users in a very short time attracting a lot of attention by researchers, psychologists, analysts, web developers and so on.
The Facebook external interface is based on three primary components:

- REST-based Application Programming Interface (API).
- Custom markup language (FML)
- Custom database query language (FQL).

Recently, Open Graph API has been introduced to encourage third-parties to generate more and more traffic from and to Facebook. Facebook releases the data kept within its database to build a social graph. In the math field, a graph theory is a graph with a set of vertexes and edges that link these vertexes.
Since 2006, the developers have been able to work directly with Facebook's API to let users connect their account to third-party applications. Facebook releases an official API library for PHP and Java, but other companies have written similar libraries in other languages, cited on Facebook and released on Facebook's developer website. The offered languages are ASP.NET, C++, C#, Java, Perl, PHP, Python, Ruby and VB.NET.
In 2011, Facebook released significant improvements to its structure with the introduction of the Graph API Explorer and a new version of the Developer app. The Graph API explorer lets developers run API calls to the system and see formatted results (JSON or XML format) in real-time or explore connections between objects and more.
User data, friends data and several permissions access tokens can be created to view private information and check APIs. Developers can switch between different Graph API URLs and between GET, POST and DELETE to make changes. The formatted results can be navigated to see details of that object. Figure 3.1 illustrates the main view of Graph Explorer with all the available possibilities.

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

FIGURE 3.1: Operating principles of Facebook using Graph Explorer.

3.3 Sentiment Analysis

The sentiment analysis is the process of extraction of opinions and subjectivity from unstructured texts. This technique evaluates whether a written text expresses a positive, negative or neutral sentiment. This is also known as opinion mining and it's used to discover how people feel about a particular topic [64].
Sentiment analysis is useful for research fields on online messages in order to automatically detect the emotion transmitted in online texts. In reviews, people express different opinions on several topics and the opinion mining process can be adjusted based on a fixed subject we are interested in. The analysis is completed on a global topic level, where the result is a general opinion on the discussed element.
The sentiment of a text could be critical in several fields of application like:

- Mining and analysis of several kind of users related to books, movies, restaurant or hotels reviews.
- Monitoring political and economical opinions
- classifying blog posts and comments

The sentiment analysis work can be divided in different views: technique used, level of detail of the text examined and so on. Taking into account the technical approaches, the identified methods are machine learning, lexicon-based, statistical and rule-based approaches.

- *The machine learning approach* is based on different learning algorithms to identify the opinion by training a well known dataset. Machine learning methods often rely on supervised classification approaches, where the sentiment can be identified as binary (positive or negative). This kind of method needs tagged data in order to train classifiers and to built trained model for specified contexts and kind of studies. For the standard machine learning, a set of texts checked for the polarity by human coders, are used to train an algorithm and to be able to

associate a positive, negative or neutral sentiment. The trained algorithm can be used in new texts in order to predict the polarity of the analyzed sentence.

- *The lexicon-based method* requires to calculate the sentiment polarity using a semantic orientation (measure of subjectivity and opinion in texts) of words or sentences. Lexical-based methods make use of a predefined list of words, where each word has a particular weight that represents the sentiment. These methods change according to the context in which they are built.

- *The rule-based approach* is based on the opinion found in a written text and it is able to classify it on the number of positive or negative words. The considered rules can be different, such as dictionary polarities, negation words, booster words, emoticons, mixed opinions.

- *Statistical models* represent each sentence as a mix of latent aspects and ratings. It can be represented by multinomial distributions and the head terms can be clustered.

Sentiment methods have been widely used for implementing applications using these different techniques. The available research presents machine learning approaches (Naive Bayes, Maximum Entropy and SVM) to be more suitable for Twitter than the lexical-based methods [63].
We illustrated the eight most popular sentiment analysis methods in the literature covering a mixture of techniques, such as the Natural Language Processing (NLP) to assign the sentiment score, the use of Amazon's Mechanical Turk (AMT) to create datasets with labels, the use of supervised or unsupervised machine learning techniques.

3.3.1 Emoticons

A rapid way to detect the sentiment on online social networks is to consider the emoticons that it contains. A lot of people use emoticons to represent sentiments and feelings and some of them are now included in the English Oxford Dictionary. For example :) stands for a smile and it expresses the happiness.
Starting from a table that contains a list of the emoticons it's possible to check the level of the positive, negative or neutral sentiment. They can be used in combination with other techniques in order to build a training dataset with a supervised machine learning approach.

3.3.2 LIWC

It's a text analysis tool that evaluates emotional, cognitive and structural components from a written text starting from a words dictionary with their classification in categories. The LIWC software, available at `http://www.liwc.net/`, is commercial with customizable options for the users.
Basically, it reads a written text and checks the percentage of words that spread different levels of emotions, style, and kind of speech in order to capture social and psychological states of users. The program is based on a main test analysis module along with a group of built-in dictionaries. This

module has been created using Java as programming language and it compares each word of a sentence with a user-based dictionary.

3.3.3 SenticNet

SenticNet is a tool able to identify the opinion of users exploring semantic Web techniques [16]. The aim is to identify the polarity of a written text by means of Natural Language Processing (NLP) approach combining commen-sense reasoning, psychology, linguistics and machine learning. SenticNet analyses the polarity and sentiment by levering on semantic and linguistic rules instead on co-occurrence frequencies.
The semantic concepts are considered the most semantically related to the input idea. Semantics are emotional values expressed in terms of four affective dimensions (Pleasantness, Attention, Sensitivity and Aptitude) and polarity conveyed as a number between -1 (extremely negative) and +1 (extremely positive). The tool can be seen as a multi-disciplinary paradigm that goes beyond statistical methods by focusing on a semantic-preserving representation of natural concepts. SenticNet has been tested and analyzed to measure the level of polarity of patients opinions about the National Health Service in England. The authors also tested the tool with data from LiveJournal blogs, categorizing the data with positive or negative sentiment [17].

3.3.4 SentiWordNet

SentiWordNet is one of the tool most used in the opinion mining field and it is a corpus-based lexical resource constructed from the perspective of WordNet [31]. It gathers adjectives, nouns, verbs and other grammatical classes into synonyms, called synsets.
The system assigns a triple of polarity score to indicate the sentiment of the examined text: positive, negative or objective (neutral), applying a semi-supervised machine learning method. For example, the triple 0, 1, 0 (positivity, negativity, objectivity) is assigned to the synset of the word “bad”. The sum of all scores related to the triple is 1.
SentiWordNet has been created automatically through a combination of linguistic and statistic classifiers. It has been applied in the sentiment analysis field with promising results.

3.3.5 PANAS-t

This is a psychometric scale proposed to investigate mood fluctuations of users on Twitter [37]. The PANAS-t is based on a large set of words where them are associated with eleven moods: sadness, joviality, assurance, serenity, surprise, fear, guilt, hostility, shyness, fatigue, and attentiveness. This approach has been created to monitor any increase or decrease in sentiments during the time.
The method calculates the P(s) score for each sentiment s for a fixed period as values between -1 and +1. Joviality, assurance, serenity, and surprise are considered as positive affect. Fear, sadness, hostility, shyness and fatigue are considered as negative.

3.3.6 HAPPINESS INDEX

This method is a sentiment scale that considers the Affective Norms for English Words (ANEW). ANEW is a set of almost 1030 words and HappinessIndex is based on this list of words and has scores for a written text between 1 and 9, representing the level of happiness found in the text.
The authors applied this approach to a dataset of song lyrics, song titles and blog sentences [14].

3.3.7 PATTERN.EN

Pattern is a web mining module for Python programming language [27]. It contains modules for data mining (Google, Twitter and Wikipedia API, HTML DOM Parser and a web crawler), natural language processing (n-gram search, sentiment analysis, WordNet), machine learning (Vector Space Model, SVM and clustering) and network analysis. Pattern.en is a particular module that includes a part-of-speech tagger for English (nouns, adjectives, verbs are discovered), sentiment analysis and a WordNet interface.
The analyzed text can be classified in facts and opinions that represent sentiments and feelings of the people regarding several aspects of the world. The module evaluates a lexicon of adjectives that occur frequently and several scores are annotated for sentiment polarity (positive – negative) and subjectivity (objective-subjective).
A particular method returns a (polarity,subjectivity)-tuple for a given sentence, where the polarity is a value between -1 and +1 and the subjectivity is between 0 and 1.

3.3.8 SENTISTRENGTH

Sentistrength is a machine learning method, suitable for applications that need content-driven or polarity identification models [79] [80]. This tool is the most suitable for social media analysis, compared to a wide range of supervised and unsupervised methods, like the simple logistic regression, SVM, J48, SVM regression and Naive Bayes.
SentiStrength estimates the strength of positive and negative sentiments in written texts. It is based on a dictionary of sentiment words and, in addition, it uses some rules non-standard grammar.
The authors considered different features to better understand the opinions expressed in social media networks.
The core of the algorithm is the list of positive and negative words. This is a collection of almost 300 positive words and 465 negative terms, classified for their sentiment strength with values from 2 to 5. These values are based on human judgements. The default manual word strength list are updated by a training algorithm in order to optimise the sentiment strengths. The key elements are illustrated below.

- List of booster words to strengthen or weaken sentiments. Each word increases the emotion by 1 or 2 (very, extremely,...) or decreases it by 1 (some,..)
- Consideration of negative word list that inverts subsequent emotion words. For example, if “very happy” has positive strength then “not very happy” would have negative strength.

- List of emoticons with related strengths.
- Consideration of repeated punctuations or characters to emphasize the strengthen sentiments.

SentiStrength has been tested on a set of 1040 MySpace comments and they were examined and classified by three human coders and a 10-fold cross correlation was used. SentiStrength was able to identify the strength of positive sentiments on a scale from 1 to 5 in 60,6% of the time, significantly above the best machine learning approaches that had a performance of 58,5%.
Compared to other methods, SentiStrength showed the highest correlation with human coders. The authors released the tool and it produces the best training model empirically obtained thanks to the combination of the main learning techniques. For this highest correlation with human coders, I decided to use SentiStrength tool, analyzing the opinion transmitted by users on social media.

Chapter 4

Bitcoin Analysis

4.1 Bitcoin

Bitcoin is a decentralized cash system, created and managed in an electronic way. It has been designed and developed by Satoshi Nakamoto in 2008 (whose name is conjectured to be fake by some, and who has not been hear before) [61].
Bitcoin can be considered as an electronic digital system based on the mathematical proof. The idea was to reproduce a currency independent of any central authority, transferable only electronically, with very low transaction fees. Bitcoin relies on digital signature to prove ownership and a public history of all transactions, in order to prevent double-spending.
Bitcoin is an innovative payment network and a new kind of digital money. It is based on the advancement on peer-to-peer networks in order to operate with no central authority or banks. It is different from the conventional money, because it's decentralized and no single institution checks the bitcoin network. It means that one central authority can't tinker with a monetary policy and cause a breakdown, or simply decide to take people's bitcoin away from them, as the central European bank decided to do in Cyprus in early 2013. The management of transactions and the issuing of Bitcoin is carried out by the network.
Bitcoin is open-source, meaning that everyone can take part of it. Its protocol affirms that only 21 million bitcoin can be created by miners [1]. Anyway, these coins can be divided into smaller parts up until the smallest part that is the Satoshi, one hundred millionth of bitcoin [71].
Users can preserve multiple addresses and they aren't connected to name, addresses, or other personal information. However, details of every single transaction are stored in the blockchain, a shared public ledger on which the Bitcoin network depends on. In fact, all the confirmed transactions are saved in the blockchain.
A personal Bitcoin wallet generates the Bitcoin address and this address could be shared to other people in order to pay or receive Bitcoin for something. Bitcoin wallet can calculate the spendable money and new transactions can be authorized to be spending the bitcoins. The integrity and the chronological order in the blockchain is enforced with the use of cryptography. Bitcoin is the first example of growing category of money, known as cryptocurrency.
A transaction is a transfer of money between different Bitcoin wallets that will be included in the blockchain. The wallet preserves a secret private key, used to sign transactions, providing a mathematical proof that this money

[1] https://en.bitcoin.it/wiki/Mining

comes from the owner of the wallet. The signature phase prevents the alteration not authorized of the transaction. All the transactions are checked and confirmed in the following 10 minutes, using a process called Mining. The cryptography can be seen in two different aspects:

- public-private key to save and spend money.
- cryptographic validation of transactions.

The public-private key allows users to create a public key with an associate private key [29]. Messages encrypted with a public key can be decrypted only by people that possesses the correspondent private key. Instead, texts encrypted with a private key can be decrypted by someone that has the correspondent public key.
The mining process works on distributed systems to confirm waiting transactions by including them in the blockchain. Bitcoin's rules were decided and implemented by a team of developers with no evident lawyer influence. Rather than save the transactions on any single server or collection of servers, Bitcoin is based on a transaction log, that is distributed across a network of powerful computers. This process enforces a particular order in the blockchain and allows to multiple computers to accept the state of the system.
The transactions are collected in a block that fits with cryptographic protocols that will be verified on the network. These kind of rules prevents previous blocks from being changed and wards off any individual from adding new blocks consecutively in the blockchain.
Each single bitcoin can be monitored through all transactions in which it has been used. In fact, all Bitcoin transactions are available and readable by everyone in several records in a widely replicated data structure. Transactions are ordered recursively starting from the input of a transaction that refers to the output of a previous transaction.
Since Bitcoin is based on advancements in peer-to-peer networks [71] and cryptographic protocols for security, Bitcoin is completely decentralized and not managed by any governments or bank, ensuring anonymity. Like any other currency, a peculiarity of Bitcoin is to facilitate transactions of services and goods with vendors that accept Bitcoins as payment [39], attracting a large number of users and a lot of media attention.
Bitcoin has been helpful for almost 62.5 million transactions between 109 million accounts. During November 2015, the daily transaction volume was roughly 216,000 bitcoins and the total amount of bitcoin in circulation was $7 billion [2]. The table 4.1 summarizes Bitcoin information to date.

The Bitcoin represents an important new phenomenon in financial markets. For instance, Mai et al. examine predictive relationships between social media and Bitcoin returns by considering the relative effect of different social media platforms (Internet forum vs. microblogging) and the dynamics of the resulting relationships using auto-regressive vector and error correction vector models [54].

[2] https://blockchain.info/

TABLE 4.1: Main information of Bitcoin to date.

Main information of Bitcoin to date	
Total blocks minted	178
Time between blocks	8,09 minutes
Bitcoins mined	4,450 BTC
Number of transactions	216.760
Trade volume	7,155,405.63 USD
Estimated transaction volume	514,998
Price expressed in dollars	420.48 USD (weighted)

4.2 Bitcoin Spread Prediction Using Social And Web Search Media

4.2.1 Introduction

Bitcoin, a decentralized electronic currency system, represents a radical change in financial systems after its creation in 2008 by Satoshi Nakamoto [61]. Bitcoin stands for an IT innovation based on the advancement in peer-to-peer networks [71] and cryptographic protocols. Due to its properties, Bitcoin is not managed by any governments or bank. Like any other currency, a peculiarity of Bitcoin is to facilitate transactions of services and goods [39], attracting a large number of users and a lot of media attention.

Nowadays, Web 2.0 services such as blogs, tweets, forums, chats, email etc. are widely used as communication media, with satisfying results. Sharing knowledge is an important part of learning and enhancing skills. Through the use of social media services, team members have the opportunity to acquire more detailed information about their peers' expertise [28]. Social media data represents a collective indicator of thoughts and ideas regarding every aspect of the world. It has been possible to assist to deep changes in habits of people in the use of social media and social network. Twitter, an online social networking website and microblogging service, has become an important tool for businesses and individuals to communicate and share information with a rapid growth and significant adoption. In addition, Twitter has rapidly grown as a mean to share ideas and thoughts on investing decisions.

In this work we analyze whether social media activity or information extracted by web search media could be helpful and used by investment professionals. There are several works that present predictive relationships between social media and bitcoin price where the relative effects of different social media platforms (Internet forum vs. microblogging) and the dynamics of the resulting relationships, are analyzed using cross-correlation such as [23] or linear regression analysis such as [12] or [58]. Social factors, that are composed of interactions among the actors of the market, may strongly drive dynamics of Bitcoin's economy [33].

We decided to apply automated Sentiment Analysis on shared short messages of users on Twitter in order to automatically analyze people's opinions, sentiments, evaluations and attitudes. We investigated whether public sentiment, as expressed in large-scale collections of daily Twitter

posts, can be used to predict the Bitcoin market. We tried to discover if the chatter of the community can be used to make qualitative predictions about Bitcoin market, attempting to establish whether there is any correlation between tweet's sentiment and the Bitcoin's price [3]. The results suggest that a significant relationship with future Bitcoin's price and volume of tweets exists on a daily level. We also used Google Trends to analyze Bitcoin's popularity under the perspective of Web search, which provides a time series index of the volume of queries made by users in Google Search. We found a striking correlation between Bitcoin's price spread and changes in query volumes for the "*Bitcoin*" search term.

The body of this paper is organized in five major sections. Section 2, describes the background, section 3 presents the research steps of our study and section 4 summarizes and discusses our results. Finally, section 5 presents conclusions and suggestions for future work.

4.2.2 Background

In these decades, social web has been commercially exploited for goals such as automatically extracting customer opinions about products or brands, to find which aspects are liked and which are disliked [79]. In their work, Ye and Wu demonstrate how particularly interesting is the influence of Twitter users and the propagation of the information related to their tweets[85].

Twitter had become the world's seventh most popular website by March 2015. Twitter is an online social networking website and microblogging service that allows users to post and read text-based messages of up to 140 characters, known as "tweets". Launched in July of 2006 by Jack Dorsey, Twitter is now in the top 10 most visited internet sites with a total amount of 645,750,000 registered users. Java et al. affirm that it seems to be used to share information and to describe minor daily activities [43]. The short format of a tweet is a defined characteristic of the service, allowing informal collaboration and quick information sharing. For business, Twitter can be used to broadcast company's latest news, posts, read comments of the customers or interact with them. A communicative feature of Twitter is the hashtag: a metatag beginning with the character #, designed to help others find a post.

Twitter is a rich source of real-time information regarding current societal trends and opinions. There are also studies that report another use of Twitter, namely as a possible predictor of market trends. Indeed, in 2010, a publication of the professor Johan Bollen showed that combining information on Wall Street with the millions of Tweets and posts, makes possible to anticipate financial performance [12]. In this work, Granger causality analysis and a Self-Organizing Fuzzy Neural Network are used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. The analysis of Tweets made by Bollen would have had 87% of chance to successfully predict prices of the stock, 3 or 4 days in advance. This study and analysis of millions of posts on Twitter represents a thermometer of emotions, on a large scale, which reflects the whole of society.

Earlier studies had found that blogs can be used to evaluate public mood,

[3] https://markets.blockchain.info/

and that tweets about movies can predict box office sales. Investigating the literature related to different uses of social media, and Twitter in particular, we collected information about the use of Twitter for seeking real world emotions that could predict real financial markets trend [44]. In their paper, Rao and Srivastava investigate the complex relationship between tweet board literature (like bullishness, volume, agreement etc) with the financial market instruments (like volatility, trading volume and stock price) [68].
The Bitcoin represents an important new phenomenon in financial markets. Mai et al. [54] examine predictive relationships between social media and Bitcoin returns by considering the relative effect of different social media platforms (Internet forum vs. microblogging) and the dynamics of the resulting relationships using vector autoregressive and vector error correction models.
In their work, Garcia et al. [33] show the interdependence between social signals and price in the Bitcoin economy, namely a social feedback cycle based on word-of-mouth effect and a user-driven adoption cycle. They provide evidence that Bitcoin's growing popularity causes an increasing search volumes, which in turn result a higher social media activity about Bitcoin. More interest inspire the purchase of bitcoins by users, driving the prices up, which eventually feeds back on the search volumes.
We compared Twitter's trending topic about Bitcoin with those in other media, namely, Google Trends. This is a feature of Google search engine that illustrates how frequently a fixed search term was looked for. Through this, you can compare up to five topics at one time to view relative popularity, allowing you to gain an understanding of the hottest search trends of the moment, along with those developing in popularity over time. Following this kind of approach, we evaluated how much *"bitcoin"* term, for the analyzed time interval, is looked for using Google's search engine.

4.2.3 Methodology

Sentiment Analysis

Tweets sometimes express opinions about different topics, and for this reason we decided to evaluate user's opinion about Bitcoin. We also investigated its power at predicting real-world outcomes. In order to evaluate if a user really appreciates the Bitcoin spread, we tried to predict sentiments analyzing tweets collection. In recent years, there is a wide collection of research surrounding machine learning techniques, in order to extract and identify subjective information in texts. This area is known as sentiment analysis or opinion mining [64]. Sentiment techniques are able to extract indicators of public mood directly from social media content [63].
Pang et al. argue that the research field of sentiment analysis has developed many algorithms to identify if the opinion expressed is positive or negative. In fact, algorithms to recognize sentiment are required to understand the role of emotions in informal communications [64]. Go et al. affirmed the strength of the sentiment analysis applied to the Twitter domain by using similar machine learning techniques to classifying the sentiment of tweets [36].
We chose to use automated sentiment analysis techniques to identify the sentiments of tweets in the matter of Bitcoin. Since the goal of this research

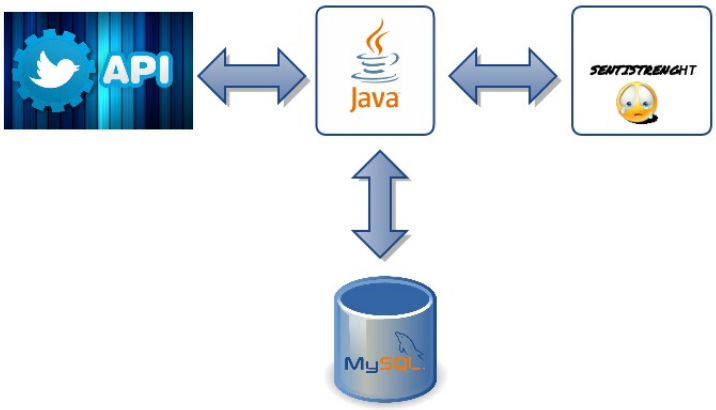

FIGURE 4.1: System Architecture

is neither to develop a new sentiment analysis nor to improve an existing one, we used "SentiStrenght", a tool developed by a team of researchers in the UK that demonstrated good outcomes [79]. SentiStrength estimates the strength of positive and negative sentiments in short texts. It is based on a dictionary of sentiment words, each one associated with a weight, which is its sentiment strength. In addition, this method uses some rules for non-standard grammar.
Based on the formal evaluation of this system on a large sample of comments from MySpace.com, the accuracy of predicting positive and negative emotions was something similar to that of other systems (72.8% for negative emotions and 60.6% for positive emotions, based on a scale of 1-5). Compared to other methods, SentiStrenght showed the highest correlation with human coders [80]. The tool is able to assess each message separately and, at the end, it returns one singular value: a positive, a negative or a neutral sentiment.

Data Collection

Tweets are available and are easily retrieved making use of Twitter Application Programming Interface (API). Composing the hashtag #Bitcoin or @bitcoin, we are able to gather all tweets that mentioned the analyzed subject. We briefly describe the different components of our system. An overview of this architecture is shown in Figure 4.1. The system consists of four components:

- *Twitter Streaming API*: it provides access to Twitter data, both public and protected, on a nearly real-time basis. A persistent connection is created between our system and Twitter. As soon as tweets come in, Twitter notifies our system in real time, allowing us to store them into our database.

- *DataStore*: our datastore consists of a back-end database engine, using MySQL as RDBMS, that repeatedly saves the incoming tweets from the Twitter Streaming API.

- *SentiStrenght tool*
- *Java Module*: this component allows us to send automated requests to Twitter Streaming API, to recover new tweets about Bitcoin, to parse the data gathered and to store them into our datastore. In a later stage, these data are sent to SentiStrenght tool in order to automatically evaluate the users' opinion.

We analyzed a collection of tweets, regarding Bitcoin, posted on Twitter between January 2015 and March 2015 (60 days). During this time 1,924,891 tweets were collected. The tweets were analyzed to determine its identifier, the date-time of the submission, its type, and its text content, which is limited to 140 characters. Comparing the timeline of tweets and the fluctuations in the Bitcoin market, we determined the specific day that provide a better correlation value. We then used SentiStrenght to evaluate comments extracted from Twitter. Given as input all tweets, the system assigned a score for each comment:

- 1 if the comment is positive
- -1 if the comment is negative
- 0 if the comment is neutral

4.2.4 Results

In order to decide the correct strategy of analysis for studying the relationship among Bitcoin's price and others meaningful parameters, the available related literature has been examined in depth. Most of articles [12] [44] [68] reports analysis about the existent relationship between the volume of tweets and the market evolution.
In general, Bollen et al. demonstrated that tweets can predict the market trend 3-4 days in advance, with a good chance of success. We analyzed the Bitcoin price's behavior comparing its variations with the number of tweets, with the number of tweets with positive mood, and with Google Trends results. The computation of cross-correlation yielded interesting results.
Our result seems to confirm that volumes of exchanged tweets may predict the fluctuations of Bitcoin's price. Furthermore, the comparison between tweets with a positive mood and trend of Bitcoin's price seems to prove this behavior. The examined literature shows different ways to highlight the existent relationship between big volumes of exchanged tweets and meaningful variations in the Bitcoin's price. Some papers show studies using regression methodology [54] or causality analysis [12]. Rao et al.[68] and Mittal et al.[58] showed how goods and stocks markets may be influenced by a big exchanged of tweet's volume. Inspired by these works, we tried to demonstrate how chatter of tweets might predict the price's variations of Bitcoin.
Figure 4.2 illustrates the curve trend of Bitcoin prices, expressed in dollars, and Twitter volume. We calculated the cross-correlation and, analyzing the results, we found that, in minimal degree, tweets volume is related to price with a maximum cross correlation value of 0.15 at a lag of 1 day (this is not very significant).

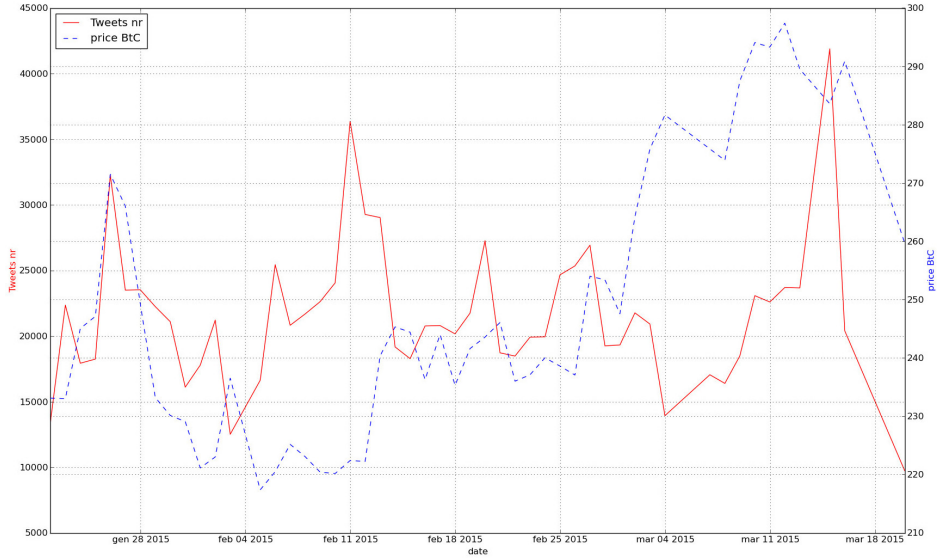

FIGURE 4.2: Similarity between Bitcoin's price and number of Tweets

TABLE 4.2: *Cross-correlation results*

Compared Systems	Cross-correlation value	delay
Bitcoin price-Tweets volume	0.15	1
Bitcoin price-Positive tweets	-0.35	3-4
Bitcoin price-Google Trends data	0.64	0

Nevertheless, if we observe Figure 4.2, we can notice how, also at a glance, there are peaks in tweets trend that precede peaks in price, suggesting a relationship between the two time series. A patent peak of tweets on 11 February, is followed by a growth of Bitcoin's price. The same circumstance is visible in the following days: January 23, February 3, February 25 and so on.
We also analyzed tweets with positive mood and we noticed a two-fold increase in cross-correlation value. Figure 4.3 shows this result and it's well rendered that positive tweets can predict the fluctuations of the Bitcoin's price. It is proven by a maximum cross correlation value of -0.35 with a positive delay of almost 4 days. We can confirm that positive mood could predict the Bitcoin's price almost 3-4 days in advance. All patent peaks in the positive tweets plot precede a significant change in the Bitcoin's price after some days.

The cross-correlation result between Google Trends data and Bitcoin's price also looks significant. The cross-correlation value increase up to a value of 0.64, that is quite substantial. This result is shown also by a little significant relationship that exists between positive tweets and Google Trends data.
Figure 4.4 shows how Google Trends proceeds in the same direction of Bitcoin's price and highlighting a striking similarity between them. Table 4.2 summarizes the cross-correlation results, obtained comparing the spread among Bitcoin price and different volumes of data.

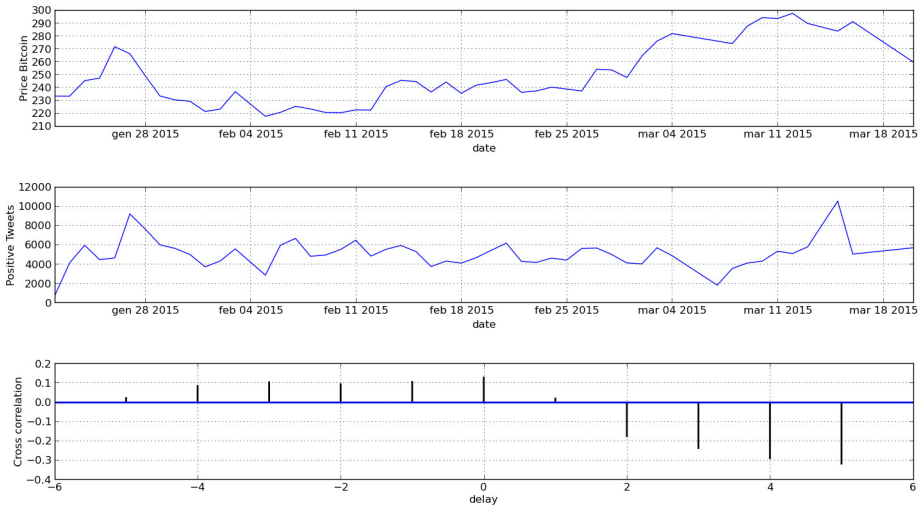

FIGURE 4.3: Cross-correlation between positive Tweets and Bitcoin's price

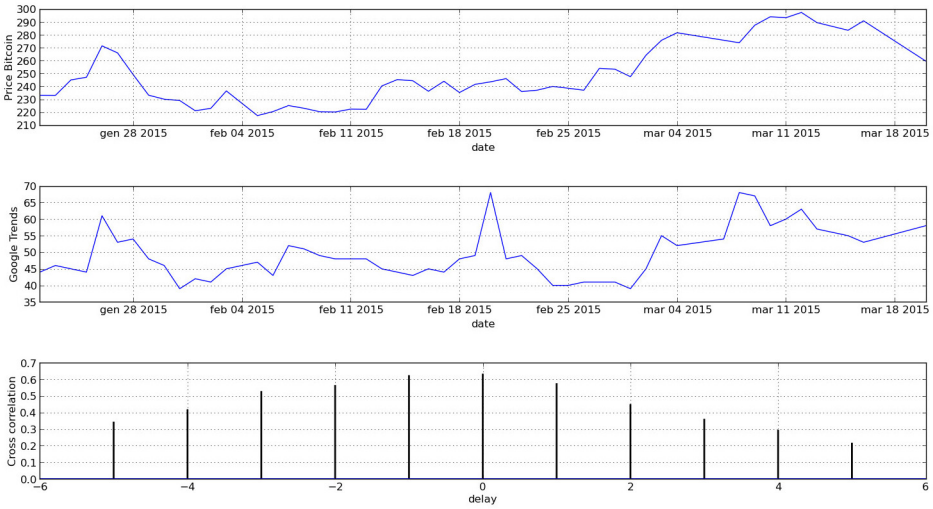

FIGURE 4.4: Cross correlation between Google Trends and Bitcoin's price, expressed in dollars

In this preliminary study, we examined the Bitcoin price's behavior comparing its variations with these of tweets volume, tweets with positive mood volume and Google Trends data. From results of a cross correlation analysis between these time series, we can affirm that positive tweets may contribute to predict the movement of Bitcoin's price in a few days. Google Trends could be seen as a kind of predictor, because of its high cross correlation value with a zero lag.
Our results confirm those found in the previous works, based on a different corpus of tweets and referred to a different Bitcoin market trend.
While the current data is only 60 days already looks promising, a consecutive analysis of more than 6 months might provide a better result quality. In further studies, we also plan to take into account the number of retweets and favorites for the tweet's corpus analyzed. Along these lines, we could check whether results stay unchanged with the addition of this variable.

4.3 The Predictor Impact of Web Search Media On Bitcoin Trading Volumes

4.3.1 Introduction

Internet has been one of the most revolutionary technologies in the last decades. The majority of daily activities radically changed, moving towards a “virtual sector", such as Web actions, credit card transactions, electronic currencies, navigators, games, etc. In recent years, web search and social media have emerged online. On one hand, services such as blogs, tweets, forums, chats, email have gained wide popularity. Social media data represent a collective indicator of thoughts and ideas regarding every aspect of the world. It has been possible to assist to deep changes in habits of people in the use of social media and social network [45].
Social media technologies have produced completely new ways of interacting [40], bringing the creation of hundreds of different social media platforms (e.g., social networking, shared photos, podcasts, streaming videos, wikis, blogs). On the other hand, due to the huge amount of available information, searching has become dominant in the use of Internet. Millions of users daily interact with search engines, producing valuable sources of interesting data regarding several aspects of the world.
Recent studies demonstrated that web search streams could be used to analyze trends about several phenomena [20] [72] [13]. In one of the most interesting works, Ginsberg et al. proved that search query volume is a sophisticated way to detect regional outbreaks of influenza in USA almost 7 days before CDC surveillance [35]. There are also studies that report another use in a search engine, namely as a possible predictor of market trends. Bollen et al. show that search volumes on financial search queries have a predictive power. They compared these volumes with market indexes such as Dow Jones Industrial Average, trading volumes and market volatility, demonstrating the possibility to anticipate financial performances [12]. In this work, Granger causality analysis and a Self-Organizing Fuzzy Neural Network are used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. Bordino et al. prove that search volumes of stocks highly correlate with trading volumes of the corresponding stocks, with peaks of search volume anticipating peaks of trading volume by one day or more [13].
Search queries prove to be a useful source of information in financial applications, where the frequency of searches of terms related to the digital currency can be a good measure of interest in the currency and it has a good explanatory power [47]. Mondria et al. proved that the number of clicks on search results stemming from a given country correlates with the amount of investment in that country [60]. Further studies showed that changes in query volumes for selected search terms mirror changes in current volumes of stock market transactions [67].
Technology always had a strong impact on financial markets and it has favored the emergence of Bitcoin, a digital currency created in 2008 by Satoshi Nakamoto [61]. It has been created for the purpose to replace cash, credit cards and bank wire transactions. It is based on advancements in peer-to-peer networks [71] and cryptographic protocols for security. Due to its

properties, Bitcoin is completely decentralized and not managed by any governments or bank, ensuring anonymity. It is based on a distributed register known as "block-chain" to save transactions carried out by users. Like any other currency, a peculiarity of Bitcoin is to facilitate transactions of services and goods with vendors that accept Bitcoins as payment[39], attracting a large number of users and a lot of media attention.
The Bitcoin represents an important new phenomenon in financial markets. Mai et al. examine predictive relationships between social media and Bitcoin returns by considering the relative effect of different social media platforms (Internet forum vs. microblogging) and the dynamics of the resulting relationships using auto-regressive vector and error correction vector models [54].
Matta et al. examined the striking similarity between Bitcoin price and the number of queries regarding Bitcoin recovered on Google search engine [56]. In their work, Garcia et al. [33] proved the interdependence between social signals and price in the Bitcoin economy, namely a social feedback cycle based on word-of-mouth effect and a user-driven adoption cycle. They provided evidence that Bitcoin's growing popularity causes an increasing search volumes, which in turn result a higher social media activity about Bitcoin. A growing interest inspires the purchase of Bitcoins by users, driving the prices up, which eventually feeds back on the search volumes.
There are several works that present predictive relationships between social media and bitcoin volume[4] where the relative effects of different social media platforms (Internet forum vs. microblogging) and the dynamics of the resulting relationships, are analyzed using cross-correlation [23] or linear regression analysis [12] [58]. Social factors, that are composed of interactions among market actors, may strongly drive the dynamics of Bitcoin's economy [33].
In this work we study the relationship that exists between trading volumes of Bitcoin currency and the queries volumes of search engine. The frequency of searches of terms about Bitcoin could be a good explanatory power, so we decided to examine Google, one of the most important search engine. We studied whether web search media activity could be helpful and used by investment professionals, analyzing the search volumes power of anticipate trading volumes of the Bitcoin currency.
We compared USD trade volumes about Bitcoin with those in a media, namely, Google Trends. This is a feature of Google search engine that illustrates how frequently a fixed search term was looked for. Following this kind of approach, we evaluated how much *"bitcoin"* term, for the specific time interval, is looked for using Google's search engine.

The body of this paper is organized in five major sections. Section 2, describes the research steps of our study, section 3 summarizes and discusses our results and, finally, section 4 presents conclusions and suggestions for future works.

[4] https://markets.blockchain.info/

4.3.2 Methodology

Google Trends

Google Trends [5] is a feature of Google Search engine that illustrates how frequently a fixed term is looked for. Through this, you can compare up to five topics at one time to view their relative popularity, allowing you to gain an understanding of the hottest search trends of the moment, along with those developing in popularity over time. The system provides a time series index of the volume of queries inserted by users into Google.
Query index is based on the number of web searches performed with a specific term compared to the total amount of searches done over time. Absolute search volumes are not illustrated, because the data are normalized on a scale from 0 to 100.
Google classifies search queries into 27 categories at the top level and 241 categories at the second level through an automatic classification engine. Indeed, queries are given out to fixed categories due to natural language processing methods.
The query index data are available as a CSV file in order to facilitate research purposes. Figure 4.9 depicts an example from Google Trends for the query “Bitcoin". We downloaded data about how much the term “Bitcoin" was referred to last year.

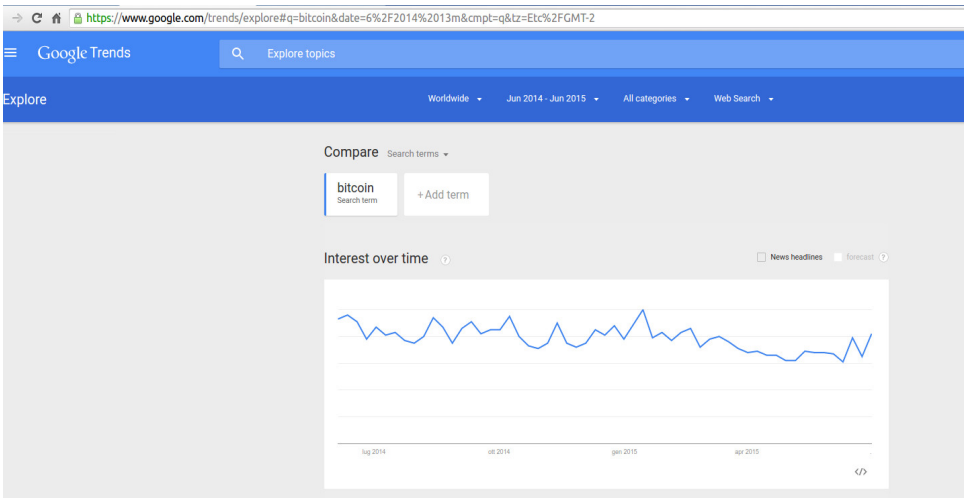

FIGURE 4.5: Example of Google Trends usage for the query “Bitcoin".

Blockchain.info

Blockchain.info[6] is an online system that provides detailed information about Bitcoin market. Launched in August 2011, this system shows data on recent transactions, plots on the Bitcoin economy and several statistics. It allows users to analyze different Bitcoin aspects:

- Total Bitcoins in circulation
- Number of Transactions

[5] http://trends.google.com
[6] http://www.blockchain.info

- Total output volume
- USD Exchange Trade volume
- Market price (USD)

We decided to study a time series regarding the USD trade volume from top exchanges, analyzing its trends.

Data Collection

Search query volumes regarding Bitcoin were collected from *Google Trends* website, capturing all searches, inserted from June 2014 to July 2015, with "Bitcoin" word as keyword .
Trading volume data were acquired from *blockchain.info* website, in order to evaluate daily trends of Bitcoin currency. We assessed the relationship over time between number of daily queries related to the trading volume of Bitcoin.
To better understand whether search engine can be seen as a good predictor of trading volumes, we applied an analysis of correlation between these data expressed in time series, a time-lagged cross-correlation study, concluding with a Granger-causality test.

4.3.3 Results

In order to decide the correct strategy of analysis for studying the relationship among Bitcoin's trading volume and others meaningful parameters, the available related literature has been examined in depth. Most of articles [12] [44] [68] reports analysis about the existent relationship between volume of media and market evolution. In general, Bollen et al. proved that tweets can predict market trend 3-4 days in advance, with a good chance of success. We extract from both data sources time series composed by daily values in the time interval ranging from June 2014 to July 2015 in order to evaluate their relationship and the capability of prediction. We run statistical analysis and the computation of correlation, cross-correlation and Granger causality test yielded interesting results.

Pearson Correlation

Pearson's correlation *r* is a statistical measure that evaluate the strength of a linear association between two time series *G* and *T*. We assumed G as query data and T as trading volumes.

$$r = \frac{\sum_i (G_i - \overline{G})(T_i - \overline{T})}{\sqrt{\sum_i (G_i - \overline{G})^2}\sqrt{\sum_i (T_i - \overline{T})^2}} \tag{4.1}$$

The correlations have values between -1 and +1, the bounds indicate maximum correlation and 0 indicating no correlation. A high negative correlation indicates a high correlation but of the inverse of one of the series. We calculated the Pearson correlation between queries search data and trading volume and we found a result equal to 0.60. This similarity is also clearly visible in the figure 2.

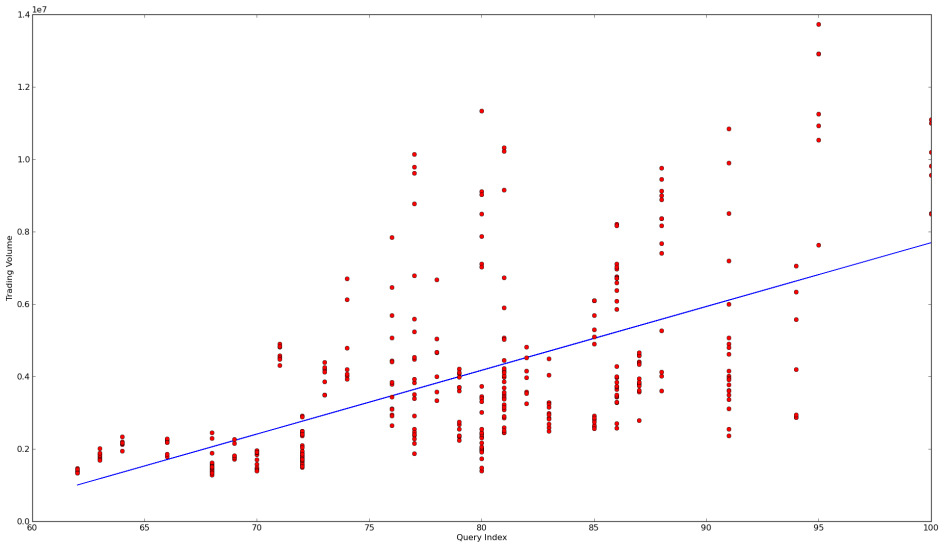

FIGURE 4.6: Correlation between Trading Volume and Queries Volume about Bitcoin.

Following this kind of analysis, we demonstrated the striking similarity existing between the time series. This result means that the trading volumes follows the same direction pace of queries volumes. Figure 4.6 reveals an obvious correlation due to peaks in one time series that occur close to peaks in the other. In this Figure it is possible to see that solid line, correspondent to search volumes, very often anticipated the dotted line correspondent to trading volumes. The most significant peaks occurred in the interval between August and September 2014, between September and October 2014, between November and December 2014 and between January and February 2015. During other periods the same phenomenon is less evident but anyway present.

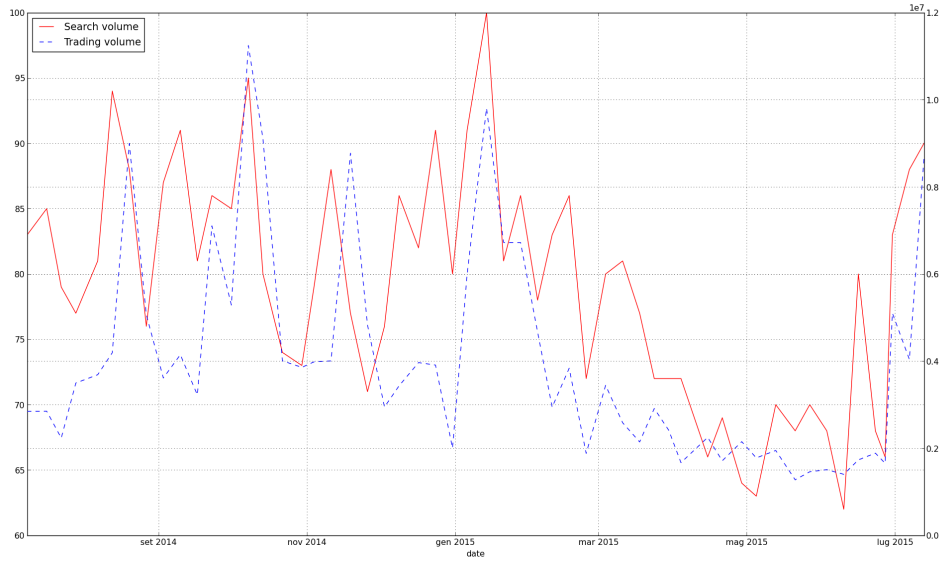

FIGURE 4.7: Correlation between Trading Volume and Queries Volume about Bitcoin.

Radical changes in peaks are due to several factors. One of the most evident peak is visible in Figure 4.7 corresponding to the interval between end of June and beginning of July. This is the period of the greek crisis acme, that causes changes also in the Bitcoin market. Indeed, a lot of people already started to invest in Bitcoin business. When people try to move money out of the country the government blocks this process, thus Bitcoin are the only way to transfer their wealth. In fact Greeks would use bitcoin to protect the value of their money at home. Ten times more Greek than usual are being recorded at the company *'German Bitcoin.de'*[7] to buy electronic currency. This situation is clearly visible in the right part of Figure 4.7, where curve correspondent to queries index volumes regarding Bitcoin considerably grew up, followed by an increase of curve correspondent to trading volumes after some days. In these mentioned cases it is clear how search volumes predict trading volumes preceding it, as confirmed by correlation values.

Cross Correlation

We investigated whether query volumes can anticipate trading volume of Bitcoin. We calculated the cross correlation values between query data G and trading volumes T as the time lagged Pearson cross correlation between two time series G and T for all delays d=0,1,2,..5.

$$r(d) = \frac{\sum_i (G_i - \overline{G})(T_{i-d} - \overline{T})}{\sqrt{\sum_i (G_i - \overline{G})^2}\sqrt{\sum_i (T_{i-d} - \overline{T})^2}} \tag{4.2}$$

We chose to evaluate a maximum lag of five days and, also in this case, the correlation ranges from -1 to 1. In Table 4.3, the results obtained from these experiments are reported. Each column shows the cross correlation result corresponding to different time-lag. We can observe that cross correlation results for positive delays are always higher than the ones with negative time lag. Indeed, the results with positive delays achieve values always higher than 0.64 and with negative delays report values always lower than 0.55. It means that query volumes is able to anticipate trading volumes in almost 3 days.

TABLE 4.3: *Cross-correlation results*

Delay	-5	-4	-3	-2	-1	0	1	2	**3**	4	5
Cross-C	0.36	0.40	0.44	0.50	0.55	0.60	0.64	0.67	**0.68**	0.67	0.64

Figure 4.8 shows the cross correlation results with a maximum lag of 30 days, just to highlight that the best result is given by a lag of almost 3.

Granger Causality

We performed a Granger causality test in order to verify whether web search queries regarding Bitcoin are able to anticipate particular trends in some days. The Granger-causality test is used to determine whether a time series G(t) is a good predictor of another time series T(t) [38]. If G Granger-causes T, then G^{past} should significantly help predicting T^{future} via T^{past} alone. We

[7] https://www.bitcoin.de/

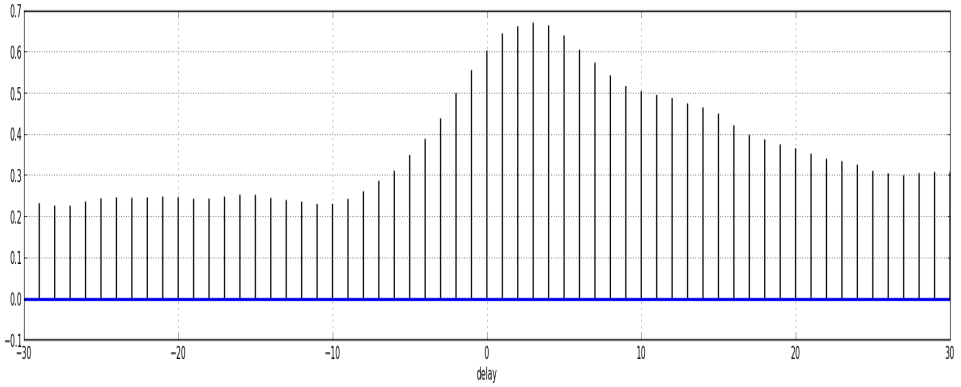

FIGURE 4.8: Cross Correlation results between Trading Volume and Queries Volume about Bitcoin with a maximum lag of 30 days.

compared query volumes G with trading volume T with the null hypothesis being that T is not caused by G. An F-test is then used to determine if the null hypothesis can be rejected.
We performed two auto-regression vectors as follows in the formula 3 and 4, where L represents the maximum time lag.

$$T(t) = \sum_{l=1}^{L} a_l T(t-l) + \epsilon_1 \tag{4.3}$$

$$T(t) = \sum_{l=1}^{L} a'_l T(t-l) + \sum_{l=1}^{L} b'_l G(t-l) + \epsilon_2 \tag{4.4}$$

We can affirm that G causes T if eq(4) is statistically better significant than eq(3). We applied the test in both directions, as an instance G $\rightarrow$ T means that the null hypothesis is “G doesn’t Granger-cause T".
Table 4.4 shows the results of the Granger causality test, where the first column represents the direction of the applied test, the second one the delay, and then the F-test result with its p-value. This parameter represents the probability that statistic test would be at least as extreme as observed, if the null hypothesis were true. So, we reject the null hypothesis if p-value is inferior to a certain threshold (p<0.05). Our analysis demonstrated that trad-

TABLE 4.4: *Granger-causality tests*

Direction	Delay	F-value Test	P-value
G→T	1	41.8135	p<0.001
	2	15.1435	p<0.001
	3	12.9332	p<0.001
	4	15.1546	p<0.001
	5	12.9279	p<0.001
T→G	1	0.5450	p=0.46
	2	2.3006	p=0.10
	3	1.4878	p=0.21
	4	1.5336	p=0.19
	5	1.2297	p=0.29

ing volumes can be considered Granger-caused by the query volumes. It is clearly shown that time-series G influences T, given by the p-value <0.001 for lags ranging from 1 to 5. So, the null hypothesis is completely rejected. On the other hand, the F-value test applied to the direction T→G reported a p-value always greater than 0.1. Trading volume T doesn't have significant casual relations with changes in queries volumes on Google search engine G. So, null hypothesis cannot be rejected.
From results of a cross correlation and Granger causality analysis between these time series, we can affirm that Google Trends is a good predictor, because of its high cross correlation value. Our results confirm those found in previous works, based on a different corpus and referred to a different Bitcoin market trend.
As future advancement, we are thinking about the possibility to apply this kind of approach to different contexts in order to better understand the predictive power of web search media. An other likelihood could be to consider not only search media but also social media like Twitter, Facebook and Google+.

4.4 Is Bitcoin's Market Predictable? Analysis of Web Search And Social Media

4.4.1 Introduction

The advent of the Internet has completely changed the way real life works. By enabling practically all Internet users to interact at once and to exchange and share information almost cost-free, more efficient decisions on several fields are possible.
The majority of daily activities radically changed, moving towards a "virtual sector", such as web actions, credit card transactions, electronic currencies, navigators, games, and so on. In recent years, web search and social media have emerged online. On one hand, services such as blogs, tweets, forums, chats, email have gained wide popularity. Social media data represent a collective indicator of thoughts and ideas regarding every aspect of the world. It has been possible to assist to deep changes in habits of people in the use of social media and social network [45]. Social media technologies have produced completely new ways of interacting [40], bringing the creation of hundreds of different social media platforms (e.g., social networking, shared photos, podcasts, streaming videos, wikis, blogs).
On the other hand, due to the huge amount of available information, searching has become dominant in the use of Internet. Millions of users daily interact with search engines, producing valuable sources of interesting data regarding several aspects of the world.
Recent studies demonstrated that web search streams could be used to analyze trends about several phenomena [20] [72] [13]. In one of the seminal works, Ginsberg et al. proved that search query volume is a sophisticated way to detect regional outbreaks of influenza in USA almost 7 days before CDC surveillance [35].
Kristoufek [48] studied the popularity of the Dow Jones stocks, measured by Google search queries for portfolio diversification. Curme et al. [24] clustered the online searches into groups and showed that mainly politics

and business oriented searches are connected to the stock market movements. Preis et al. [66] demonstrated that Google searches, for financial terms, can support profitable trading strategies. Dimpfl et al. found a strong relationship between internet search queries and the leading stock market index. In addition they found a strictly correlation between the Dow Jones' realised volatility and the volume of search queries [30].
There are also studies that report another use of a search engine, namely as a possible predictor of market trends. Bollen et al. showed that search volumes on financial search queries have a predictive power. They compared these volumes with market indexes such as Dow Jones Industrial Average, trading volumes and market volatility, demonstrating the possibility to anticipate financial performances [12]. In this work, Granger causality analysis and a Self Organizing Fuzzy Neural Network are used to investigate the hypothesis that public mood states, as measured by the Opinion Finder and GPOMS mood time series, are predictive of changes in DJIA closing values. Kristoufek proposed the study of Power-law correlations for Google searches queries for Dow Jones Industrial Average (DJIA) component stocks, and their cross-correlations with volatility and traded volume [49].
Bordino et al. proved that search volumes of stocks highly correlate with trading volumes of the corresponding stocks, with peaks of search volume anticipating peaks of trading volume by one day or more [13]. In his work [15], Bulut described that internet search data, via Google Trends, is utilized to nowcast the known variates of two structural exchange rate determinations models. By using internet search data, the author aims to get a timely description of the state of the economy way before the official data are released to the market participants. Kim et al. [46] introduced an analysis system to predict the value fluctuations of virtual currencies used in virtual worlds, and based on user opinion data in selected online communities. In their proposed method, data of user opinions on a predominant community are collected by employing a simple algorithm and guaranteeing a stable prediction of value fluctuations of more than one virtual currency.
Search queries prove to be a useful source of information in financial applications, where the frequency of searches of terms related to the digital currency can be a good measure of interest in the currency [47]. Mondria et al. proved that the number of clicks on search results stemming from a given country correlates with the amount of investment in that country [60]. Further studies showed that changes in query volumes for selected search terms mirror changes in current volumes of stock market transactions [67].
In recent years, social media data assume the role of a collective indicator of thoughts and ideas regarding every aspect of the world. It has been possible to assist to deep changes in habits of people in the use of social media and social network. In particular, we deeply analysed the transmitted sentiment of users regarding a particular topic. Twitter [8], an online social networking website and microblogging service, has become an important tool for businesses and individuals to communicate and share information with a rapid growth and significant adoption. In fact, Java et al. affirmed that it seems to be used to share data and to describe minor daily activities [43].

[8] https://twitter.com/

Twitter and other social media offer a plethora of opportunities to reveal business intuitions, where it remains a challenge to identify the potential social audience. In their work Ling et al. [53] analyzed the Twitter content of an account owner and its list of followers through various text mining methods and machine learning approaches in order to identify a set of users with high-value social audience members. In their paper, Ciulla et al. [21] assessed the usefulness of open source data that come from Twitter for prediction of societal events by analysing in depth the microblogging activity surrounding the voting behaviour on a specific event. Mocanu et al. performed a comprehensive survey of the worldwide linguistic landscape emerging from mining the Twitter microblogging platform [59]. Hick et al. explored the opportunities and challenges in the use of Twitter as platform for playing games, through crawling game that uses Twitter for collaborative creation of game content[42].
Additionally, Twitter has rapidly grown as a mean to share ideas and thoughts on investing decisions. Analyzing in deep the literature related to different uses of social media, and Twitter in particular, we collected information about its use for seeking real world emotions that could predict real financial markets trend [44]. In their paper, Rao and Srivastava studied the complex relationship that exists between tweet board literature (like bullishness, volume, agreement etc) with the financial market instruments (like volatility, trading volume and stock price) [68].
One of the fascinating phenomena of the Internet era is the emergence of digital currencies. Bitcoin, the most popular among these, has been created in 2008 by Satoshi Nakamoto [61] for the purpose to replace cash, credit cards and bank wire transactions. A digital currency can be defined as an alternative currency which is exclusively electronic and thus has no physical form.
Bitcoin is based on advancements in peer-to-peer networks [71] and cryptographic protocols for security. Due to its properties, Bitcoin is completely decentralized and is not managed by any government or central bank. Moreover, it ensures anonymity. So, it is practically detached from the real economy. Bitcoin is based on a distributed register known as "block-chain" to save transactions carried out by users. Like any other currency, a peculiarity of Bitcoin is to facilitate transactions of services and goods with vendors that accept Bitcoins as payment [39], attracting a large number of users and a lot of media attention.
The Bitcoin represents an important new phenomenon in financial markets. Mai et al. examined predictive relationships between social media and Bitcoin returns by considering the relative effect of different social media platforms (Internet forum vs. microblogging) and the dynamics of the resulting relationships using auto-regressive vector and error correction vector models [54]. Matta et al. examined the striking similarity between Bitcoin price and the number of queries regarding Bitcoin recovered on Google search engine [56]. In their work, Garcia et al. [33] proved the interdependence between social signals and price in the Bitcoin economy, namely a social feedback cycle based on word-of-mouth effect and a user-driven adoption cycle. They provided evidence that Bitcoins growing popularity causes an increasing search volumes, which in turn result a higher social media activity about Bitcoin. A growing interest inspires the purchase of Bitcoins by users, driving the prices up, which eventually feeds back on the search

volumes. There are several works that present predictive relationships between social media and Bitcoin volumes where the relative effects of different social media platforms (Internet forum vs. microblogging) and the dynamics of the resulting relationships, are analyzed using cross-correlation [23] or linear regression analysis [58]. Social factors, that are composed of interactions among market actors, may strongly drive the dynamics of Bitcoin's economy [33].
In this work we decided to investigate whether social media activity or information collected by web search media could be profitable and used by investment professionals. We also evaluated the possibility to find a relationship between Bitcoins trading volumes and volumes of exchanges tweets.
We first studied the relationship that exists between trading volumes of Bitcoin currency and the volumes of search engine, then we analyzed a corpus of 2,353,109 tweets in order to discover if the chatter of the community can be used to make qualitative predictions about Bitcoin market, attempting to establish whether there is any correlation between tweet's sentiment and the Bitcoin's trading volume. The frequency of searches of terms about Bitcoin could have a good explanatory power, so we decided to examine Google, one of the most important search engine. We studied whether web search media activity could be helpful and used by investment professionals, analyzing the search volumes power of anticipate trading volumes of the Bitcoin currency.
We compared USD trade volumes about Bitcoin with search volumes using Google Trends. This is a feature of Google search engine that illustrates how frequently a fixed search term was looked for. Following this kind of approach, we evaluated how much "Bitcoin" term, for the specific time interval, is looked for using Google's search engine.
Simultaneously, we decided to apply automated Sentiment Analysis on shared short messages of users on Twitter in order to automatically analyze people's opinions, sentiments, evaluations and attitudes. We wondered whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can be used to predict the Bitcoin market. The results of our previous analysis suggest that a significant relationship with future Bitcoin's price and volume of tweets exists on a daily level. We found a striking correlation between Bitcoin's price spread and changes in query volumes for the "Bitcoin" search term [56].
The body of this paper is organized in five major sections. Section 2, describes the methodology applied in our study, section 3 summarizes and discusses our results and, finally, section 4 presents conclusions and suggestions for future work.

4.4.2 Methodology

Google Trends

Google Trends [9] is a feature of Google Search engine that illustrates how frequently a fixed term is looked for. Through this, you can compare up to five topics at one time to view their relative popularity, allowing you to gain an understanding of the hottest search trends of the moment, along

[9] http://trends.google.com

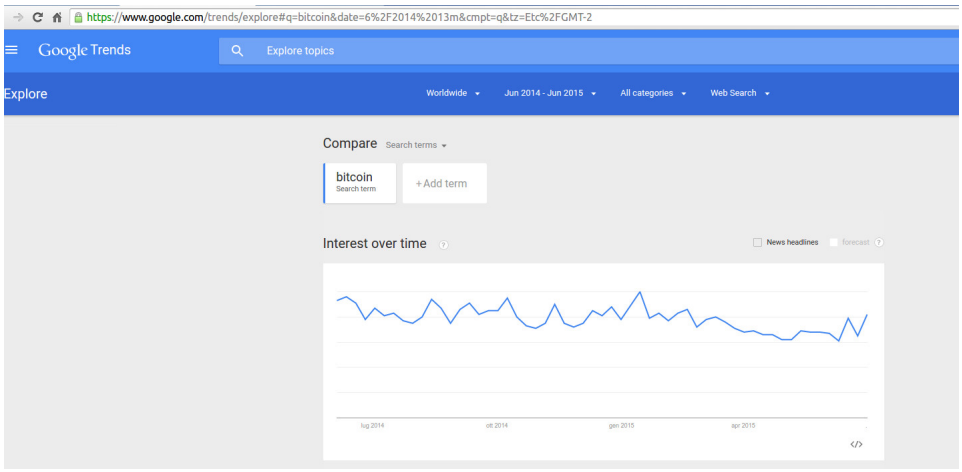

FIGURE 4.9: Example of Google Trends usage for the query "Bitcoin"

with those developing in popularity over time. This system provides a time series index of the volume of queries made by the users with Google.
Query index is based on the number of web searches, performed with a specific term, and compared to the total amount of searches done over time. Absolute search volumes are not shown, because the data are normalized on a scale from 0 to 100.
Google classifies search queries into 27 categories at the top level and 241 categories at the second level through an automatic classification engine. Indeed, queries are given out to fixed categories, due to natural language processing methods.
The query index data are available as a CSV file in order to facilitate research activities. Figure 1 depicts an example from Google Trends for the query "Bitcoin".

Blockchain.info

Blockchain.info[10] is an online system that provides detailed information about Bitcoin market. Launched in August 2011, this system shows data on recent transactions, plots on Bitcoin economy and several statistics. It allows users to analyze different Bitcoin aspects:

- Total number of Bitcoins in circulation
- Number of Transactions
- Total output volume
- USD Exchange Trade volume
- Market price (USD)

We decided to study a time series regarding the USD trade volume from top exchanges, analyzing its trends.

[10]http://www.blockchain.info

Twitter API

The Twitter space can be explored by means of its Application Programming Interface (API) [11], implementing a simple crawler that allows developers to collect the required data. Twitter servers send back XML or JSON responses, that are parsed and processed by the implemented system. Twitter gives the opportunity to work with its API with three different approaches that are listed below.

- *Original REST API* allows users to analyze the Twitter core, in order to update their statuses or profile information in realtime.
- *Search API* is read-only in the search database, and search queries return up to 1500 tweets from up to seven/eight days before.
- *Streaming API* provides real time data access to tweets in filtered forms that return public statuses that match with one or more filters. This approach doesn't allow developers to find all the historical tweets, but it enables the access to data as it is being tweeted.

While Twitter REST API is available and suitable for different applications, we decided to monitor tweets with Twitter Streaming API, that provides immediate updates. As soon as tweets come in, Twitter notifies the implemented system allowing us to store them into the database without the delay of polling the REST API.

Opinion Mining

The opinion mining is a particular technique that detects automatically the sentiment and subjectivity transmitted in written texts. The user's tweets could express the opinion regarding different topic, trends or brands [64]. For this reason, we decided to monitor the sentiment expressed, day after day, by users on the matter of Bitcoin.
Since the goal of this research is neither to develop a new sentiment analysis nor to improve an existing one, we used "SentiStrenght", a tool developed by a team of researchers in the UK that demonstrated good outcomes [80]. SentiStrength estimates the strength of positive and negative sentiments in short texts. It is based on a dictionary of sentiment words, each one associated with a weight, which is its sentiment strength. In addition, this method uses some rules for non-standard grammar.
Based on the formal evaluation of this system on a large sample of comments from MySpace.com, the accuracy of predicting positive and negative emotions was something similar to that of other systems (72.8% for negative emotions and 60.6% for positive emotions, based on a scale of 1-5). Compared to other methods, SentiStrenght showed the highest correlation with human coders [**mike**]. The tool is able to assess each message separately and, at the end, it returns one singular value.

- +1 if the system identifies a positive sentiment
- -1 if the system identifies a negative sentiment
- 0 if a neutral opinion is identified

[11] https://dev.twitter.com/overview/documentation

Data Collection

Search query volumes regarding Bitcoin were collected from *Google Trends* website, capturing all searches, inserted from June 2014 to July 2015, with "Bitcoin" word as keyword .
Trading volume data were gathered from *blockchain.info* website, in order to evaluate daily trends of Bitcoin currency. We assessed the relationship over time between number of daily queries and trading volume of Bitcoin.
We collected a dataset of tweets regarding Bitcoin in the period between January and April 2015 using *Streaming Twitter API*, achieving almost two millions of statuses. The system has been set to raise in real time the tweets that contain "Bitcoin" word in the sentence. Twitter API provides several fields in JSON format. We decided to hold the following tweet's information, saving it in our database.

- Tweet ID
- Tweet text
- Date of creation
- User who posted the status
- User location

For each collected tweet, we detected the language in order to use the correct sentiment dictionary. "Language-detection" is a particular library, implemented in Java language, able to identify the language of a given sentence using Naive Bayesian filter [12]. The system has been tested, achieving a 99% precision for 53 languages. After this step, the sentiment was calculated using the correct language dictionary by means of SentiStrength.
To better understand whether a search engine can be a good predictor of trading volumes, we analysed the correlation between these data expressed as time series, performing a time-lagged cross-correlation study, and a Granger-causality test. We applied the same analysis to the number of tweets, and to the sentiment expressed by users. These analyses based on Twitter have been applied on a shorter period than the search media analysis.

4.4.3 Results

We extracted from Google Trends and Blockchain.info data sources time series composed by daily values in the time intervalranging from June 2014 to July 2015, in order to evaluate their relationship and prediction capability. The same approach has been applied to the Twitter, dataset but in the time interval ranging from January 2015 to April 2015. We run statistical analysis and the computation of correlation, cross-correlation and Granger causality test, obtaining interesting results.

Tweets Analysis

We collected 2,353,109 tweets covering the period between January and April 2015. Using Sentistrength, we computed the sentiment of each tweet,

[12] https://code.google.com/p/language-detection/

obtaining 418,949 positive tweets and 270,669 negative ones. The remaining tweets are neutral. There are more positive messages than negative ones. We also found that positive messages are almost 2 times more likely to be forwarded than negative ones. After a careful analysis, it was observed that the number of neutral tweets is very high because people very often write non-expressive comments, the price of Bitcoin or simple links that lead to other web pages.
Figure 4.10 shows the two time series for the period under consideration, representing the positive tweets with a dotted line and the negative tweets with a solid line. Taking a look to the two time series it's possible to see some negative or positive peaks, corresponding to price variations [13]. For instance, the peak of January 26^{th} is due to the top price of the Bitcoin for the same day, 278$. An other example can be seen on 12 February when there is a negative peak, corresponding to a price decrease at 221.85 dollars. This Figure clearly shows that, most of the times, positive and negative time series grow up and decrease with the same pace in a given day. This is related to the total amount of tweets of the evaluated day. To solve this problem, we developed a simple metric called PT-NT ratio x to predict the trends of the Bitcoin trading volume. We defined the sentiment score x_t of day t as the ratio of positive versus negative messages on the Bitcoin topic. A message is defined as positive if it contains more positive than negative words, and negative in the opposite case.

$$x_t = \frac{count_t(pos.tweet \wedge Bitcoin\ topic)}{count_t(neg.tweet \wedge Bitcoin\ topic)} \tag{4.5}$$

$$= \frac{p(pos.tweet \mid Bitcoin\ topic, t)}{p(neg.tweet \mid Bitcoin\ topic, t)} \tag{4.6}$$

With this approach, it's possible to determine the ratio of positive versus negative tweets on a fixed day. The resulting time series were used to study their correlation with the Bitcoin trading volume over a given period of time.

Pearson Correlation

Pearson's correlation *r* is a statistical measure that evaluate the strength of a linear association between two time series *G* and *T*. Initially, we assumed G as query data and T as trading volumes.

$$r = \frac{\sum_i (G_i - \overline{G})(T_i - \overline{T})}{\sqrt{\sum_i (G_i - \overline{G})^2}\sqrt{\sum_i (T_i - \overline{T})^2}} \tag{4.7}$$

The Pearson correlation coefficient has values between -1 and +1, the bounds denoting maximum anti-correlation or correlation, respectively, whereas 0 indicates no correlation. We calculated the Pearson correlation between queries search data and trading volume and we found a result equal to 0.60, which is quite high. The correlation is also clearly visible in the Figure 4.11.

This result is confirmed by Figure 4.12, that shows the two time series. Here peaks in one time series typically occur close to peaks in the other. The solid line, that represents search volumes, very often anticipates the dotted

[13] https://blockchain.info/it/charts/market-price

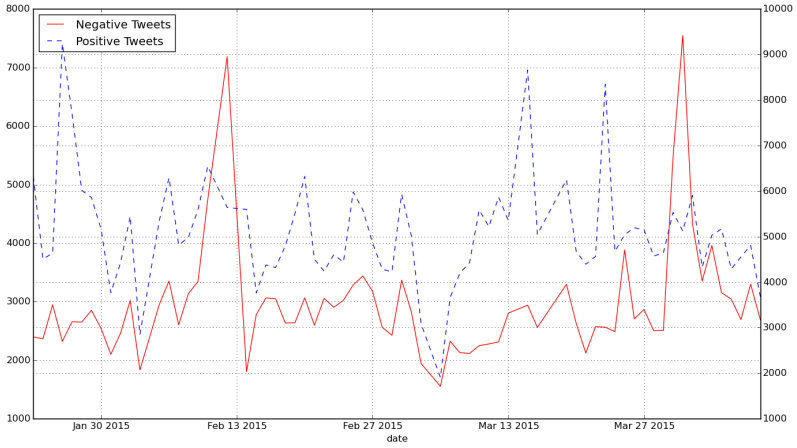

FIGURE 4.10: Representation of the positive tweets with a dotted line and the negative tweets with a solid line for the period between January and April 2015

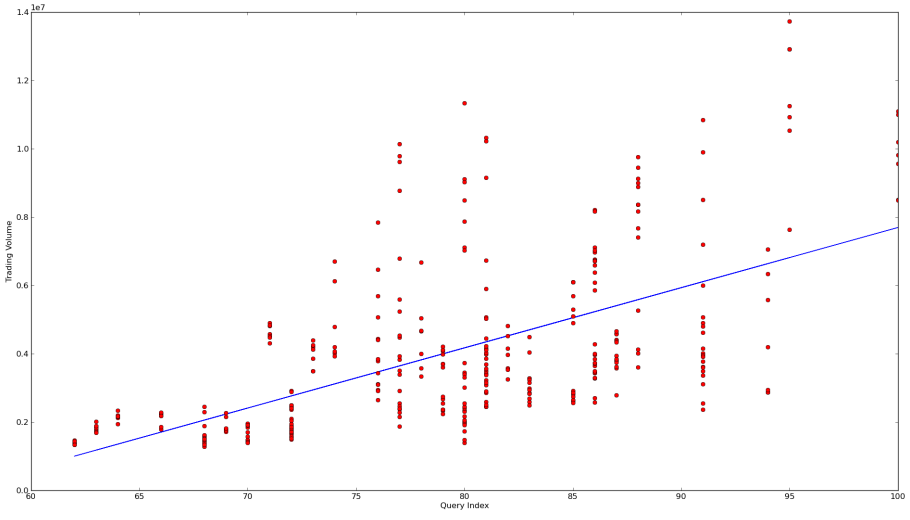

FIGURE 4.11: Correlation between Trading Volume and Queries Volume about Bitcoin

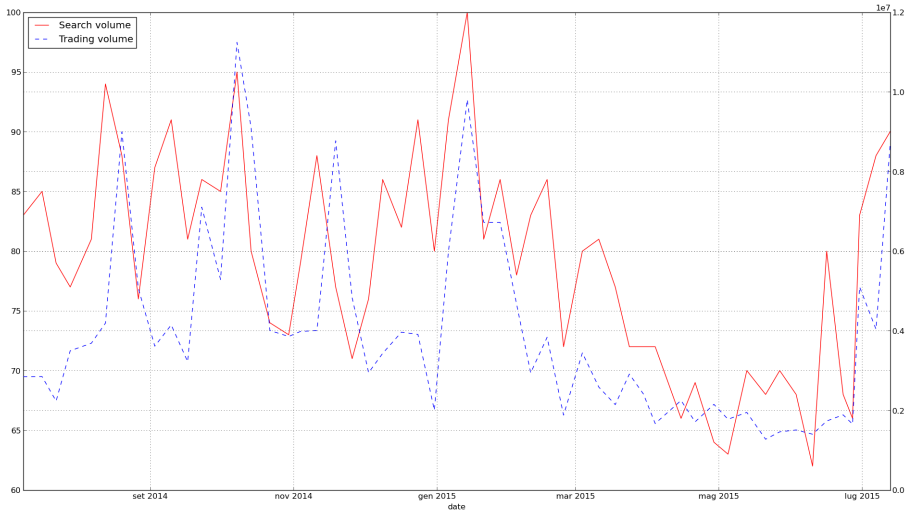

FIGURE 4.12: Bitcoin Trading Volume and Queries Volume about Bitcoin.

line, that represents trading volumes. The most significant peaks occurred in the interval between August and September 2014, between September and October 2014, between November and December 2014 and between January and February 2015. During other periods, this phenomenon is less evident, but anyway it is present.
Radical changes in peaks are due to several factors. One of the most evident peaks in Figure 4.12 corresponds to the interval between end of June and beginning of July. This is the period of the Greek crisis acme, that caused changes also in the Bitcoin market. Indeed, a lot of people started to invest in Bitcoin because people tried to move money out of the country, and Greek government tried to block this process. Bitcoins were seen by many as the only way to move their wealth to other currencies.
In fact, Greeks would use Bitcoin to protect the value of their money at home. Ten times more Greek than usual were found at the company *'German Bitcoin.de'*[14] to buy electronic currency. This situation is clearly visible in the right part of Figure 4.12 , where the curve corresponding to the volumes of queries regarding Bitcoin considerably increases, followed by an increase of the curve of trading volumes after some days. This is confirmed by correlation values.
Comparing the PT-NT ratio of tweets volume with the trading volume of Bitcoins, we found a Pearson correlation equals to 0.37, which is still a remarkable value. The two time series are again quite similar, and this is shown in Figure 4.13, that shows a fair correlation, with some points far from the line of best fit. Figure 4.14 highlights the period between January and February 2015, where we can notice how social peaks correspond to following peaks in the trading volume. This is particularly evident on January, 25 and on February, 14. The prediction power of the PT-NT ratio on Bitcoin trading volumes, however, is lower than that of web search volumes.

[14]https://www.bitcoin.de/

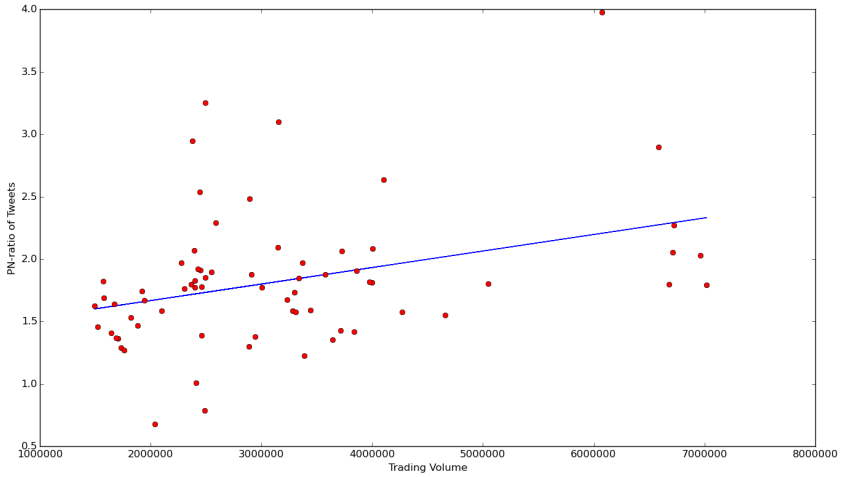

FIGURE 4.13: Pearson Correlation coefficient between Trading Volume and the Volume of Tweets about Bitcoin.

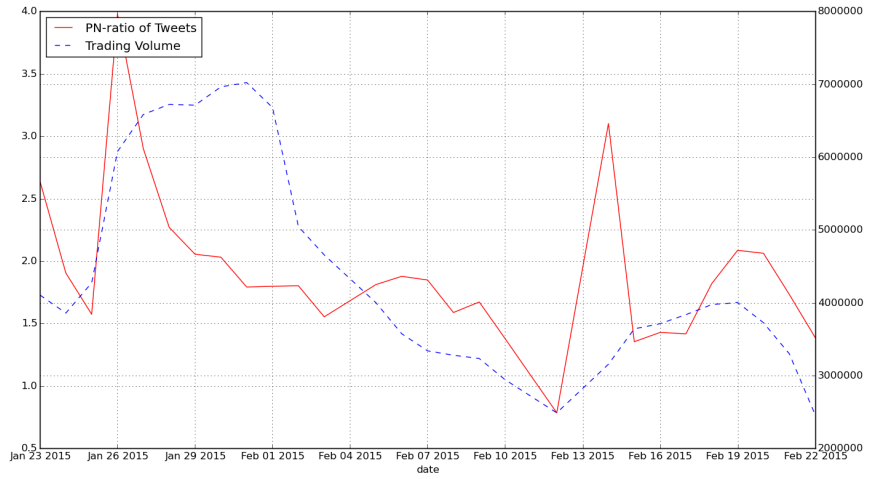

FIGURE 4.14: Bitcoin Trading Volume and PT-NT ratio in the period January, 23 to February, 22, 2015.

TABLE 4.5: *Cross-correlation with the trading volume of Bitcoin, compared to the search volume, and to the PT-NT ratio*

Delay	-4	-3	-2	-1	0	1	2	3	4
Search	0.402	0.447	0.502	0.559	0.609	0.649	0.670	**0.682**	0.674
PT-NT Ratio	0.122	0.165	0.169	0.221	0.353	**0.379**	0.374	0.373	0.366

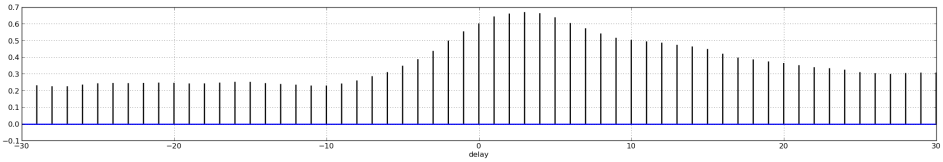

FIGURE 4.15: Cross-Correlation between Trading Volume and Queries Volume about Bitcoin, with a maximum lag of 30 days

Cross Correlation

We investigated whether query or search volumes can anticipate trading volumes of Bitcoin. We calculated the cross correlation comparing the trading volume T with the query volume G as the time lagged Pearson cross correlation between two time series G and T for all delays d between -5 and 5. We also applied the same analysis, substituting the search data with the social data, according to eq. 4.8 .

$$r(d) = \frac{\sum_i (G_i - \overline{G})(T_{i-d} - \overline{T})}{\sqrt{\sum_i (G_i - \overline{G})^2} \sqrt{\sum_i (T_{i-d} - \overline{T})^2}} \tag{4.8}$$

We chose to evaluate a maximum lag of five days and, also in this case, the correlation ranges from -1 to 1.
In Table 4.5 we report the results obtained from these experiments. Each column shows the cross-correlation corresponding to different time lags. We can observe that cross-correlation for positive lags is always higher than for negative lags. Taking a look to the raw search volume, the results with positive delays take values always higher than 0.64, whereas those with negative delays take values always lower than 0.55. This means that query volumes are able to anticipate trading volumes of 3 days, or even more
Different outcomes are visible in the bottom raw, corresponding to the the PT-NT ratio social volume, where the highest cross correlation result is given with a delay equal to 1. It means, that the social volume is able to anticipate the trading volume in one or two days with 0.37 as the best result. The results with positive delays achieve outcomes always higher than 0.35 and with negative delays report values always lower than 0.22. Although we achieved good outcomes with both search and Twitter functions, the search volume has better predictive power.

Figure 4.15 and 4.16 show the cross correlation results with a maximum lag of 30 days, just to highlight that the media volume (social or search) anticipates the trading volume and that the best result is given by a lag of almost 3 in the Figure 4.15 and by a lag of one in the other picture.

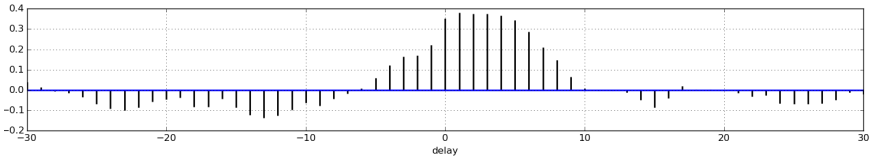

FIGURE 4.16: Cross-Correlation between Trading Volume and PT-NT ratio of social volume about Bitcoin, with a maximum lag of 30 days

Granger Causality

We performed a Granger causality test in order to verify whether web search queries or the PT-NT ratio of social volume, regarding Bitcoin, are able to cause particular trends in some days. The Granger-causality test is used to determine whether a time series G(t) is a good predictor of another time series T(t) [38]. If G Granger-causes T, then G^{past} should significantly help predicting T^{future} via T^{past} alone.
We compared the two media volumes G, one after the other, with trading volume T with the null hypothesis being that T is not caused by G. An F-test is then used to determine if the null hypothesis can be rejected.
We computed two auto-regression vectors as follows in the eqs. 4.9 and 4.10, where L represents the maximum time lag.

$$T(t) = \sum_{l=1}^{L} a_l T(t-l) + \epsilon_1 \tag{4.9}$$

$$T(t) = \sum_{l=1}^{L} a'_l T(t-l) + \sum_{l=1}^{L} b'_l G(t-l) + \epsilon_2 \tag{4.10}$$

We can affirm that G causes T if eq. 6 is statistically better significant than eq. 5. We applied the test in both directions, as an instance G $\rightarrow$ T means that the null hypothesis is "G doesn't Granger-cause T".
Table 4.6 and Table 4.7 show the results of the Granger causality test, applied to the different media, social and web search. The first column represents the direction of the applied test, the second one the delay, and then the F-test result with its p-value. This parameter represents the probability that statistic test would be at least as extreme as observed, if the null hypothesis were true. So, we reject the null hypothesis if p value is inferior to a certain threshold ($p < 0.05$).

Table 4.6 demonstrates that trading volumes can be considered Granger-caused by the query volumes. It is clearly shown that time-series G influences T, given by the p value <0.001 for lags ranging from 1 to 5. So, the null hypothesis is strongly rejected. On the other hand, the F-value test applied to the direction T$\rightarrow$G reported a p-value always greater than 0.1. Trading volume T doesn't have significant casual relations with changes in queries volumes on Google search engine G. So, null hypothesis cannot be rejected.
Table 4.7 shows that the Granger causality tests report a p value always higher than 0.25, meaning that the null hypothesis cannot be rejected, so the PT-NT ratio of social volume cannot cause the trading volume. Also in this case, the main problem could be the short temporal period that we took in account, not enough to evaluate the predictive power of anticipation or

TABLE 4.6: *Granger-causality tests between trading volume T and web search volume G*

Direction of Causality	Delay	F-value Test	P-value
G→T	1	41.8135	p<0.001
	2	15.1435	p<0.001
	3	12.9332	p<0.001
	4	15.1546	p<0.001
	5	12.9279	p<0.001
T→G	1	0.5450	p=0.46
	2	2.3006	p=0.10
	3	1.4878	p=0.21
	4	1.5336	p=0.19
	5	1.2297	p=0.29

TABLE 4.7: *Granger-causality tests between PT-NT ratio of social volume G and trading volume T*

Direction of Causality	Delay	F-value Test	P-value
G→T	1	2.5175	p=0.1173
	2	0.1210	p=0.8863
	3	0.2512	p=0.8602
	4	1.3807	p=0.2519
	5	1.1938	p=0.3243
T→G	1	1.3501	p=0.24
	2	0.4841	p=0.61
	3	0.6937	p=0.55
	4	0.0899	p=0.98
	5	0.7991	p=0.55

causation. Nevertheless, we can confirm that search volume is the best predictor, able to cause changes in the trading volume.
As future improvements, we are working on the possibility to apply a similar approach to other contexts, in order to better understand the predictive power of web search and social media. We are also working to extend our analysis to evaluate the correlation of Bitcoin Market Volatility.

Chapter 5

Agile Project Management Tools prediction

5.1 Agile

Agile development methodology provides opportunities to evaluate the evolution of a project during the development life cycle. This is evaluable through fixed cadences of work, well known as iterations or sprints, where each component of the team could show any kind of project increment. A software project is composed by a collection of activities, able to return a certain outcome. The project management consists of planning, executing, and checking these activities. Anyway, the high costs and the time required is really interesting for researchers and practitioners .
A lot of new organizations are looking for competitive advantages making use of several deployments of Internet-based services. For this reason, software developers are under increasing pressure to produce new and innovative implementations very quickly [3] [25].
Agile software development processes were created to solve this kind of problem, utilizing technical and managerial processes that continuously adapt changes from different experiences gained during the development phase. Agile process support the production during the working code phase. This is possible structuring the development process into iterations, where each iteration focuses on the working code that provide value to the customers and to the project.
Agile methodology is becoming really popular during the software development of a project. In 2001, the Agile Manifesto has been published to lead the realisation of the project with agile practices [1].
The main aim of the Agile Methodologies is to increase the ability to react to changing customers, business and technological needs. Several companies are moving toward agile software development in order to improve quality and productivity.
In spite of the huge amount of available tools that support Agile functions, little is known about the habits and needs of the software companies during the tools usage. Teams and companies follow the agile processing making use of agile tools. There is a huge amount of tools to choose from, but the companies need could be different. Nowadays, there is a lack of surveys regarding Agile tools usage but nothing regarding the Agile tools most useful.

5.1.1 Agile Development

The agile development is defined as a set of different principles and methodologies that lead toward iterative and incremental development. The iterative development is a kind of development where a portion of a project is build dividing the life-cycle into small sequences back to back in time. These kind of iterations keeps each almost 2 weeks with a maximum of six weeks. The majority of the activities, like analysis, design, implementation and testing are repeated in each sprint [52].
An incremental development is the consequent phase, due to the grown of the system that adds new features to the development process of the project. The Agile Manifesto was written by a team of developers that already tried the several methodologies. It mainly consists of twelve principles [1]:

1. The highest priority is to satisfy the customer through early and continuous delivery of valuable software.
2. Welcome changing requirements, even late in development. Agile processes harness change for the customer's competitive advantage.
3. Deliver working software frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale.
4. Business people and developers must work together daily throughout the project.
5. Build projects around motivated individuals. Give them the environment and support they need, and trust them to get the job done.
6. The most efficient and effective method of conveying information to and within a development team is face-to-face conversation.
7. Working software is the primary measure of progress.
8. Agile processes promote sustainable development. The sponsors, developers, and users should be able to maintain a constant pace indefinitely.
9. Continuous attention to technical excellence and good design enhances agility.
10. Simplicity–the art of maximizing the amount of work not done–is essential.
11. The best architectures, requirements, and designs emerge from self-organizing teams.
12. At regular intervals, the team reflects on how to become more effective, then tunes and adjusts its behaviour accordingly.

In order to define daily development work there exist different kind of practices to achieve agility, based on methods and methodologies, like Scrum, Extreme-Programming, Feature-Driven development and so on.

SCRUM

Scrum is one of the most popular Agile software development model created to help multiple small teams that hardly develop software systems. The term 'Scrum' represents the formation in rugby used to restart the game after the event that caused the stop play.
The project work is divided in fixed-length iterations or sprints, in order to release the product software respecting a regular cadence. Each sprint starts with a Sprint planning, where the work to do during the sprint is decided and is concluded by a Sprint review, where a prefixed demo is shown to the stakeholders.
An important meeting is the daily Scrum, where doubts regarding the development of the project are solved. The sprint retrospective is a particular phase where the team considers new aspects in order to adjust something during the next phase of the project. The product backlog is a prioritized list of the final product requirements, which contains a subset with the estimated items to be done during the current sprint.
The backlog is completely managed by a Product Owner, responsible for the state of progress of the product. An other important role is the Scrum Master that verifies if the Scrum theory and rules are respected and understood [75].

EXTREME PROGRAMMING

Extreme programming is a model based on values of communication, respect, courage and simplicity [8]. It can be seen as a set of principles, values, rules to develop in the fastest way possible high quality of the software. XP is based on 12 well-known practices:

1. The Planning Game: Business and development work together to produce the maximum business value as rapidly as possible.

2. Small Releases: It starts with the smallest useful feature collection. The release takes place early and often, adding a few features each time.

3. System Metaphor: Each project has an organizing metaphor, which provides an easy way to remember naming convention.

4. Simple Design: Always use the simplest possible design that gets the work done.

5. Continuous Testing: Before software developers add a feature, they write a test for it. When the suite runs, the job is done. Two kind of tests are available XP.

6. Unit Tests are automated tests written by the developers to test functionality as they write it. Each unit test typically tests only a single class, or a small cluster of classes. Unit tests are typically written using a unit testing framework, such as Junit.

7. Acceptance Tests (also known as Functional Tests) are specified by the customer to test that the overall system is functioning as specified. Acceptance tests typically test the entire system, or some large chunk

of it. When all the acceptance tests pass for a given user story, that story is considered complete.

Refactoring: Refactor out any duplicate code generated in a coding session.

Pair Programming: All production code is written by two programmers sitting at one machine and all code is reviewed as it is written.

8. Collective Code Ownership: No single person "owns" a module. Any developer is expect to be able to work on any part of the codebase at any time.
9. Continuous Integration: All changes are integrated into the codebase at least daily. The tests have to run 100% both before and after integration.
10. 40-Hour Work Week.
11. On-site Customer: Development team has continuous access to a real live customer, that is, someone who will actually be using the system. For commercial software with lots of customers, a customer proxy (usually the product manager) is used instead.
12. Coding Standards: Everyone codes to the same standards.

LEAN AND KANBAN

Lean is a customer-centric methodology used to maximize the customer value while minimizing waste with fewer resources. This is based on the reduction of the waste, control of variability and flow optimization of delivered software. Lean achieves value to the customer more efficiently by removing waste. This situation is well explained in a set of basic principles that are shown below [65].

- Focus on the important value of the customer
- Respect for the people involved in the project.
- Removal of any kind of waste
- Optimization of the flow
- Creation of knowledge
- Use of empirical methods to take important decisions.

The Lean approach requires to divide a process into individual steps, identifying which ones add waste (muda). So, the aim is to remove the waste and to improve the added value steps (kaizen).
The kanban system is based on the Lean software development, highlighting an approach to maximize flow and avoiding bottlenecks and other kind of issues [77]. Kanban software process can be defined as WIP (Work In Progress) limited pull system visualized by the Kanban board. Kanban is based on five principles that largely overlap with the principles of Lean:

1. Show the workflow
2. Limit of the work in progress
3. Management of the flow
4. Collaboration improvement
5. Process policies explicit

5.1.2 Agile Tools

An agile tool can be defined as a project and general work management tool, able to help a team to support their use of agile practices. As pre announced before, there is a wide number of tools, from different categories and are classified in different ways. One of these categories is the agile project management. The majority of the teams make use of several tools to manage all the activities during the development phase. Modern agile project management software combines common activities providing a lot of services:

- Users stories and epics management.
- Priorities of backlog.
- High level release planning and low level iteration planning.
- Progress tracking
- Management of bugs, tests, customer's requests and so on.

There is the possibility to get continually support for one or more integrated processes, starting from the planning through deployment and several kind of operations.

5.2 Understanding Approval Rating of Agile Project Management Tools Using Twitter

5.2.1 Introduction

The role of managing a software project can be extremely complicated, requiring many teams and organizational resources. Software projects tend to raise many issues and problems throughout their life cycles. The quality of the final software product is related to how the project has been managed [57]. Project Management, then, is the application of knowledge, skills, and techniques to execute project effectively and efficiently. This action typically includes facing the needs and expectations of the project stakeholders as the project is scheduled and built up, as well as identifying the requirements of the project and balancing the project constraints [41]. Recently, the evolution of Project Management tools for both software and non-software applications is speeding up at a rapid pace and, then, the number of available products is growing considerably [32].

Many people are acknowledging that Agile development is helpful to business, with an high increase over the last years in the number of people who believe that Agile helps companies to complete projects faster.

Teams and organizations often support their Agile practices using an Agile Project Management tool, defined as a project and work management tool that helps a team or an organization to improve their quality and enhance project agility [9]. More and more software companies shift towards agile methodologies to achieve speed, efficiency and quality of the software. They take advantage of the various agile methods such as Extreme Programming (XP), Scrum[74] and Lean [55]. Since there is a multitude of available tools that supports Agile methods, we decided to evaluate directly which tools are most looked for using Web Search media, most mentioned and most appreciated through Social media.

Nowadays, Web 2.0 services such as blogs, tweets, forums, chats, email etc. are widely used as media for communication, with great results. Research has established that software engineers use Twitter, one of the most popular social network, in their work to communicate about software engineering topics. Through use of social media services, team members have opportunities to acquire more detailed information about their peers' expertise [28].

The popularity of Twitter is attracting the attention of researchers. Several recent studies examined Twitter from different point of views, including the sentiment prediction power [81], the topological characteristics of Twitter [50] or tweets as social sensors of real-time events [73]. Another stream of research focuses on corporate applications of microblogging such as the company internal use for project management [11] or the analysis of Twitter as electronic word of mouth in the area of product marketing [**jansen**]. Our aim is to evaluate the power of Twitter in order to understand the level of popularity and appreciation of Agile Project Management tools. We decided to apply automated Sentiment Analysis on shared short messages of users on Twitter in order to analyze automatically people's opinions, sentiments, evaluations and attitudes. We compared our results with one of the most used web search media, Google Trends, and with the related results of published surveys [83] [5] [57].

The body of this paper is organized in five major sections. Section 2, describes the background, section 3 presents the research steps of our study and section 4 summarizes and discusses our results. Finally, Sections 5 presents the conclusions and suggestions for future work.

5.2.2 Background

We analyzed the existing literature and surveys on tool usage in the context of Agile development, looking also for "common" sources such as websites, white-papers and published surveys. In the last years, two relevant surveys concerning Agile Project Management tools have been published.

The first is a study conducted by the tool vendor VersionOne in 2013. This is the latest of a series of similar surveys made every year by VersionOne, that is also a tool producer. Most of the survey is focused on the state of Agile development itself [83]. A total of 3500 responses were collected, analyzed and described in a summary report. The survey is very detailed, including information such as reasons for adopting Agile methods, resulting benefits, roles and so on. They declare that the most commonly used tools still are standard office productivity tools ,such as Excel (66%)

followed by tools like Microsoft Project (48%), VersionOne (41%), Atlassian/Jira (36%), Microsoft TFS (26%), IBM ClearCase (10%), LeanKit (5%), Xplanner (4%) and Trello (outside the survey choices). In addition to tool usage, the respondents were asked whether they would recommend the tools they are using based on their past or present use. VersioneOne had the highest satisfaction of any other tool evaluated (93%), followed by Atlassian/Jira (87%), LeanKit (84%), TargetProcess (83%), Microsoft TFS (79%) and ThoughtWorks Mingle (69%).

The other relevant survey is a study conducted in 2011[5]. It is an Agile Project Management Tool survey with 121 answers from 35 countries. The survey reports also features, a list of favored tool types and a list of tools that the respondents felt most satisfied with. This survey reports that the most commonly used tool within the companies studied is the physical wall (26%) followed by Microsoft Project (8%), Rally (5%), Mingle (3%), VersionOne (2%), JIRA (2%) and Team Foundation Server (2%).

In order to verify the accuracy of the information reported on the cited surveys, we decided to check these data through a public web service, called "Google Trends" [1]. It is a feature of the search engine that illustrates how frequently a fixed search term was looked for using Google. Using Google Trends you can compare up to five topics at one time to view relative popularity, allowing you to gain an understanding of the hottest search trends of the moment, along with those developing in popularity over time. Following this kind of approach and taking into consideration the list of previously mentioned tools, we are able to evaluate which ones are the most looked for using Google's search engine.

Figure 5.1 shows the top five of the most mentioned Agile Project Management tools in Google Trends. We found that, nowadays, the most searched tools are, in descending order of search, Jira, Trello, Mingle, Microsoft Project and Team Foundation Server. This graph shows also that, in the time period from 2004 to today, Microsoft Project has been losing popularity in favour of a more and more increasing fame of Jira and Trello.

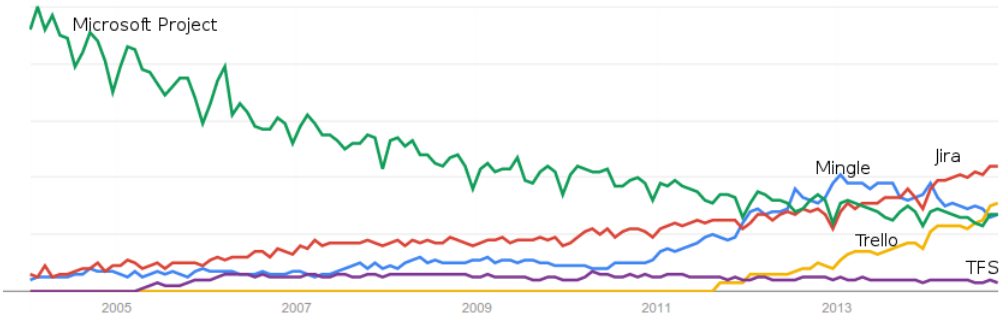

FIGURE 5.1: Most mentioned Agile Project Management tools in Google Trends

[1] http://www.google.it/trends/

5.2.3 Methodology

Sentiment Analysis and Twitter

Besides analyzing which are the most searched tools, we wanted to examine what customers really think about them. In the last years, the social web has been commercially exploited for goals such as automatically extracting customer opinions about products or brands, to find which aspects are liked and which are disliked [79].

Twitter is an online social networking website and microblogging service that allows users to post and read text-based messages of up to 140 characters, known as "tweets". It seems to be used to share information and to describe minor daily activities [43]. The short format of tweet is a defined characteristic of the service, allowing informal collaboration and quick information sharing. Black et al. presented a survey conducted to collect information on social media use in global software systems development. Twitter was found to be the most popular media and respondents affirmed that specification, source codes and design information were shared over social media [10]. Singer et al. showed that Twitter helps developers keep up with the fast-paced development landscape. They use it to stay aware of industry changes, for learning and for building work relationships [76]. Romero et al. showed that the aspect responsible for the popularity of certain topics is the influence of users of the network on the spread of content. Some members produce content that resonates very strongly with their followers thus causing the content to propagate and gain popularity [70].

So, the tweets sometimes express opinions about different topics, and for this reason we decided to evaluate how much users speak about Agile Project Management tools. In order to evaluate if a user really appreciates the tool, we tried to predict the sentiment analyzing the collection of tweets. By recent years, there is a wide collection of research surrounding machine learning techniques, in order to extract and identify subjective information in texts. This area is known as sentiment analysis or opinion mining. The research field of sentiment analysis has developed many algorithms to identify if the opinion expressed is positive or negative [64]. The strength of the sentiment analysis applied to the Twitter domain by applying similar machine learning techniques to classifying the sentiment of tweets [36].

For these reasons, we chose to use automated sentiment analysis techniques to identify the sentiments of tweets regarding Agile Project Management tools. Since the goal of this research is neither to develop a new sentiment analysis nor to improve an existing one, we to used "SentiStrenght", a tool developed by a team of researchers in the UK [78]. This tool was implemented to analyze informal short messages. Based on the formal evaluation of this system on a large sample of comments from MySpace.com, the accuracy of predicting positive and negative emotions was something similar to that of other systems (72.8% for negative emotions and 60.6% for positive emotions, based on a scale of 1-5), and compared to other methods, SentiStrenght showed the highest correlation with human coders [80]. The tool SentiStrenght is able to assess each message separately and, at the end, it returns one singular value: a positive (1), a negative (-1) or a neutral sentiment (0).

Data Collection

Starting from the list of the most used Agile Project Management tools found in the existing literature, we evaluated how many people speak on Twitter about these tools, and what they think about them. The tweets are available and are easily retrieved making use of Twitter Application Programming Interface (API). Composing the hashtags # or @ with the name of the tool, we can see all the tweets that mentioned the analyzed system. The system architecture consists of four components:

- *Twitter Streaming API*: it provides access to Twitter data, both public and protected, on a nearly real-time basis. A persistent connection is created between our system and Twitter. As soon as tweets come in, Twitter notifies our system in real time, allowing us to store them into our database.
- *DataStore*: our datastore consists of a back-end database engine, using MySQL as RDBMS, that repeatedly saves the incoming tweets from the Twitter Streaming API.
- *SentiStrenght tool*
- *Java Module*: this component allows us to send automated requests to Twitter Streaming API, to recover new tweets about analyzed tools, to parse data gained and to store them into our datastore. In a later stage, these data are sent to SentiStrenght tool in order to evaluate automatically the user's opinion.

We analyzed a collection of tweets posted on Twitter between September 2014 and March 2015 regarding the Agile project management tools most mentioned. We found a total of 84837 tweets. We then used the SentiStrenght tool to evaluate the comments extracted from Twitter. Given as input all tweets of every tool previously mentioned, a score for each comment was assigned.

- 1 if the comment is positive
- -1 if the comment is negative
- 0 if the comment is neutral

5.2.4 Results and discussion

Number of Tweets Analysis

Of the total 84837 tweets, 39756 were neutral (46%), 35409 were positive (42%) and 9672 were negative (11%). In Table 5.1, the most quoted Agile Project Management tools, ordered by number of tweets, are shown. XPlanner and IBM ClearCase were excluded since few comments were found about them.

Going back to the results of Google Trends about the Agile Project Management tools reported in Figure 1, we observed a striking similarity between it and the number of tweets we found. In fact, Google Trends reported, in order of search, Jira, Trello, Mingle, Microsoft Project and Team Foundation Server. Taking a look to Table 5.1, we can see that the rank is

TABLE 5.1: *Most quoted Agile Project Management tools ordered by number of tweets*

Number of Tweets	
Trello	32613
Jira	21903
VersionOne	7887
Microsoft TFS	6730
Rally	5684
Mingle	4629
Microsoft Project	2249
LeanKit	1595
TargetProcess	1547

similar, with the exception of VersionOne. So, we found that there is a relation between the number of tweets posted for each tool and the Google searches about it. TargetProcess turns out to be the least tweeted, maybe due to the fact that it is a recent tool and it is still little known worldwide. Consequently, it is possible to stress that the most used and quoted Agile Project Management tools on Twitter and Google are Jira, Trello and VersionOne.

Comparison Between Negative and Positive Tweets

From the list of tweets, we evaluated the sentiment using SentiStrenght tool, in order to understand if the users really appreciated these Agile Project Management tools. Among all tweets, we observed that, on average, 40% of the tweets don't represent a sentiment, or Sentistrenght is not able to identify it. After a careful analysis, it was observed that a lot of tweets are neutral because often people wrote texts asking help, non-expressive comments, tweets in a different language than English, or simple links that lead to other web pages.

An Agile Project Management tool that has more positive than negative tweets is likely to be successful. After this evaluation, we determined that, for all tools, there were more positive messages than negative ones, and that positive messages were almost 3 times more likely to be forwarded than negative messages. In Table 5.2 we show the data found about positive and negative tweets using SentiStrenght. We observed that, in general, the percentage of negative tweets found in each tool is lower than 14%; Jira, however, achieved 16% of negative comments. In fact, users of Jira sometimes posted tweets in which they explained their problems using the tool.

Based on this result, we found that Trello is clearly the most appreciated tool, getting a high percentage of positive comments (51%) and only 8% of negative comments. Beyond this tool, the other tools with a high percentage of positive comments are LeanKit (52%) and Microsoft Project (47%).

To better quantify the sentiments, we defined the score *PNRatio* as the ratio of positive versus negative tweets on each tool.

$$PNRatio = \frac{|Tweets\ with\ Positive\ Sentiment|}{|Tweets\ with\ Negative\ Sentiment|} \tag{5.1}$$

TABLE 5.2: *Comparison between negative and positive tweets*

Tool	Negative		Positive	
	Total	%	Total	%
Trello	2760	8 %	16569	51 %
Jira	3445	16 %	7043	32 %
VersionOne	808	10 %	2612	33 %
MicrosoftTFS	859	13 %	2290	34 %
Rally	673	12 %	2612	46 %
Mingle	554	12 %	1971	43 %
MicrosoftProj	247	11 %	1055	47 %
LeanKit	89	6 %	829	52 %
TargetProcess	237	15 %	428	28 %

TABLE 5.3: *Comparison between PN ratio of eq.5.1 and VersionOne satisfaction survey results by tools*

Tool Name	Ratio P/N	[83]
LeanKit	9.31	84%
Trello	6.00	-
Microsoft Project	4.27	53%
Rally	3.88	-
VersionOne	3.23	93%
Mingle	3.55	69%
Microsoft TFS	2.66	79%
Jira	2.04	87%
TargetProcess	1.80	83%

The indicator of eq.5.1 was applied to all tools and Table 5.3 shows the results. We compared our results to the satisfaction achievement obtained by VersionOne survey, where it is possible. In our study the most quoted (excluding Trello and Rally, since they were not included in VersionOne survey) are LeanKit, Microsoft Project and VersionOne. We noticed that LeanKit and VersionOne also exhibit a high percentage of satisfaction from the survey, greater than 84%. Nevertheless, Trello turns out to be one of the most popular tool, showing a ratio P/N of 6.00, with a total of 16569 positive comments.

In Table 5.3 we observed a relation between PN ratio and its satisfaction, for most tools. As a matter of fact, the majority of the tweets represents a positive sentiment while the negative comments are less than 16% and PN Ratio is always greater than 2. Also in this case, we can confirm that the favorite tools are Trello and VersionOne.

Relationship with Agile Methodologies and Approaches

We decided to analyze all tweets in order to assess whether some Agile methodology was mentioned by someone and, in the case, which one of these and in which tool. So, for each tweet, we evaluated if the user cited one of the major agile methodologies and approaches. We chose to test Scrum, Kanban approach, Lean and eXtreme Programming (XP) [74] [55].

TABLE 5.4: *Most mentioned Agile methodologies in the tweets*

Methodology	Tot tweets	Tool contribution
Scrum	2021	VersionOne 53% Rally 20% Jira 10% Trello 6%
Kanban	807	LeanKit 34% Trello 40% Jira 12%
Lean	439	VersionOne 30% LeanKit 21%
eXtreme Programming	111	VersionOne 42% Jira 27% Trello 14%

Table 5.4 shows the most cited methodologies within the tweets; we can notice that the methodology most mentioned is Scrum, followed by Kanban, Lean and finally eXtreme Programming. About Scrum, we found that the main contribution is given by VersionOne, Rally, Jira and Trello. On the other hand Trello, Jira and LeanKit quoted Kanban approach.

We checked these results using Google Trends, to confirm the ranking found using Twitter. Figure 5.2 shows the results. Scrum is clearly the most searched word, so we can say that it's the most popular methodology, and its fame is still growing. The popularity of Lean and Kanban are quite similar while eXtreme Programming, over the years, lost most of its fame. In the end, Google Trends confirmed the rankings related to Agile methodologies and approaches found using Twitter. It affirms the growing popularity of Scrum in the last years, and the effectiveness of using tweets to assess the popularity of something.

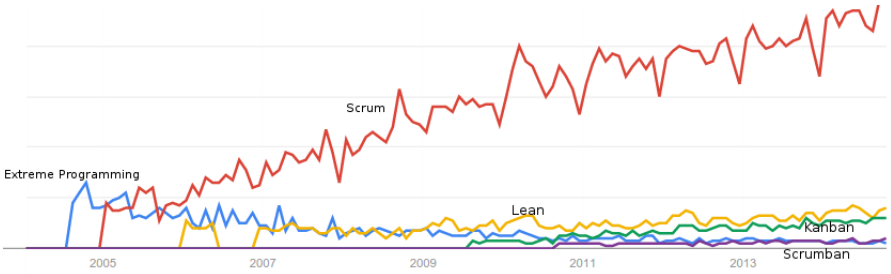

FIGURE 5.2: Most mentioned Agile metodologies in Google Trends

The aim of this work was to understand which tools are most mentioned and looked for, and, if those who use it really appreciate it. Using Twitter, it was possible to recognize the most mentioned tools, and to evaluate their level of appreciation. This approach is able to give immediate results, reducing the need to submit surveys to users. We can conclude that Jira, Trello and VersionOne are the most twitted and, at the same time, the most appreciated Agile Project Management tools. Google Trends confirms that the fame of Jira and Trello is ever increasing in these years.

In future we plan to extend the increase the number of analyzed tweets, and to perform a complete correlation analysis of the results. We also plan to gather tweets along a much longer time interval, performing trend analysis on the tweets. In this way, we could assign a specific weight to each tweet and check whether the results remain unchanged with the addition of this variable.

Chapter 6

Understanding Economic Uncertain Policy in Belgium through Social Media

6.1 Introduction

Twitter is an online social networking website and microblogging service that allows users to post and read text-based messages of up to 140 characters, known as "tweets". Launched in July of 2006 by Jack Dorsey, Twitter is now in the top 10 most visited internet sites with a total amount of 645,750,000 registered users.
The short format of a tweet is a defined characteristic of the service, allowing informal collaboration and quick information sharing. For business, Twitter can be used to broadcast company's latest news, posts, read comments of the customers or interact with them. A communicative feature of Twitter is the hashtag: a metatag beginning with the character #, designed to help others find a post.
Due to its growing popularity, tweets are about different topics, ranging from economic news to brand information. Twitter's fame is attracting a lot of attention from researchers of several disciplines. There are several fields of research, one of them focuses on understanding its usage and community structure. Another stream of research is about the influence of users and the propagation of information. Finally, there is a field of research focused on the predictive power and the potential application to different areas [73]. There are several studies that investigate the Twitter sentiment during elections or economical/political events [81], analyze discussions performances [18], measure user influence or study the predictive power of Twitter in several fields, like financial markets.
Huberman et al. analyzed the social interaction on Twitter, demonstrating that the driver of usage is a sparse hidden network among friends and followers, while most of the interaction links are meaningless [4]. Romero et al. showed that the correlation between popularity and influence is weaker than it might be expected, because most users are passive information consumers and do not forward the content to the network [70].
By building a model capturing the speed, scale and range of information diffusion, Yang et al. [84] claimed that some properties of the tweets themselves predict greater information propagation. Beyond the overall comprehension of Twitter, other researchers are interested in its prediction power and potential application to other areas. Asur and Huberman used Twitter to forecast box-office revenues of movies [4]. They demonstrated that a

simple model constructed from the rate at which tweets are created about particular topics could achieve better results than market-based predictors. In their study, Tumasjan et al. studied Twitter messages that cite parties and politicians prior to the German federal election 2009 and they found that the number of tweets reflects voter preferences and comes close to traditional election polls [81]. Other researchers found that Twitter also could be used in areas such as tracking the pace of epidemic disease [51]. There is also prior work on analyzing correlation among web buzz and stock market. Antweiler and Frank determine correlation between activity in Internet message boards and stock volatility and trading volume [2].
Other researches worked on blog posts to predict stock market behaviour. Gilbert and Karahalios used over 20 million posts from the LiveJournal website to develop an index of the US national mood, which they nominate the Anxiety Index [34]. They found that when this index rose sharply, the SeP 500 ended the day marginally lower than is expected. Besides the posts' contents itself, other properties of communication such as the number of comments, the length and response time of comments etc. are also helpful. Choudhury et al. elaborated such contextual properties as a regression problem in a Support Vector Machine framework and trained it with stock movement [26]. Their results are promising, yielding about 87% accuracy in predicting the direction of movement.
According to the international institutions, economic policy uncertainty increased to historically high levels after the 2007-2009 recession due to uncertainty about tax, spending, regulatory, and monetary policies. This kind of uncertainty has slowed the upswing from the recession by causing businesses and households to reduction or postpone investment, hiring and consumption.
In his work, In't Veld modelled the impact on GDP of fiscal consolidation under different uncertainty and learning scenarios [69]. In a scenario of uncertainty on the credibility of the fiscal consolidation, the short term negative impact on GDP is up to 3 times higher than in a scenario of immediate credibility. Balta et al. found that uncertainty has an important effect on both investment and consumption in the euro zone with the effect of uncertainty on activity increasing since the crisis and going beyond traditional cyclical effects [7].
Economic research has come up with several ways of constructing uncertainty measures based on stock market volatility [22], dispersion in forecasts by professional forecasters or in expectations of consumers or producers, or the prevalence of terms such as economic uncertainty in the media [6]. Tobback et al. focused on the third methodology and contribute to the economic literature by using the latest state-of-the-art text mining methods to construct uncertainty indicators [1]. This methodology allows to identify the main factors with which uncertainty is associated.
Recently, Baker, Bloom and Davis [6] have constructed an Economic Policy Uncertainty index (EPU) as a proxy for movements in policy related economic uncertainty over time. This index combines the frequency of newspaper references to EPU with the deviation of future inflation expectations. The authors found that their index peaks near important events such as

[1] https://www.ecb.europa.eu/events/pdf/conferences/140407/BelgianEPU-textminingversion2403.pdf

9/11 and the bankruptcy of Lehman Brothers. The index has given rise to numerous studies concerning the influence of economic uncertainty on macroeconomic indicators.
In this paper we tried to understand if social media volume can be able to predict the economic policy uncertainty with a particular focus on Belgium country. Tweets and Facebook posts, written by the most famous Belgian economists, have been analyzed in more detail applying an automatic sentiment analysis technique. We found a striking similarity between this data and the Belgium Government Bond 5Y demonstrating our initial ideas.

6.2 Methodology

6.2.1 Opinion Mining

The opinion mining is a particular technique that detects automatically the sentiment and subjectivity transmitted in written texts. The user's tweets could express the opinion regarding different topic, trends or brands [64]. For this reason, we decided to monitor the sentiment expressed, day after day, by the belgian economists on the matter of Economic Policy Uncertainty.
Since the goal of this research is neither to develop a new sentiment analysis nor to improve an existing one, we used "SentiStrenght", a tool developed by a team of researchers in the UK that demonstrated good outcomes [80]. SentiStrength estimates the strength of positive and negative sentiments in short texts. It is based on a dictionary of sentiment words, each one associated with a weight, which is its sentiment strength. In addition, this method uses some rules for non-standard grammar.
Based on the formal evaluation of this system on a large sample of comments from MySpace.com, the accuracy of predicting positive and negative emotions was something similar to that of other systems (72.8% for negative emotions and 60.6% for positive emotions, based on a scale of 1-5). Compared to other methods, SentiStrenght showed the highest correlation with human coders [79]. The tool is able to assess each message separately and, at the end, it returns one singular value.

- +1 if the system identifies a positive sentiment
- -1 if the system identifies a negative sentiment
- 0 if a neutral opinion is identified

6.2.2 Topic Modeling

Topic modeling is gaining increasingly attention in different text mining communities. Latent Dirichlet Allocation (LDA) [3] is becoming a standard tool in topic modeling. Latent Dirichlet Allocation is an unsupervised machine learning technique which identifies latent topic information in large document collections. It uses a "bag of words" approach, which treats each document as a vector of word counts. Each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a number of words. LDA defines the following generative process for each document in the collection:

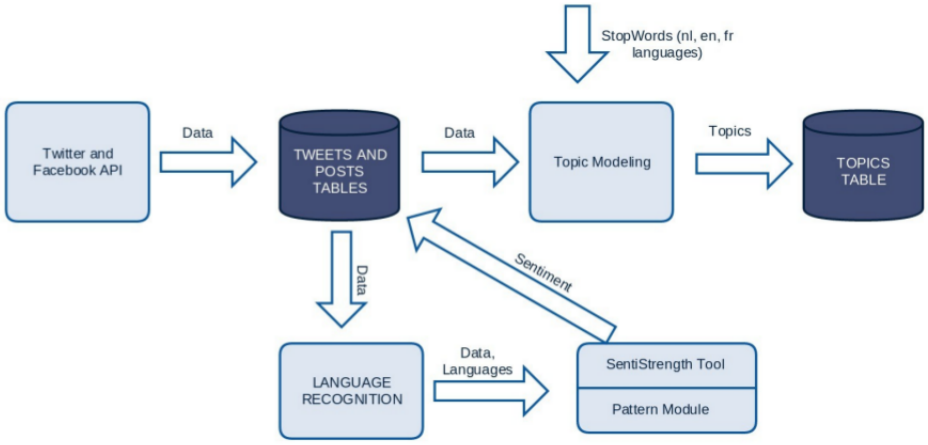

FIGURE 6.1: System architecture

1. For each document, pick a topic from its distribution over topics.
2. Sample a word from the distribution over the words associated with the chosen topic.
3. The process is repeated for all the words in the document.

We applied this technique in order to understand which topics are posted on social media during the time period considered in a rapid way.

6.2.3 Data Collection

Starting from the list of the main economists of Belgium, we checked who of them are the most active on Twitter and Facebook, and we found 26 economists. For each of them we downloaded all the posts with Facebook API and the tweets, using Twitter API. We implemented a system able to connect itself to Twitter and Facebook API and to download automatically the posts for each user. Then, the data are saved in a database for later queries. Figure 6.1 shows the system architecture explained above. The data are preprocessed in order to identify automatically the language and then the sentiment is calculated using the appropriate dictionary with SentiStrength tool.
All the available data regarding the most important belgian economists have been downloaded and used to build the topic modelling. For each topic, the list of authors who have contributed, with the number of words assigned to the topic, has been identified and saved in the database.

6.3 Results

We collected almost 60000 posts covering the period between 2010 and 2015. Statistics about the number of posts per year are shown below.

- 2010 $\rightarrow$ 404 posts
- 2011 $\rightarrow$ 890 posts

FIGURE 6.2: Words cloud that illustrates the top 50 words in Social media

- 2012 → 4339 posts
- 2013 → 15847 posts
- 2014 → 13864 posts
- 2015 → 12092 posts

It's well visible that the number of comments increases year after year due to the increasing popularity of social media like Facebook and Twitter. The words cloud in the Figure 6.2 illustrates the top 50 words. In that way, there is the guarantee that these posts contains topics related to the economy of the analyzed country.

We used the standard topic modelling technique, in order to extract the main topics after an automatic analysis of the available data. The data are preprocessed with the methods illustrated below.

- *Tokenizing*: a document is converted to its atomic elements.
- *Stopping*: removal of meaningless words.
- *Stemming*: merge of words that are equivalent in meaning.
- *Removing* links and words starting with '@'

We found several topics related to the economy and some examples are shown below.

- ECB and FED monetary policy
- European Council
- Analysis from KBC: KBC ECONOMICS and BOLERO
- EMU, Spain and Greece
- Greek crisis-ECB reaction
- Debt crisis Ukrain Greece
- Qoin agency for community currencies

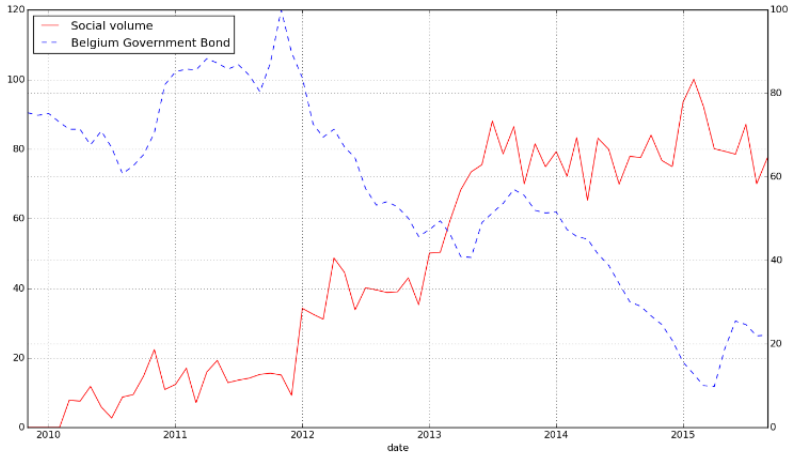

FIGURE 6.3: Comparison between social data index and Belgium Government Bond 5Y

Different lexical items and constructions could be used to express uncertainty, including auxiliary verbs, main verbs, adjectives, adverbs and others. Given a list of modal items expressing uncertainty in Dutch and English language, we selected posts that contain these elements.
From these preselected posts, we analyzed the sentiment expressed using SentiStrength tool. I found that 46% of posts represents a negative sentiment and only 12% shows a positive mood. We filtered an other time the data keeping only those that express a negative sentiment.
It's important to take into account even the number of economists that gave the contribution in a fixed period. Then, I tried to consider a new index where, for each month, we consider the following formula:

$$Index = \frac{numberofuncertainandnegativeposts * numberofEconomists}{numberofuncertainposts} \tag{6.1}$$

All the indexes are normalized in a range between 0 and 100. The data refers to the period between 2010 and 2015 and I chose to compare them with the normalized Belgium Government Bond 5Y. We applied a Pearson correlation analysis in order to verify if a relationship exists between them. The Figure 6.3 shows the two time series.

Pearson's correlation *r* is a statistical measure that evaluate the strength of a linear association between two time series *G* and *T*. Initially, we assumed G as social data index and T as Belgium Government Bond 5Y.

$$r = \frac{\sum_i (G_i - \overline{G})(T_i - \overline{T})}{\sqrt{\sum_i (G_i - \overline{G})^2}\sqrt{\sum_i (T_i - \overline{T})^2}} \tag{6.2}$$

The Pearson correlation coefficient has values between -1 and +1, the bounds denoting maximum anti-correlation or correlation, respectively, whereas 0 indicates no correlation. We calculated the Pearson correlation and we found a result equal to -0.85, which is quite high. Indeed, the data are inversely related: when authors reveal uncertainty the stock market goes down and viceversa. The Figure 6.3 demonstrates the existing correlation with a significant result, due to the P-value<0.00001.

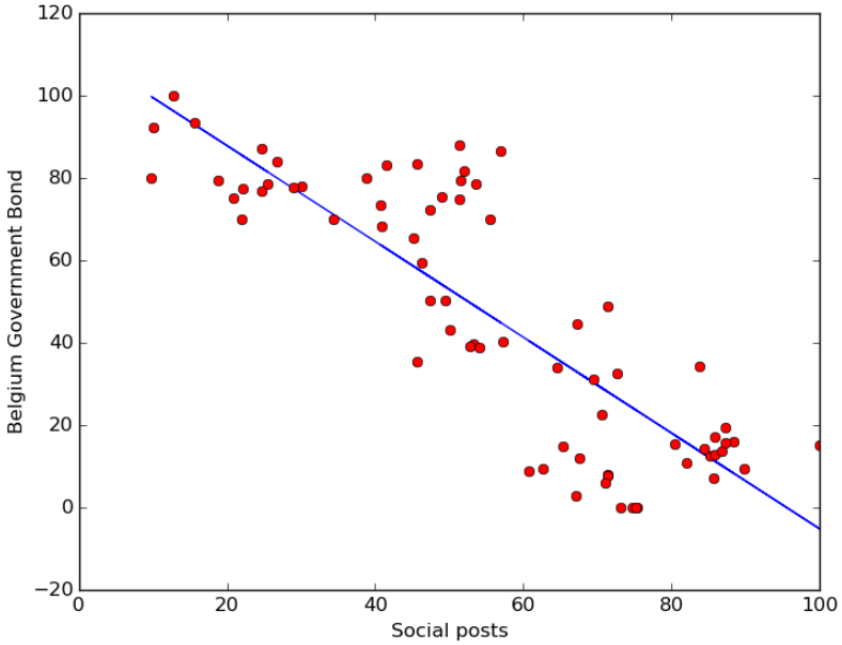

FIGURE 6.4: Correlation between social data index and Belgium Government Bond 5Y

From results of this analysis, we can affirm that the new social media index is a good predictor, because of its high Pearson correlation value. Indeed, the social index reproduces the Belgium Government Bond 5Y in a optimal way. Thus, applying the same approach, we expect the same optimal result with different phenomena.

Chapter 7

Conclusions

The objective, described in this thesis and conducted during these three years of PhD, was to evaluate the predictive power of web search and social media analyzing their data volumes during the time.
I investigated whether the chatter of the social community can be used to make qualitative predictions about a particular phenomena, attempting to establish whether there is any correlation between them.
Simultaneously, I applied an automated Sentiment Analysis on shared short messages of users on Twitter in order to automatically analyze peoples opinions, sentiments, evaluations and attitudes. We wondered whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can be used to predict particular trends.
In addition, the frequency of searches of terms on search engines could have a good explanatory power. Web search media activity could be helpful and used by investment professionals. So, I decided to examine Google, one of the most important search engine, analyzing the popularity of particular topics under its perspective.
In the Chapter 4, I presented two case studies about Bitcoin prediction. In the first work, I studied whether Bitcoin's trading volume is related to the web search and social volumes about Bitcoin. I investigated whether public sentiment, expressed in large-scale collections of daily Twitter posts, can be used to predict the Bitcoin market too.
I achieved significant cross correlation outcomes, demonstrating the search and social volumes power to anticipate trading volumes of Bitcoin currency. From results of cross correlation analysis between the time series, I found that the ratio between positive and negative tweets may contribute to predict the movement of Bitcoin's trading volume in a few days. Anyway, I found that Google Trends can be seen as the best predictor, because of its high cross correlation value with three days of lag.
In the second work, I investigated if the spread of the Bitcoin's price is related to the volumes of tweets or Web Search media results. I compared trends of price with Google Trends data, volume of tweets and particularly with those that express a positive sentiment. From results of analysis between these time series, I can affirm that positive tweets may contribute to predict the movement of Bitcoin's price in a few days. Google Trends could be seen as a kind of predictor, because of its high cross correlation value.
In the Chapter 5, I presented an analysis of a dataset of tweets about Agile project management in order to identify automatically the tools most used by companies.
Using Twitter, it has been possible to recognize the most mentioned tools, and to evaluate their level of user's appreciation. This approach is able

to give immediate results, reducing the need to submit surveys to users. I found that social media is a good predictor of popularity regarding the Agile project management tools. Consequently, I can affirm that the same approach can be applied to several phenomena and other contexts.
In the Chapter 6, I investigated whether social media volume can be able to predict the economic policy uncertainty with a particular focus on Belgium country. Tweets and Facebook posts, written by the most famous Belgian economists, have been analyzed in more detail applying an automatic sentiment analysis technique. I found a striking similarity between this data and the Belgium Government Bond 5Y, demonstrating our initial ideas.
Finally, I can conclude this thesis confirming the first initial ideas. By the time, social and web search media are always more used by worldwide users, and for this reason it is possible to capture a lot of personal information and identify the predictive power of the media.
As future advancement, I'm thinking about the possibility to apply this kind of approach to different contexts in order to better understand the predictive power of web search and social media. An other likelihood could be to consider not only search media but also social media like Twitter, Facebook and Google+.

Bibliography

[1] A. Alliance. "Agile manifesto". In: *Online at http://www. agilemanifesto. org* 6.6.1 (2001).

[2] W. Antweiler and M. Z. Frank. "Is all that talk just noise? The information content of internet stock message boards". In: *The Journal of Finance* 59.3 (2004), pp. 1259–1294.

[3] M. Aoyama. "Web-based agile software development". In: *Software, IEEE* 15.6 (1998), pp. 56–65.

[4] S. Asur, B. Huberman, et al. "Predicting the future with social media". In: *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*. Vol. 1. IEEE. 2010, pp. 492–499.

[5] G. Azizyan, M. K. Magarian, and M. Kajko-Matsson. "Survey of agile tool usage and needs". In: *Agile Conference (AGILE), 2011*. IEEE. 2011, pp. 29–38.

[6] S. R. Baker, N. Bloom, and S. J. Davis. "Measuring economic policy uncertainty". In: *Chicago Booth research paper* 13-02 (2013).

[7] N Balta, I. V. Fernández, and E Ruscher. "Assessing the impact of uncertainty on consumption and investment". In: *Quarterly Report of the Euro Area* (2013), pp. 7–16.

[8] K. Beck. *Extreme programming explained: embrace change*. Addison-Wesley Professional, 2000.

[9] K. Beck et al. "Manifesto for agile software development". In: (2001).

[10] S. Black, R. Harrison, and M. Baldwin. "A survey of social media use in software systems development". In: *Proceedings of the 1st Workshop on Web 2.0 for Software Engineering*. ACM. 2010, pp. 1–5.

[11] M. Böhringer and A. Richter. "Adopting social software to the intranet: a case study on enterprise microblogging". In: *Proceedings of the 9th Mensch & Computer Conference*. 2009, pp. 293–302.

[12] J. Bollen, H. Mao, and X. Zeng. "Twitter mood predicts the stock market". In: *Journal of Computational Science* 2.1 (2011), pp. 1–8.

[13] I. Bordino et al. "Web search queries can predict stock market volumes". In: *PloS one* 7.7 (2012), e40014.

[14] M. M. Bradley and P. J. Lang. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Tech. rep. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, 1999.

[15] L. Bulut et al. *Google Trends and Forecasting Performance of Exchange Rate Models*. Tech. rep. 2015.

[16] E. Cambria et al. "SenticNet: A Publicly Available Semantic Resource for Opinion Mining." In: *AAAI fall symposium: commonsense knowledge*. Vol. 10. 2010, p. 02.

[17] E. Cambria et al. “Towards crowd validation of the UK national health service”. In: *WebSci10* (2010), pp. 1–5.

[18] M. Cha et al. “Measuring User Influence in Twitter: The Million Follower Fallacy.” In: *ICWSM* 10.10-17 (2010), p. 30.

[19] H. J. Cheong and M. A. Morrison. “Consumers’ reliance on product information and recommendations found in UGC”. In: *Journal of Interactive Advertising* 8.2 (2008), pp. 38–49.

[20] H. Choi and H. Varian. “Predicting the present with google trends”. In: *Economic Record* 88.s1 (2012), pp. 2–9.

[21] F. Ciulla et al. “Beating the news using social media: the case study of American Idol”. In: *EPJ Data Science* 1.1 (2012), pp. 1–11.

[22] S. Claessens, M. A. Kose, and M. E. Terrones. “How do business and financial cycles interact?” In: *Journal of International economics* 87.1 (2012), pp. 178–190.

[23] E. Constantinides, C. L. Romero, and M. A. G. Boria. “Social media: a new frontier for retailers?” In: (2009), pp. 1–28.

[24] C. Curme et al. “Quantifying the semantics of search behavior before stock market moves”. In: *Proceedings of the National Academy of Sciences* 111.32 (2014), pp. 11600–11605.

[25] M. Cusumano, D. B. Yoffie, et al. “Software development on Internet time”. In: *Computer* 32.10 (1999), pp. 60–69.

[26] M. De Choudhury et al. “Can blog communication dynamics be correlated with stock market activity?” In: *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*. ACM. 2008, pp. 55–60.

[27] T. De Smedt and W. Daelemans. “Pattern for python”. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 2063–2067.

[28] K. G. Dessai, M. S. Kamat, and R. Wagh. “Application of social media for tracking knowledge in agile software projects”. In: *Available at SSRN 2018845* (2012).

[29] W. Diffie and M. E. Hellman. “New directions in cryptography”. In: *Information Theory, IEEE Transactions on* 22.6 (1976), pp. 644–654.

[30] T. Dimpfl and S. Jank. “Can internet search queries help to predict stock market volatility?” In: *European Financial Management* (2015).

[31] A. Esuli and F. Sebastiani. “Sentiwordnet: A publicly available lexical resource for opinion mining”. In: *Proceedings of LREC*. Vol. 6. Citeseer. 2006, pp. 417–422.

[32] J. Fortune et al. “Looking again at current practice in project management”. In: *International Journal of Managing Projects in Business* 4.4 (2011), pp. 553–572.

[33] D. Garcia et al. “The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy”. In: *Journal of the Royal Society Interface* 11.99 (2014), p. 20140623.

[34] E. Gilbert and K. Karahalios. “Predicting tie strength with social media”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2009, pp. 211–220.

[35] J. Ginsberg et al. "Detecting influenza epidemics using search engine query data". In: *Nature* 457.7232 (2009), pp. 1012–1014.

[36] A. Go, R. Bhayani, and L. Huang. "Twitter sentiment classification using distant supervision". In: *CS224N Project Report, Stanford* 1 (2009), p. 12.

[37] P. Gonçalves, F. Benevenuto, and M. Cha. "Panas-t: A pychometric scale for measuring sentiments on twitter". In: *arXiv preprint arXiv:1308.1857* (2013).

[38] C. W. Granger. "Investigating causal relations by econometric models and cross-spectral methods". In: *Econometrica: Journal of the Econometric Society* (1969), pp. 424–438.

[39] R. Grinberg. "Bitcoin: an innovative alternative digital currency". In: *Hastings Sci. & Tech. LJ* 4 (2012), p. 159.

[40] D. Hansen, B. Shneiderman, and M. A. Smith. *Analyzing social media networks with NodeXL: Insights from a connected world*. Morgan Kaufmann, 2010.

[41] J. Heagney. *Fundamentals of project management*. AMACOM Div American Mgmt Assn, 2011.

[42] K. Hicks et al. "Exploring Twitter as a game platform; strategies and opportunities for microblogging-based games". In: *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. ACM. 2015, pp. 151–161.

[43] A. Java et al. "Why we twitter: understanding microblogging usage and communities". In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM. 2007, pp. 56–65.

[44] J. Kaminski and P. Gloor. "Nowcasting the Bitcoin Market with Twitter Signals". In: *arXiv preprint arXiv:1406.7577* (2014).

[45] A. M. Kaplan and M. Haenlein. "Users of the world, unite! The challenges and opportunities of Social Media". In: *Business horizons* 53.1 (2010), pp. 59–68.

[46] Y. B. Kim et al. "Virtual world currency value fluctuation prediction system based on user sentiment analysis". In: *PloS one* 10.8 (2015), e0132944.

[47] L. Kristoufek. "BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era". In: *Scientific reports* 3 (2013).

[48] L. Kristoufek. "Can Google Trends search queries contribute to risk diversification?" In: *Scientific reports* 3 (2013).

[49] L. Kristoufek. "Power-law correlations in finance-related Google searches, and their cross-correlations with volatility and traded volume: Evidence from the Dow Jones Industrial components". In: *Physica A: Statistical Mechanics and its Applications* 428 (2015), pp. 194–205.

[50] H. Kwak et al. "What is Twitter, a Social Network or a News Media?" In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: ACM, 2010, pp. 591–600. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772751. URL: http://doi.acm.org/10.1145/1772690.1772751.

[51] V. Lampos and N. Cristianini. "Tracking the flu pandemic by monitoring the social web". In: *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*. IEEE. 2010, pp. 411–416.

[52] C. Larman. *Agile and iterative development: a manager's guide*. Addison-Wesley Professional, 2004.

[53] S. L. Lo, D. Cornforth, and R. Chiong. "Identifying the high-value social audience from Twitter through text-mining methods". In: *Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems, Volume 1*. Springer. 2015, pp. 325–339.

[54] F. Mai et al. "From Bitcoin to Big Coin: The Impacts of Social Media on Bitcoin Performance". In: (2015).

[55] P. Mary and P. Tom. "Lean Software Development: An Agile Toolkit for Software Development Managers". In: *Lean Software Development: An Agile Toolkit for Software Development Managers* (2003).

[56] M. Matta, I. Lunesu, and M. Marchesi. "Bitcoin Spread Prediction Using Social And Web Search Media". In: *Proceedings of DeCAT* (2015).

[57] A. Mishra and D. Mishra. "Software project management tools: a brief comparative view". In: *ACM SIGSOFT Software Engineering Notes* 38.3 (2013), pp. 1–4.

[58] A. Mittal and A. Goel. "Stock prediction using twitter sentiment analysis". In: *Standford University, CS229* (2012).

[59] D. Mocanu et al. "The twitter of babel: Mapping world languages through microblogging platforms". In: *PloS one* 8.4 (2013), e61981.

[60] J. Mondria, T. Wu, and Y. Zhang. "The determinants of international investment and attention allocation: Using internet search query data". In: *Journal of International Economics* 82.1 (2010), pp. 85–95.

[61] S. Nakamoto. "Bitcoin: A peer-to-peer electronic cash system". In: *Consulted* 1.2012 (2008), p. 28.

[62] T. O'Reilly. "Web 2.0: compact definition". In: *Message posted to http://radar.oreilly. com/archives/2005/10/web_20_compact_definition. html* (2005).

[63] A. Pak and P. Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." In: *LREC*. Vol. 10. 2010, pp. 1320–1326.

[64] B. Pang and L. Lee. "Opinion mining and sentiment analysis". In: *Foundations and trends in information retrieval* 2.1-2 (2008), pp. 1–135.

[65] M. Poppendieck and T. Poppendieck. *Lean software development: an agile toolkit*. Addison-Wesley Professional, 2003.

[66] T. Preis, H. S. Moat, and H. E. Stanley. "Quantifying trading behavior in financial markets using Google Trends". In: *Scientific reports* 3 (2013).

[67] T. Preis, D. Reith, and H. E. Stanley. "Complex dynamics of our economic life on different scales: insights from search engine query data". In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 368.1933 (2010), pp. 5707–5719.

[68] T. Rao and S. Srivastava. "Analyzing stock market movements using twitter sentiment analysis". In: *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society. 2012, pp. 119–123.

[69] M. Ratto, W. Roeger, and J. in't Veld. "QUEST III: An estimated open-economy DSGE model of the euro area with fiscal and monetary policy". In: *economic Modelling* 26.1 (2009), pp. 222–233.

[70] D. M. Romero et al. "Influence and passivity in social media". In: *Machine learning and knowledge discovery in databases*. Springer, 2011, pp. 18–33.

[71] D. Ron and A. Shamir. "Quantitative analysis of the full bitcoin transaction graph". In: *Financial Cryptography and Data Security*. Springer, 2013, pp. 6–24.

[72] D. E. Rose and D. Levinson. "Understanding user goals in web search". In: *Proceedings of the 13th international conference on World Wide Web*. ACM. 2004, pp. 13–19.

[73] T. Sakaki, M. Okazaki, and Y. Matsuo. "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors". In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: ACM, 2010, pp. 851–860. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772777. URL: http://doi.acm.org/10.1145/1772690.1772777.

[74] K. Schwaber. *Agile project management with Scrum*. Microsoft Press, 2004.

[75] K. Schwaber and J. Sutherland. "The scrum guide". In: *Scrum Alliance* (2011).

[76] L. Singer, F. Figueira Filho, and M.-A. Storey. "Software engineering at the speed of light: how developers stay current using twitter". In: *Proceedings of the 36th International Conference on Software Engineering*. ACM. 2014, pp. 211–221.

[77] Y Sugimori et al. "Toyota production system and kanban system materialization of just-in-time and respect-for-human system". In: *The International Journal of Production Research* 15.6 (1977), pp. 553–564.

[78] M Thelwall, K Buckley, and G Paltoglou. *SentiStrength, 2011. Available online*.

[79] M. Thelwall, K. Buckley, and G. Paltoglou. "Sentiment in Twitter events". In: *Journal of the American Society for Information Science and Technology* 62.2 (2011), pp. 406–418.

[80] M. Thelwall et al. "Sentiment strength detection in short informal text". In: *Journal of the American Society for Information Science and Technology* 61.12 (2010), pp. 2544–2558.

[81] A. Tumasjan et al. “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.” In: *ICWSM* 10 (2010), pp. 178–185.

[82] *Twitter API*. https://dev.twitter.com/rest/tools/console. [Online; accessed 30-November-2014].

[83] VersionOne. *Inc. 8th Annual State of Agile Development Survey*. http://www.versionone.com/pdf/2013-state-of-agile-survey.pdf. [Online; accessed 8-April-2015]. 2014.

[84] J. Yang and S. Counts. “Predicting the Speed, Scale, and Range of Information Diffusion in Twitter.” In: *ICWSM* 10 (2010), pp. 355–358.

[85] S. Ye and S. F. Wu. “Estimating the size of online social networks”. In: *International Journal of Social Computing and Cyber-Physical Systems* 1.2 (2011), pp. 160–179.