Explaining Stock Exchange Prices using supervised learning and sentiment analysis

Álvaro Antón Blanco
Master in BI & Big Data
Social Network Analysis

Director: Dra. Laia Subirats Maté
Teachers responsible of the subject: Dra. Maria Pujol Jover and Dra. Teresa Sancho Vinuesa

30.06.2017

© Álvaro Antón Blanco
All rights reserved. Reproduction in whole or in part of this work is prohibited by any means or procedure, including printing, reprography, microfilm, computer processing or any other system, as well as distribution of copies by rental and loan, without the written permission of the Author or limits authorized by the Intellectual Property Law.

Universitat Oberta de Catalunya
uoc.edu

Plot of the Master´s Theses

Title:	*Explaining Stock Exchange Prices*
Name of the author:	*Álvaro Antón Blanco*
Name of the associate Professor:	*Dra. Laia Subirats Maté*
PRA:	*Dra. Maria Pujol Jover and Dra. Teresa Sancho Vinuesa*
Deadline:	07/2017
Degree:	*Master in BI & Big Data*
Area of Master's Theses:	*Analysis of social networks*
Language of the final degree:	*English*
Key Words	*Data Capture, Data Processing, Databases, Social Media, Algorithms*

Abstract

In an interconnected world, the wingbeat of a hummingbird in one side of the world can cause a dramatic earthquake in the opposite side of the world, hence the important of political, social and financial events. In a panorama with increasing information exchange, headlines and news, seems to be the new soil for finance.

Within this context, this document aims to explain the stock market development in contrast with headlines and news data. For doing this, sentiment analysis and time series analysis along with other approaches are used.

After the analysis of the performance, it is going to be proven that there is no strong evidence, with the available data, to show a clear common trend between headlines and the stock index.

However, open remarks and recommendations are going to be given and explained for a future development and further researches and analyses. Also, the code is posted in Kaggle, for other users to develop the idea and to make the work completely open and reproducible.

The analyses and explorations were performed on R and the data were gathered from the Kaggle and yahoo finance.

En un mundo interconectado como el contemporáneo, el aleteo de un colibrí en un lado del mundo, puede causar un terremoto en el lado opuesto. Debido a esto, los eventos políticos y sociales por muy insignificantes que parezcan, pueden tener consecuencias inesperadas. Con este trasfondo, en un mundo con un panorama de aumento de intercambio de información, los titulares de prensa y las noticias son el nuevo nutriente del mundo de las finanzas.

Este documento trata de explicar el desarrollo en el mercado de capital en contraste con los datos de titulares de prensa. Para ello, se utilizará las herramientas de análisis de series temporales y el “sentiment analysis”.

Después de realizar el análisis, se podrá concluir que se puede rechazar dicha presencia de una tendencia común entre el desarrollo de los titulares y su impacto sobre el mercado de capital.

Sin embargo, se van a plantear y explicar ciertas recomendaciones para futuros desarrollos y posteriores análisis e investigaciones. El código ha sido subido a Kaggle para que otros usuarios puedan captar la idea y para hacer las líneas de trabajo reproducibles y transparentes.

Los análisis y las exploraciones han sido hechas en R, mientras que los datos se han obtenido de Yahoo Finance y Kaggle.

Index

Index of Figures

Index of Tables

1. Introduction

1.1 Background

Social media data is nowadays a substantial part of all generated data in the world. Today, almost everyone posts a text, a video or just music into one owns social media account.

Every day, Facebook grows by 500000 users, 6 new profile openings each second. Almost 1.3 thousand million people have a twitter account and each single minute 300 hours of video are uploaded to YouTube.

With that kept in mind, one should realize which proportion of data nowadays is shared instantly. And this data, again, can have multiple applications for the world to come.

Whether someone just asks for a loan and the BO of the financial institution is makes a real-time customer profiling or companies performing a market research, social data has become the world's new fuel. [1]

Nowadays online newspapers, *aka*. media industry, (and the corresponding information channels such as yahoo!, or google news among others) produce more information than ever before in a short time frame. And again, this information has an impact in the financial market.

On the one hand, the stock exchanges of all over the world have, whether we like it or not, a profound impact on the economy. On the other hand, we are not able to handle or generate an algorithm to determine the prices and fluctuations in the stock market. This document aims to get insights of such correlation using modern advanced analytics and sentiment analysis.

The goal is to find any correlation that can explain the development of stock market exchange prices with the news headlines.

1.2 Objectives

The objectives of this work are the following:

- Obtain news headlines and mine its content
- Text mining techniques
- To correlate the stock exchange index and the news data

- To visualize the results in multiple forms such as charts, graphs and tables. The visualization will include brute information and the processed “inter steps” like word clouds or already processed data.

1.3 Methodology

Keeping in line with the aim of the document, the data is going to be extracted from one of the main sources headlines and news, “Reddit WorldNews Channel”. The news are already packed into a dataframe in Kaggle [2] The text is going to be analyzed via a process of “text mining”.

The first thing that is going to be done, is to capture the data such as the S&P index out of the internet. For this reason, specific packages are going to be installed in R.

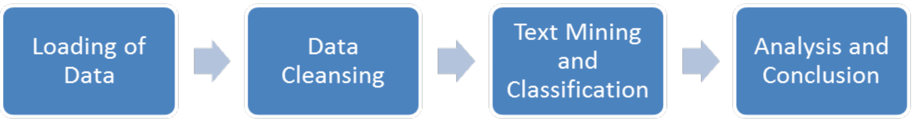
Afterwards a sentiment analysis on the news is going to be performed. The sentiment analysis is going to be performed with a package that quantifies polarity. An explanation on this topic is also going to be given.

The resulting variable is going to be correlated with the stock market exchange to see then if there is a clear correlation between the news and the price development.

The data that we are going to use for the stock exchange is the index of the S&P500, as one of the main indexes for the US. As mentioned above the news data that is going to be used is a “kaggle” data set, being this source also one of the main source of information for finance.

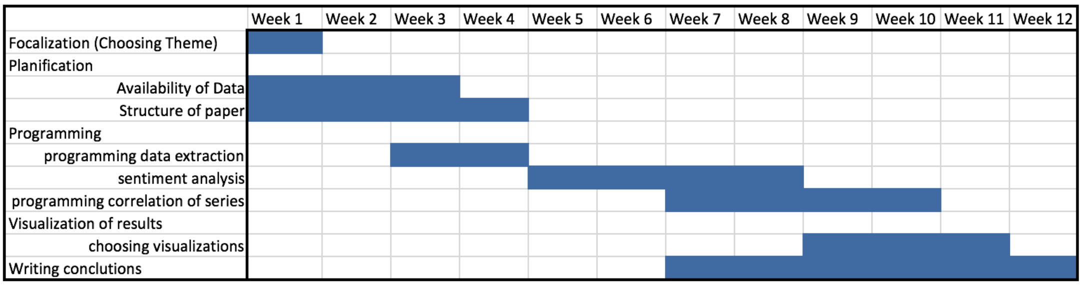
To make an unbiased analysis, the timeframe of the S&P500 and the Yahoo Finance Data is going to be the same i.e. the both data are going to be captured sequentially using the same period frame.

A glimpse of the methodology is pictured below;



All analyses are going to be performed with R and the code is going to be provided via *Github* (https://github.com/aablanco/stock_forecasting).

1.4 Work plan



	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12
Focalization (Choosing Theme)	■											
Planification												
Availability of Data	■	■	■									
Structure of paper	■	■	■	■								
Programming												
programming data extraction			■	■								
sentiment analysis					■	■	■	■				
programming correlation of series							■	■	■	■		
Visualization of results												
choosing visualizations									■	■	■	
Writing conclutions							■	■	■	■	■	■

1.5 Summary of the obtained products

Given that the nature of the analysis software, the following packages were installed to complete the analyses and data visualizations (additional packages were installed due to the existing dependencies on other libraries)

- Library to create Wordclouds
- Library for text mining
- Library for determining polarity and sentiment
- Library for calculating correlations and visualizing data
- Library for extracting(scraping) web data

1.6 Summary of the structure of the manuscript

In the next chapters, the analysis is going to be performed as well as the different steps to complete it. The next chapter "Data Analysis" contains the following under points:

Sub-chapter	Topic
Methodology	The programming methods and statistical analyses are going to be explained and referenced.
State of the Art	The recent status of the methods used to analyze the data
Data Description	Description of the data used for the analyses
Data Processing	Description of the development of the analyses for each part separately
Joining Data	Analyses for the joint data sources and extraction of information

Table 1. Content of document

Afterwards, to accomplish the purpose of this document, in the last chapter the conclusions regarding the analysis are going to be written and explained.

2. Data Analysis

2.1 State of the art

Text mining is a discipline that began during the mid-80s and has gained in the last years an important role with the improvement and the cost reduction of the processing power.

Until today, numerous researchers try to obtain a formula to quantify the linkage between the market development and people's discussions and thoughts. Within this context, numerous studies show the persistence of researchers for identifying patterns between discussions in social media and stock market indexes:

- Predicting Stock Market Indicators through Twitter "I hope it is not as bad as I fear" [3]
- Twitter mood predicts the stock market [4]
- Investor sentiment and the near-term stock market [5]

But moreover, prediction of the stock market returns has led to a huge amount of articles and scientific reports, and has been analyzed as something semi-magical and a lot of effort has been put over the past years into predicting the development of the market.

Since the early 80s a lot of news providers, being aware that news have a massive power on people's behavior, have been creating products that correlate and gives valuable insights information out of news and headlines such as Bloomberg and Reuters, but today's technology allow anybody in possession of a computer, to capture real-time data mine it.

Today's challenges of market prediction incorporate almost all knowledge areas such as psychology, computation, economics, statistics… One of the most repeated words in the past years in prediction of stock market is the so called "high frequency trading" (HFT).

HFT is characterized by algorithmic trading within short-term investment horizons. This is only possible subjugating all technological advances possible to make in mass low margins of profits and having a very fast algorithm. HFT is explained in the following articles:

- High-Frequency Trading and Price Discovery [6]
- Prediction based – high frequency trading on financial time series [7]

Although, not all HFT algorithms incorporate text mining and sentiment analysis to predict the stock market development, these algorithms have become more important in the last ten years.

But all in all, as mentioned before, the market faces a lot of challenges that can only be mastered including all sort of knowledge from other disciplines: Data mining (i.e. text mining), neuronal networks or machine learning algorithms have been adjusted and implemented to predict the returns of the stocks. The reader can find examples for this within these documents:

- Stock market prediction system with modular neural networks [8]
- A Hybrid Machine Learning System for Stock Market Forecasting [9]
- The use of data mining and neural networks for forecasting stock market returns [10]

2.2 METHODOLOGIES USED AND EXPLAINED

Text Mining

Slowly started during the 1980s, text mining is the process of deriving high quality information out of texts. This information is derived through exploration and devising of patterns with help of statistical methods.

Given that 80% of all information generated in text format, text mining has become in the past few years, together with the improvement of the computing power, very relevant and popular.

Being text mining, a data mining method, it has lot of practical applications such as social media monitoring, Business Intelligence or information management among others.

Papers which explain more in detail the text mining methods and also the package used in this analysis can be found at:

- The state of the art and the challenges [11]
- Text Mining [12]
- Text Mining in R [13]

Sentiment analysis

Sentiment analysis refers to the processing of natural language to systematically identify, extract and quantify states and subjective information. Sentiment analysis is used for a lot of purposes being CRM or Marketing departments, main users of this technique. In fact, the importance of sentiment analysis has an analogy to the development of social media.

One of the basic tasks in sentiment analysis is the classification of the polarity of the texts i.e. if the sentiment ("emotional status") captured in the texts are "angry", "sad" etc…

Actually, the challenges that face the polarity detection are to identify irony or jokes or subjectivity; however the algorithms and methods are accurate and can classify successful around 70 to 80 percent of the texts. Usually, sentiment analysis comes along with strong mathematical classification methods. A deep-dive into sentiment analysis (also the methodology of the package "sentimentr" used for this instance) can be found in the papers below:

- Package 'sentimentr' – Methodology for polarity in text mining using R [14]
- Sentiment Analysis and Opinion Mining - Bing Liu [15]

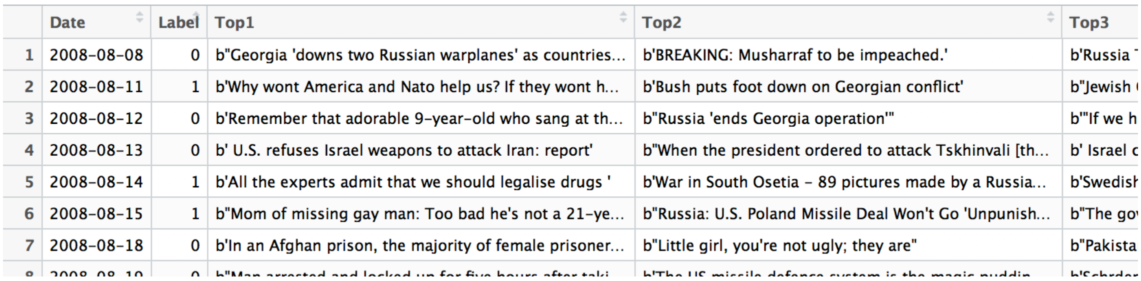
In this document, the goal of text mining is going to be set for a sentiment analysis, which is a process of the text mining. The sequence of the text mining analysis is going to be as follows:

Data in dataframe > Cleaning data and extracting Data tm package > Wordcloud/Sentiment Analysis

In our context, the Text information is the news extraction from the table “News”, which is gathered from the Kaggle web page and contains the top 25 news headlines for one day.

For our purpose, the package “tm” of text mining is going to be used. This package aims to offer all functionalities for performing a text mining analysis: data import, corpus handling, preprocessing and creation of term-document matrices – afterwards processing is also performed, so data is turned into information.

In the document case, as seen above, there are a lot of columns “Top1, Top2… Top25” containing a selected daily news. The first step that is going to be performed is to concatenate the string variables and to make one out of the 25 variables.



	Date	Label	Top1	Top2	Top3
1	2008-08-08	0	b"Georgia 'downs two Russian warplanes' as countries...	b'BREAKING: Musharraf to be impeached.'	b'Russia '
2	2008-08-11	1	b'Why wont America and Nato help us? If they wont h...	b'Bush puts foot down on Georgian conflict'	b"Jewish (
3	2008-08-12	0	b'Remember that adorable 9-year-old who sang at th...	b"Russia 'ends Georgia operation'"	b'"If we h
4	2008-08-13	0	b' U.S. refuses Israel weapons to attack Iran: report'	b"When the president ordered to attack Tskhinvali [th...	b' Israel c
5	2008-08-14	1	b'All the experts admit that we should legalise drugs '	b'War in South Osetia - 89 pictures made by a Russia...	b'Swedish
6	2008-08-15	1	b"Mom of missing gay man: Too bad he's not a 21-ye...	b"Russia: U.S. Poland Missile Deal Won't Go 'Unpunish...	b"The go
7	2008-08-18	0	b'In an Afghan prison, the majority of female prisoner...	b"Little girl, you're not ugly; they are"	b"Pakista
8	2008-08-19	0	b"Man arrested and locked up for five hours after taki...	b'The US missile defence system is the magic puddin...	b'Schrder

Figure 1. Capture of the headline dataframe

Having done this, invoking the “tm” package the text-data is “cleaned” so the program can create the term-document-matrix (TDM) for further analyses.

After having performed such operations, a word cloud is going to be created from the TDM and a sentiment analysis[1] is going to be executed.

Web Scraping

Within the context of sentiment analysis, the first attempt to perform the data gathering was the web scrapping. Even though this methodology, finally, was not used, it is important to mention the methodology.

[1] See polarization analysis

Within the last decades, a vast amount of information is being increasingly posted on the web, whether opinions or reviews or key facts, a huge amount of information is being shared.

Keeping this in mind, web scraping consists in the reverse process of uploading information: Getting data from the web i.e. turning the information posted in the web into structured data.

Web scraping are a set of methods for extracting data from web sites. Taking into consideration this definition manual "copy and paste" would be considered the most rudimentary form of web scraping, although this technique is not considered by all authors as web scraping:

- Exploiting web scraping in a collaborative filtering based approach to web advertising [16]

Currently a set of software has been developed for the purpose of gathering information available in the web. Also, numerous of enterprises have focused as providers of "web scraping services" turning this technique into a "data as a service" (*DaaS*).

In the context of this document, the issue of web scraping, is that the server where the web is hosted, is not always configured the same way as others. For instance, when running a web-scraping-script, the server may block or blacklist an IP while the scrapping is being done, so essentially one may get few headlines.

For preventing this, tools (or enterprises, as mentioned above) are being used to scrape the web properly obtaining the desired data, which are adjusted ad-hoc, so the end user/customer get the desired data.

Given that an important number of headlines were needed, to obtain robust results, the solution to the issue, was to download a kaggle dataframe[2] containing the 25th most important news headlines related to financial data.

Univariate Analysis of Series

The nature of the S&P500 Data is a time series. Given the features of a time series, some techniques of time series analysis are going to be used as well within the document.

A profound manual for time series analysis can which is recommended to follow all the concepts is

- Time Series Analysis – James D. Hamilton [17]

[2] https://www.kaggle.com/aaron7sun/stocknews

The main focus, is to handle the series with the methods described such as taking differences and explaining variations. Time series analysis techniques are well known to statisticians and vastly used nowadays to handle data.

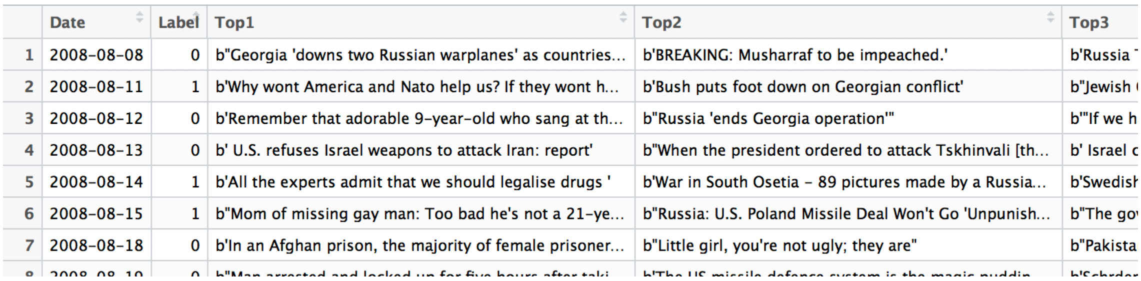
3. Data Gathering Process

3.1 Data Description

The data gathering was a substantial part of the work, given the difficulty of obtaining data news and headlines to mine. The first attempt was to perform the so called "web-scrapping", but due to technical reasons explained above the option of getting the data directly from Kaggle was chosen.

The Kaggle Dataset (the link to the dataframe is provided also above) which contains the top 25 headlines within the range of 2008-08-08 to 2016-07-01. The source of the data is the reddit word news channel[3].

A glimpse of the information of the dataset was given above,



	Date	Label	Top1	Top2	Top3
1	2008-08-08	0	b"Georgia 'downs two Russian warplanes' as countries...	b'BREAKING: Musharraf to be impeached.'	b'Russia
2	2008-08-11	1	b'Why wont America and Nato help us? If they wont h...	b'Bush puts foot down on Georgian conflict'	b"Jewish
3	2008-08-12	0	b'Remember that adorable 9-year-old who sang at th...	b"Russia 'ends Georgia operation'"	b"'If we h
4	2008-08-13	0	b' U.S. refuses Israel weapons to attack Iran: report'	b"When the president ordered to attack Tskhinvali [th...	b' Israel
5	2008-08-14	1	b'All the experts admit that we should legalise drugs '	b'War in South Osetia - 89 pictures made by a Russia...	b'Swedish
6	2008-08-15	1	b"Mom of missing gay man: Too bad he's not a 21-ye...	b"Russia: U.S. Poland Missile Deal Won't Go 'Unpunish...	b"The go
7	2008-08-18	0	b'In an Afghan prison, the majority of female prisoner...	b"Little girl, you're not ugly; they are"	b"Pakista
8	2008-08-19	0	b"Man arrested and locked up for five hours after taki...	b'The US missile defence system is the magic puddin...	b'Schrder

Figure 1. Capture of the headline dataframe

As one can see, the dataframe has the following structure

Name	Type	Description
Date	Numeric – Date Format	Date
Top1 – Top25	Character	Headlines – Top 25 of day
Label	Dummy	This variable is not relevant for the document

Table 2. Description of headlines dataframe

Also, the dataframe has dimensions 1989x27 variables and contains daily news.

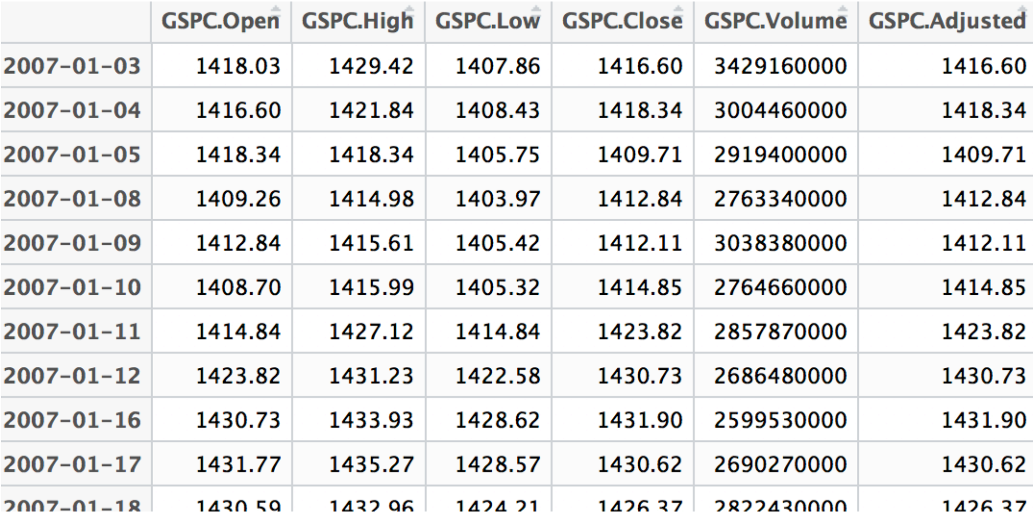
On the other hand, we´ve obtained thanks to a package (quantmod[4]) the financial data. As a note, and to a certain extend we can call this "web

[3] https://www.reddit.com/r/worldnews/?hl=

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

scrapping” because there is a retrieval of information of the yahoo finance server.

Nevertheless, the obtained data consists in the different magnitudes captured for the maximum period available within the Yahoo Finance server. A glimpse of the data can be seen with the following screenshot;



	GSPC.Open	GSPC.High	GSPC.Low	GSPC.Close	GSPC.Volume	GSPC.Adjusted
2007-01-03	1418.03	1429.42	1407.86	1416.60	3429160000	1416.60
2007-01-04	1416.60	1421.84	1408.43	1418.34	3004460000	1418.34
2007-01-05	1418.34	1418.34	1405.75	1409.71	2919400000	1409.71
2007-01-08	1409.26	1414.98	1403.97	1412.84	2763340000	1412.84
2007-01-09	1412.84	1415.61	1405.42	1412.11	3038380000	1412.11
2007-01-10	1408.70	1415.99	1405.32	1414.85	2764660000	1414.85
2007-01-11	1414.84	1427.12	1414.84	1423.82	2857870000	1423.82
2007-01-12	1423.82	1431.23	1422.58	1430.73	2686480000	1430.73
2007-01-16	1430.73	1433.93	1428.62	1431.90	2599530000	1431.90
2007-01-17	1431.77	1435.27	1428.57	1430.62	2690270000	1430.62
2007-01-18	1430.59	1432.96	1424.21	1426.37	2822430000	1426.37

Figure 2. Capture of financial dataframe

The structure of the data can be seen in the table below

Name	Type	Description
Open	Numeric	Value at opening market
High	Numeric	Highest value at session
Low	Numeric	Lowest value at session
Close	Numeric	Value at closing market
Volume	Numeric	Volume of the S&P market
Adjusted	Numeric	Adjusted index
Date	Numeric – Date	Date of the session

Table 3. Description of financial dataframe

The dimension of the xts-type dataframe is 6 variables and 2594 observations, given that the date range (2007-01-03 to 2017-04-21) is more extend that the dataframe shown before.

The main variable which one is going to work with is “Adjusted Index” variable, which is represented below. The main transformation which is going to be performed is the variation in logarithms.

The reason for taking the differences is, because the process is not stationary, that means, that it has a unitary root and is not stable. We can see a clear trend among time.

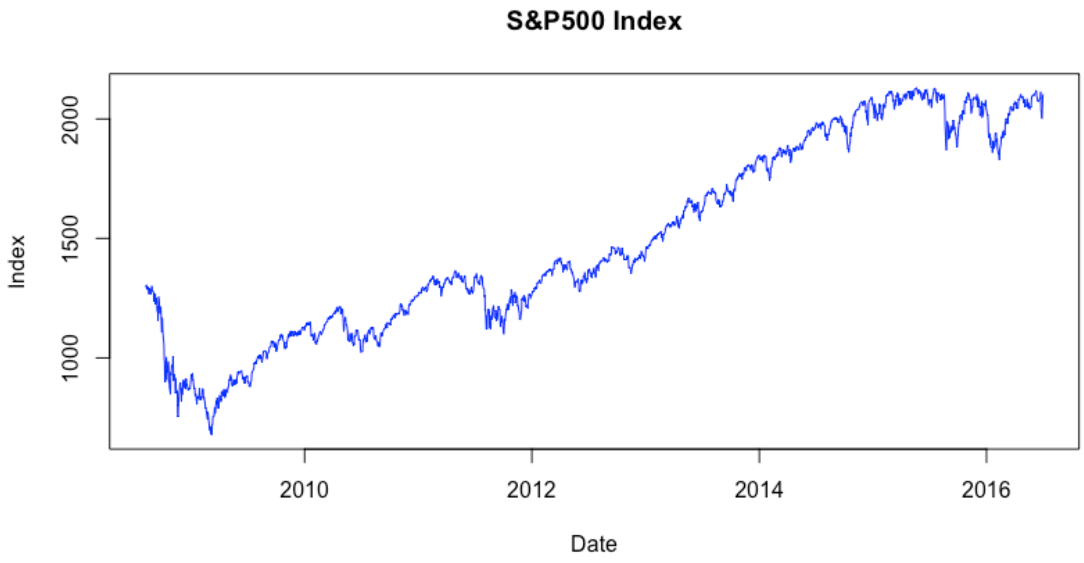


Figure 3. Plot of the S&P500 Dataframe

3.2 Data processing

Given that the dataframes are not located within the same ranges, the data must be firs homogenized and brought to their common denominator. In this, case the range of the analysis, is going to be reduced to 2008-08-08 to 2016-07-01.

The next thing to do, once the data is gathered, is to perform a text mining analysis. To do this, all the variables in form of strings are joint in a variable for each day. Afterwards, a corpus is created and simplified erasing punctuation, numbers, using just minuscule fonts and strange characters. Moreover, words that are not useful are also eliminated from the corpus, such as adverbs or prepositions.

Drawing a chart that represent the frequency of words, one may see the frequency of words.

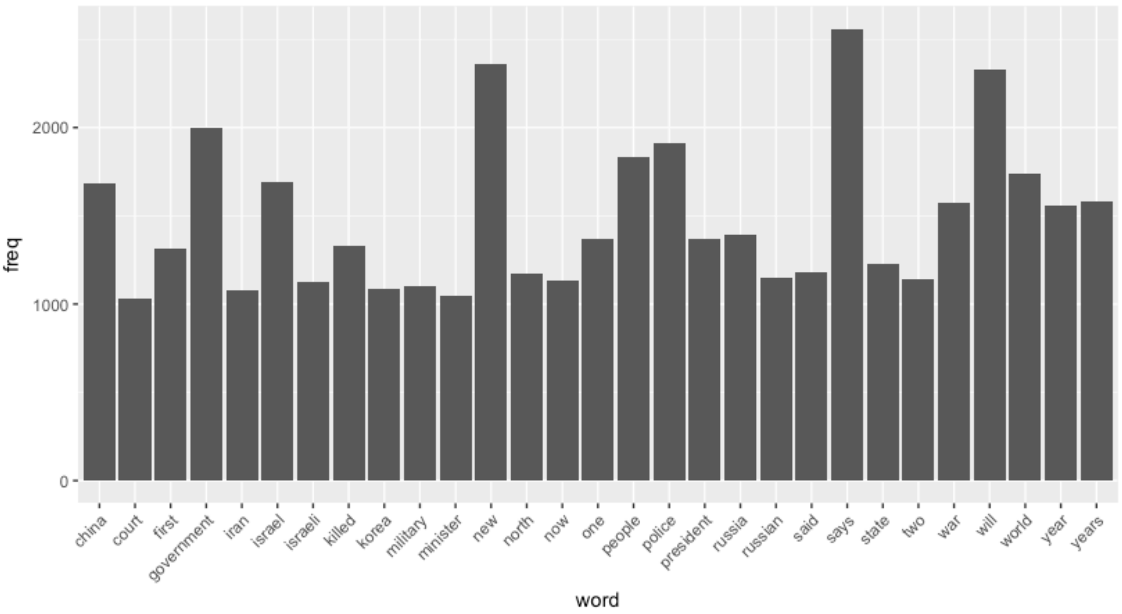


Figure 4. Frequency Plot

One may see, that the words contained in the corpus make reference to geopolitical events as well as politics in general.

The wordcloud of the corpus can also give a more clarifying hint to this matter.

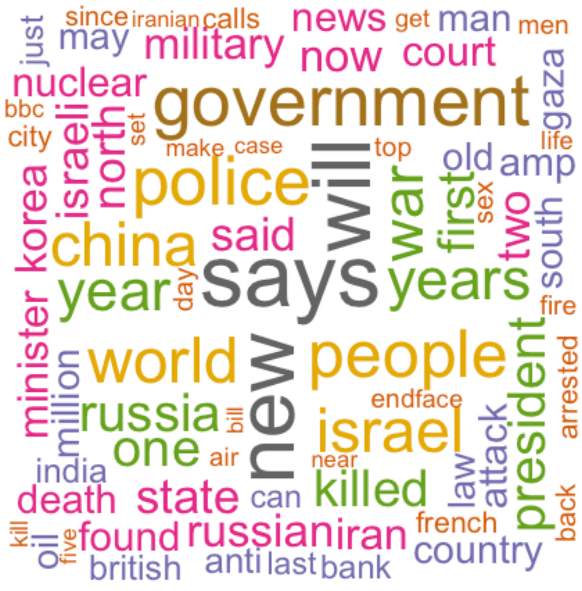
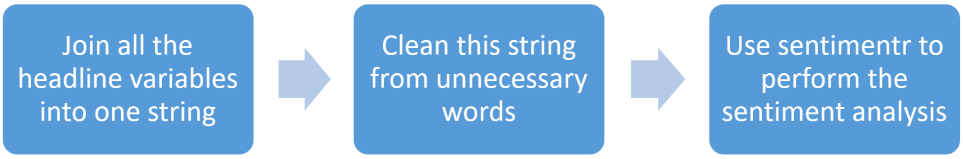


Figure 5. Wordcloud Plot

As commented before, the wordcloud gives maybe a more expresive hint of the subject.

The main thing to do after the creation of the frequencies and the wordcloud to roughly see what the text is handling of, is, and what's more important due to the subject of this document, to capture the sentiment of each and one of the headlines.

For this purpose the following steps were executed:



Join all the headline variables into one string

Clean this string from unnecessary words

Use sentimentr to perform the sentiment analysis

For consistency purposes, the corpus is used as the clear input strings for the analysis and merged with the “News” dataframe.

A loop is build afterwards among the dataframe to calculate the polarity with the help of the sentimentr[5] package. The sentimentr package calculates a polarity score which takes positive, negative and 0 values being around “-1,-2” negative/ or an extremely adverse sentiment, and “1 or 2” an extremely positive sentiment. Also it takes value 0 which means neutral or 99 as sarcasm.

Performing the sentiment analysis for the headlines variable and ploting the histogram of the polarity score variable,

[5] https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf

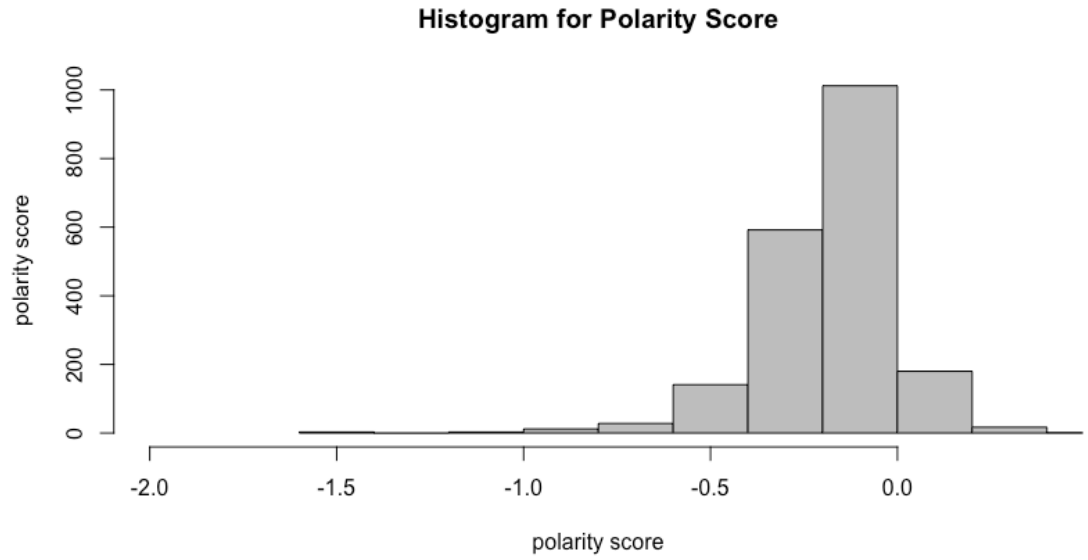


Figure 6. Histogram Plot

The histogram reveals, that the mass of the sentiment is in the negative area. This means, that the mayority of news tend to be pesimistic about the corresponding topic. However this may be because journalists try to be cautious about information.

The second dataframe contains the variables described in the „data description“ chapter. Given that the purpose is to try to explain stock prices, there is just the need to spread the dataframe and take just the variable “Date” and “Adjusted index.”

In the classic theory, the time series can all be divided by a trend component, a seasonal component and an irregular component:

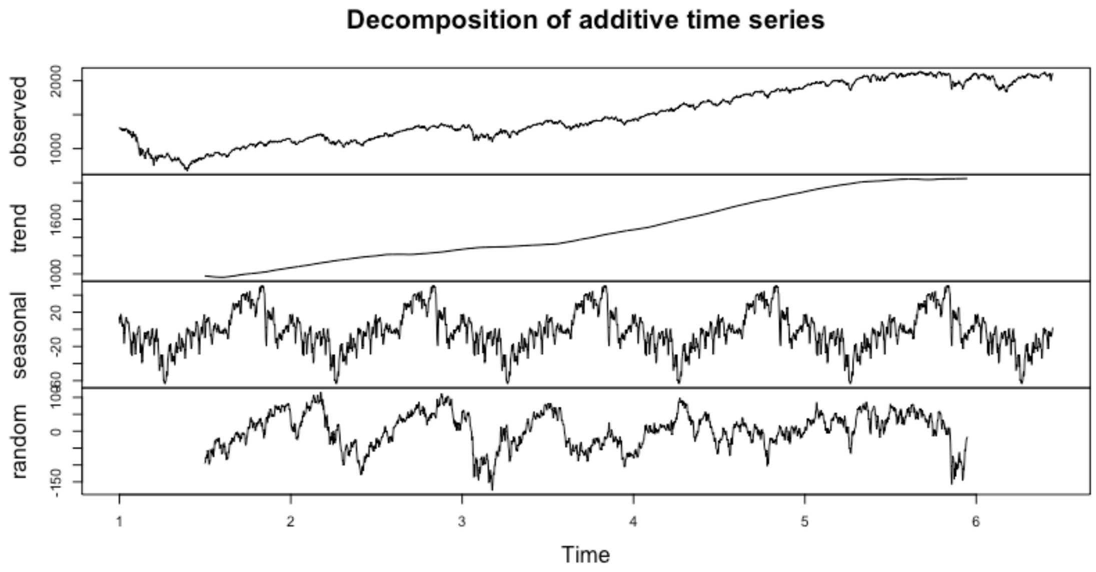


Figure 7. Time series decomposition

The seasonal and irregular (or random) component, look pretty much the same and especially the seasonal component, seems to remain very stable among time. To get more insights about the series, the autocorrelation, is shown

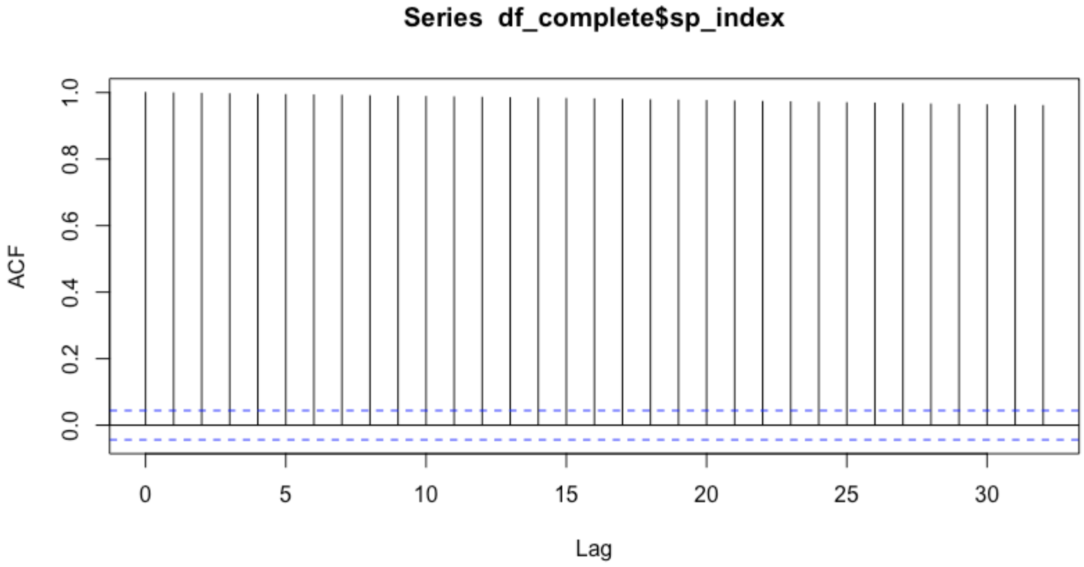


Figure 8. ACF 1

As one may see the process is not stable among time i.e. it has a unitary root (trend). Therefore to eliminate the trend and smooth the series, logaritmic differences were taken leaving the data like this;

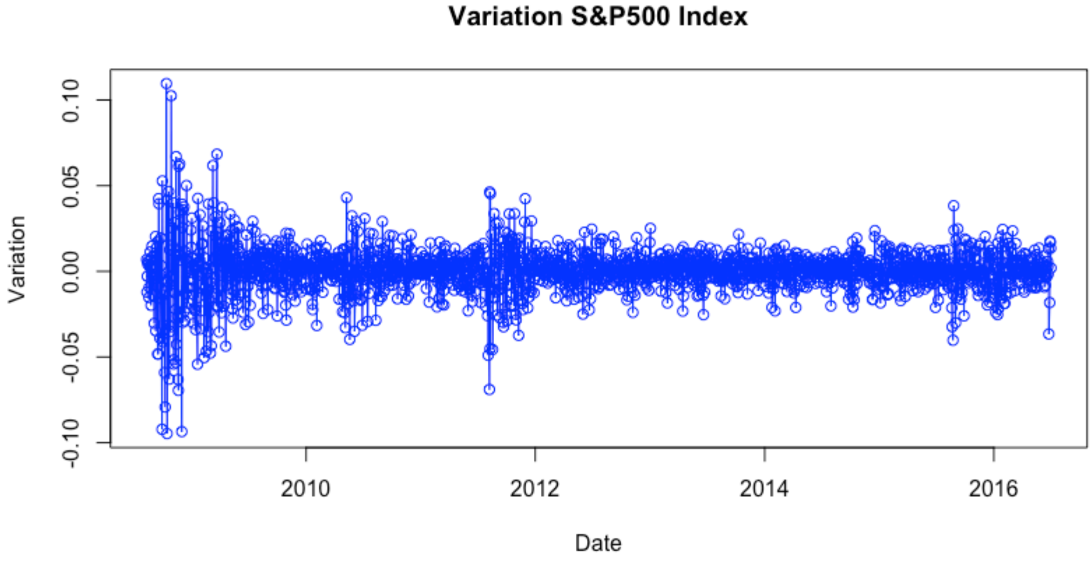


Figure 9. Variation of index

Computing again, the autocorrelation for the variation, one may see that the process is now handeable and stationary.

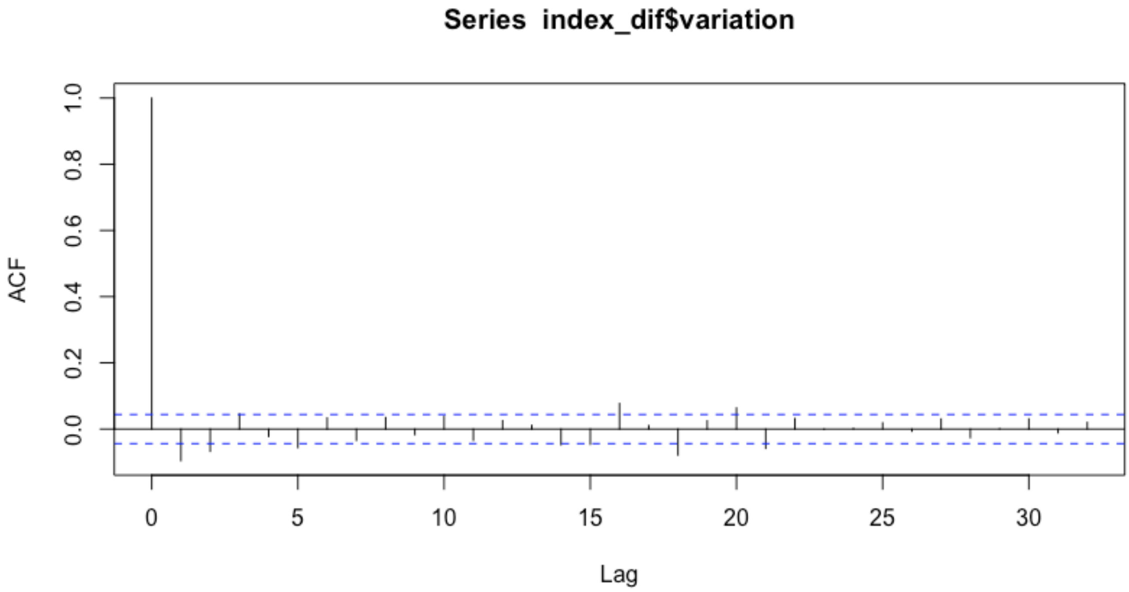


Figure 10. ACF 2

3.3 Joining the Data and correlations

The interesting fact about this document is to see essentially how the variation reacts tho the headlines. For this purpose a new dataframe is created with all relevant variables

Name	Type	Description
Date	Numeric	Date
All_News	Character	String – All headlines
Sentiment	Numeric	Polarity Score
SP_Index	Numeric	S&P500 Index Adjusted
Variation	Numeric	Variation of S&P500 Index

Table 4. description of joint dataframe

Now some plots to show the data,

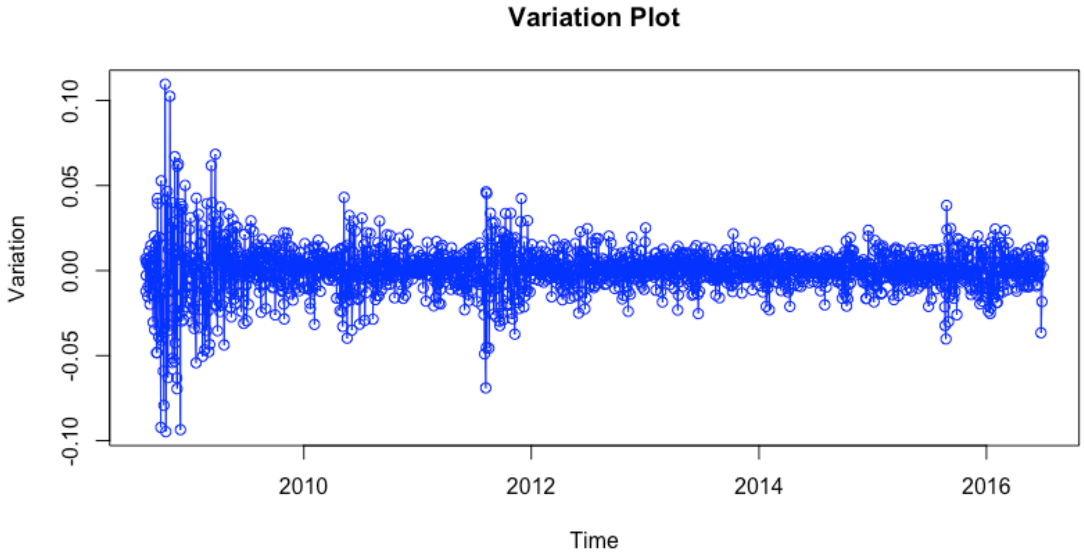


Figure 11. Variation of index

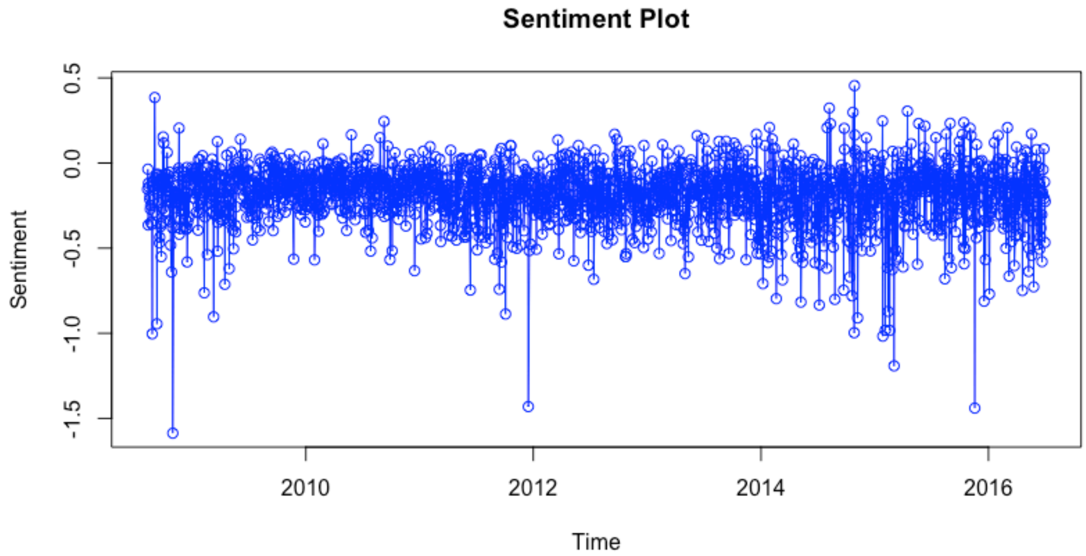
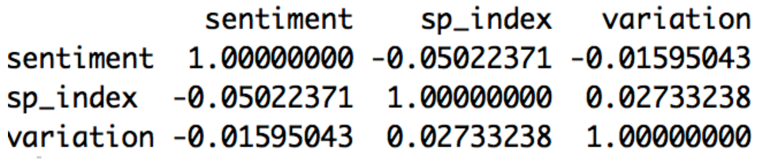


Figure 12. Sentiment Plot

Now, the plot of the correlation shows that there is no correlation what so ever between the polarity score and the variation of the stock market.



```
            sentiment    sp_index   variation
sentiment  1.00000000 -0.05022371 -0.01595043
sp_index  -0.05022371  1.00000000  0.02733238
variation -0.01595043  0.02733238  1.00000000
```

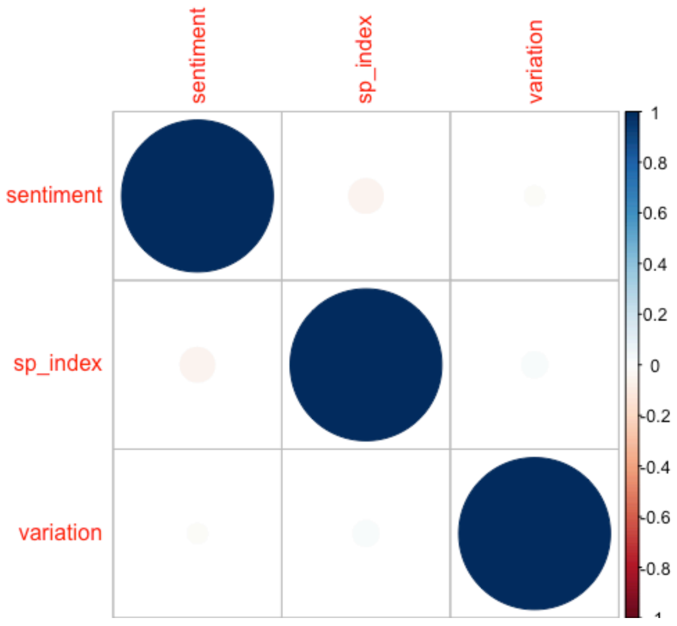


Figure 13. correlation plot between sentiment, index and variation

However, it would be too soon to extract some conclusions, since there is not a complete understanding how news can affect the stock market development. For instance, it could be that news headlines have a delay compared to the cycle of the index. Therefore, for analysis completion 10 lags of the index is going to be added (ten lags would be an impact of a headline after 10 days of being happened – this time frame was just set up because 10 days appears as the most reasonable time frame for a cause-consequence-relation).

Hence, lags of the variation were computed,

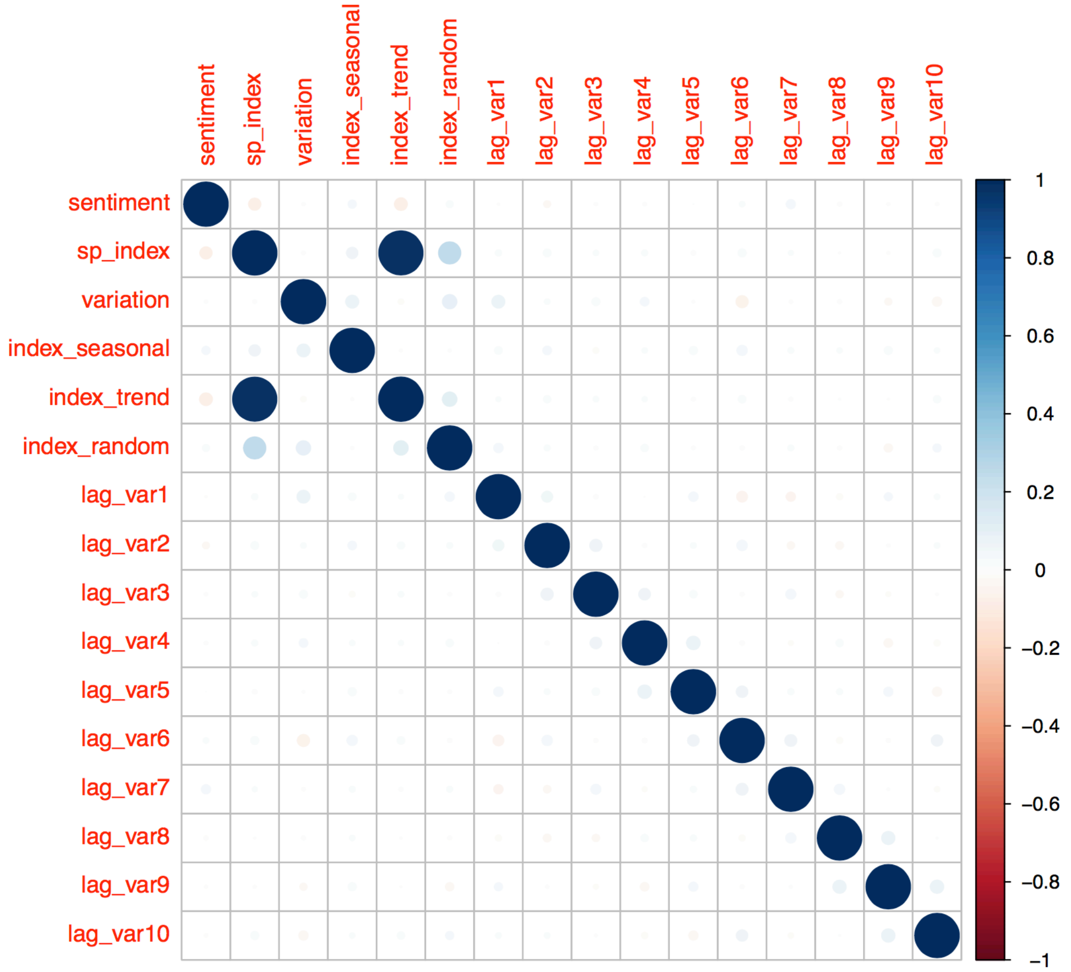


Figure 14. Correlation plot of all variables

The only strong correlation in the matrix is between the S&P index and its own trend component.

Afterwards, there is a check if the sentiment series is correlated with the differences of the trend component of the index series, i.e. if the sentiment has some sort of correlation with the variation of the trend.

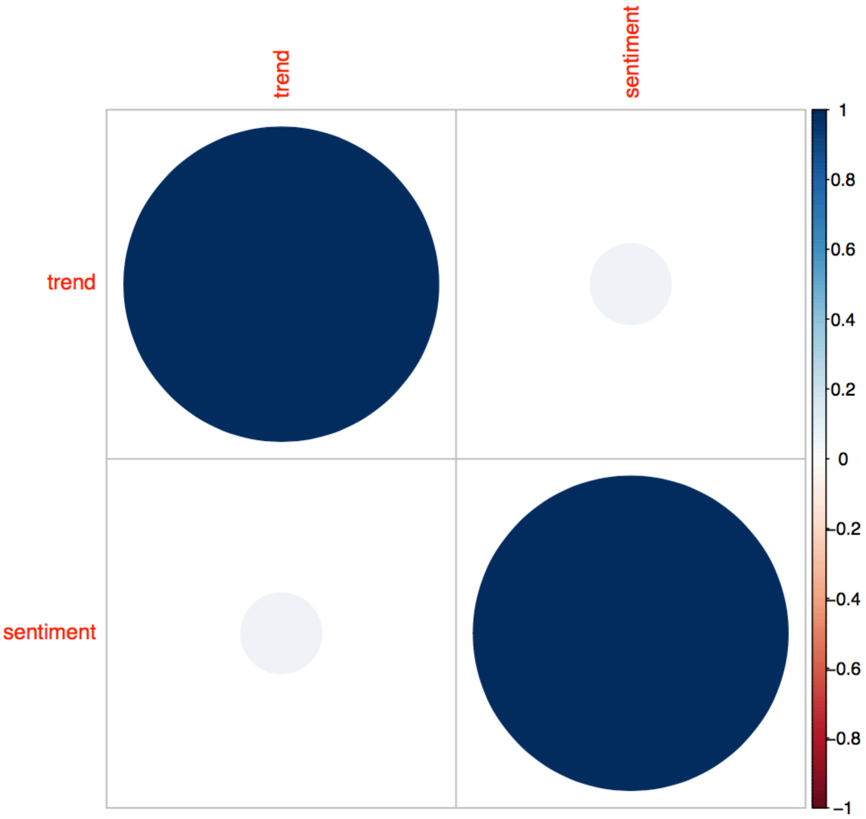


Figure 15. Correlation Plot of sentiment and trend component

The chart above, shows a weak correlation between the variation of the trend component and the sentiment series. Not even a quotient of 0.07 shows that both series are independent from another.

The next check, is to see if there is a co-operation between the variation of the index and the sentiment. The idea behind this is to capture “positive” or “negative”, i.e to create a binary variable for the sentiment and another for the variation of the index.

This would eliminate some effects behind the magnitudes, and make them more comparable. However, in the chart below it is also shown, that both dummy series are also independent from another.

The last computed chart, the cross-correlation-function is shown. Following the same logic as with the ACF (Auto Correlation Function), the cross correlation captures lineal dependencies among the variables instead of the variable with itself (ACF).

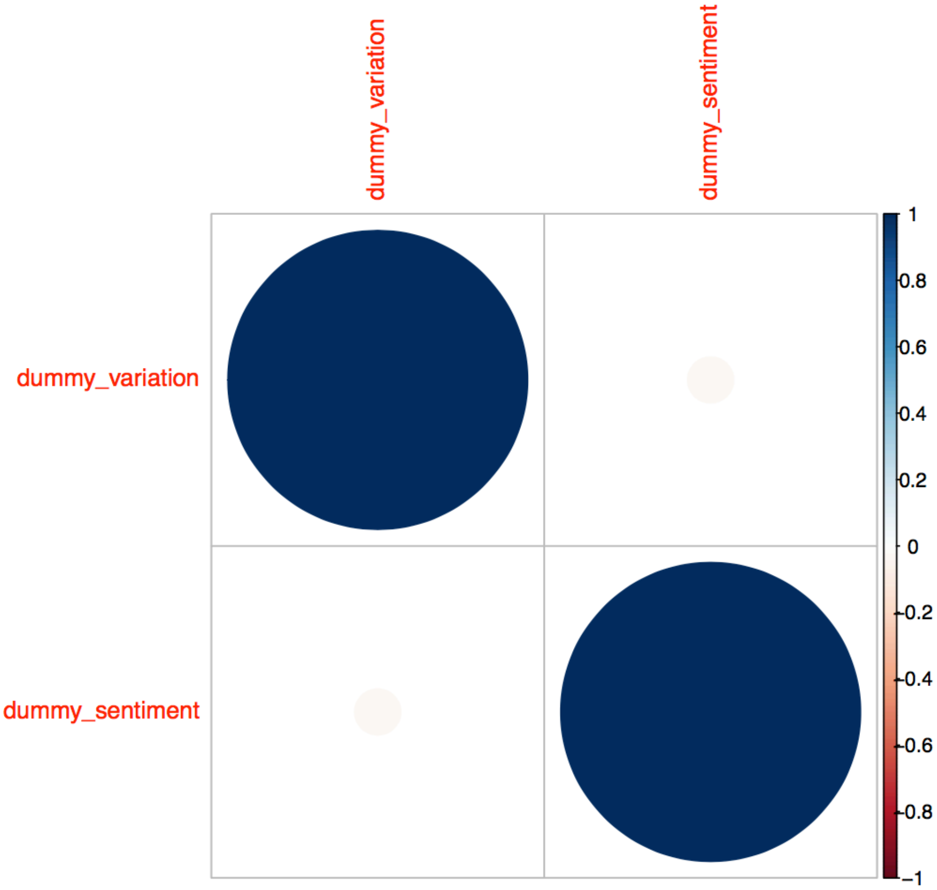


Figure 16. Correlation Plot of dummy variables

In the last analysis for searching a relation, the cross-correlation function (covariance function) is computed. As one may see there is no strong evidence at a 5% level that there is a correlation between both series.

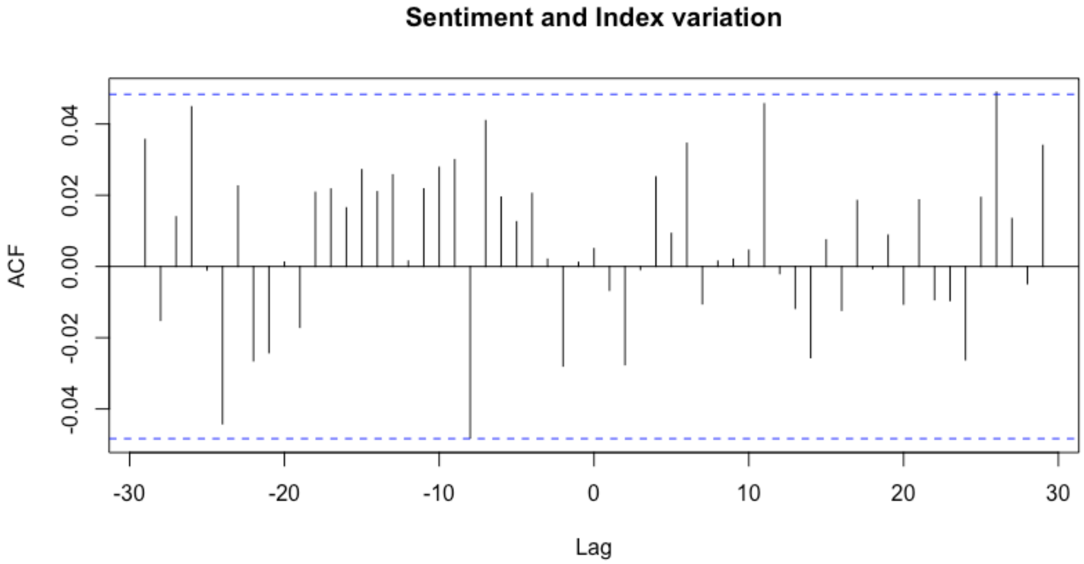


Figure 17. Cross Correlation Plot

4. Conclusions

The analysis above shows, that there is no strong evidence neither for correlation nor for any other type of relation between the news and the stock market.

But furthermore, the computation of the correlation between the different components of the index time series and the lags of the variations show that there is absolutely no evidence for correlation between the sentiment score (or its individual components) and any type of the lagged variation of the index.

Afterwards, plot of the correlation between the trend component and the sentiment just shows a rough 0.07 of correlation for these two variables, and it becomes worse for a computation of dummy variables that captures the effects of the variation of the index and the sentiment.

Computing the cross-correlation function shows, that at a significance level of 5% one can reject the null hypothesis that both series are correlated at a lag level 0. However, some spikes turn between lag -5 to -10. This would mean that the headlines precede the movement of the index, and indeed show some sort of relationship, but again, it’s too low to affirm such fact and even more if one keeps in mind the analyses made before.

On the other hand, the charts of the headlines show that the subject of the headlines is geopolitical/political mainly, rather than actual finance news headlines. So, in this extend there is a clear limitation of the conclusions that one can take for this topic.

With help of the Wordcloud, one can briefly get the idea of the topics that are listed within the headlines, and again, there are not the ideal ones for deal with financial data.

Briefly recalling, that the aim of this document was to explore the relation between the stock market and the sentiment of the headlines, the conclusion is, that there is no evidence (with the data used) that this correlation exists.

Keeping in mind the initial aim of the document, the lessons learned are that to show the correlation between sentiment analysis and the actual world, is much harder than it seems. There are other things that may alter and important to keep in mind for the analysis, such as the type of information where to measure the polarity score or exogenous effects such as economic cycle etc.

Now taking into consideration the planning and schedule composed for this document, there was the firm conviction to follow it as much as the author could, however, it was not possible to adopt the planning 1:1 given that the original idea was to code a web-scraping script and it generated delay within the deliverables.

Following this last argument, there were issues as well with the "StemCompletion" package in R, given that it clusters words to avoid to using roots. Some words were falsely clustered and some other words were just falsely outputted and appeared like "policy" instead of "police" for instance. Also, the impossibility of code a web scrapping script had a negative impact for both the conclusions of this document and for the analyses performed, given that the aim was to create a solid analysis of financial data.

Even though it's difficult to say that with other data there are different conclusions. There is a real struggle between the researchers if stock market can be predicted via sentiment analysis or, there are some pre-requisites that must be met on beforehand not to speak about the quantification and the lack of a consistent model that proves this relationship.

All in all, this document is a good set up point to sharpen the analysis taking more data into consideration such as social network data (Twitter, Facebook, etc.) or even enrich the existing data with other polarity scores computed in a different way.

Also, one of the questions still pending, and that took very long to clarify, was to code and develop a proper web scrapper which can gather information from one of the main source of finance headlines: "Yahoo Finance". Ideally, the scraper should be able to get data of other sources, this would improve the conclusions of the analysis.

Additionally, the code was uploaded to Kaggle and made public under the address https://www.kaggle.com/alvaroanton/headlines-and-s-p500-index for the community to improve the analysis and as a contribution to the data source, making the results reproducible and the data available.

5. Glossary

1. **S&P500**: The index known as S&P500 is an American Stock Market Index based on the market capitalization of 500 large companies having a common stock listed in the NYSE and the NASDAQ. It is one of the most followed equity index.
2. **Term-Document Matrix**: A document term matrix is a matrix that describes the frequency of terms in a collection of documents i.e. in a text or set of texts. For further information: https://en.wikipedia.org/wiki/Document-term_matrix
3. **DaaS (Data as a Service)**: analogously of "SaaS" Software as a service, DaaS builds on the concept that the product can be provided on demand, in this case "Data".
4. **R**: Analytical software which is going to be used to process the data.
5. **Polarity score**: refers to a score computed taking into consideration key words in a sentence
6. **ACF**: the ACF is the acronym for auto correlation function. The autocorrelation function tells you pattern of the data, and correlates the series with a lagged copy of itself.

6. Bibliography

1. Smith, K. 96 estadísticas y datos increíbles de las redes sociales para 2016. Available online: https://www.brandwatch.com/es/2016/08/96-estadisticas-redes-sociales-2016. (Accessed on 04/04/2017)
2. News data https://www.kaggle.com/aaron7sun/stocknews (Accessed on 03/05/2017)
3. Predicting Stock Market Indicators Through Twitter http://www.sciencedirect.com/science/article/pii/S1877042811023895 (Accessed on 04/05/2017)
4. Twitter mood predicts the stock market http://www.sciencedirect.com/science/article/pii/S187775031100007X (Accessed on 04/05/2017)
5. Investor sentiment and the near-term stock market http://www.sciencedirect.com/science/article/pii/S0927539803000422 (Accessed on 04/05/2017)
6. High-Frequency Trading and Price Discovery https://academic.oup.com/rfs/article-abstract/27/8/2267/1582754/High-Frequency-Trading-and-Price-Discovery (Accessed on 05/05/2017)
7. Prediction based – high frequency trading on financial time series https://pp.bme.hu/eecs/article/viewFile/7165/6177 (Accessed on 05/05/2017)
8. Stock market prediction system with modular neural networks http://ieeexplore.ieee.org/abstract/document/5726498/ (Accessed on 05/05/2017)
9. A Hybrid Machine Learning System for Stock Market Forecasting http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.307.6153&rep=rep1&type=pdf (Accessed on 05/05/2017)
10. The use of data mining and neural networks for forecasting stock market returns http://www.sciencedirect.com/science/article/pii/S0957417405001156 (Accessed on 05/05/2017)
11. The state of the art and the challenges http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf(Accessed on 05/05/2017)
12. Text Mining https://en.wikipedia.org/wiki/Text_mining (Accessed on 02/05/2017)
13. Text Mining in R https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf (Accessed on 02/05/2017)
14. Package ‘sentimentr’ – Methodology for polarity in text mining using R https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf (Accessed on 02/05/2017)
15. Sentiment Analysis and Opinion Mining Sentiment Analysis and Opinion Mining Bing Liu Synthesis Lectures on Human Language Technologies, May 2012, Vol. 5, No. 1 , Pages 1-167 (doi: 10.2200/S00416ED1V01Y201204HLT016)
16. Exploiting web scraping in a collaborative filteringbased approach to web advertising http://www.sciedu.ca/journal/index.php/air/article/view/1390/1115 (Accessed: 04/06/2017)
17. Time Series Analysis Time Series Analysis - James D. Hamilton , Princeton University Press, 1994, Pages (1-43,544-571)