Explaining Stock Exchange Prices using supervised learning and sentiment analysis

By Álvaro Antón Blanco

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

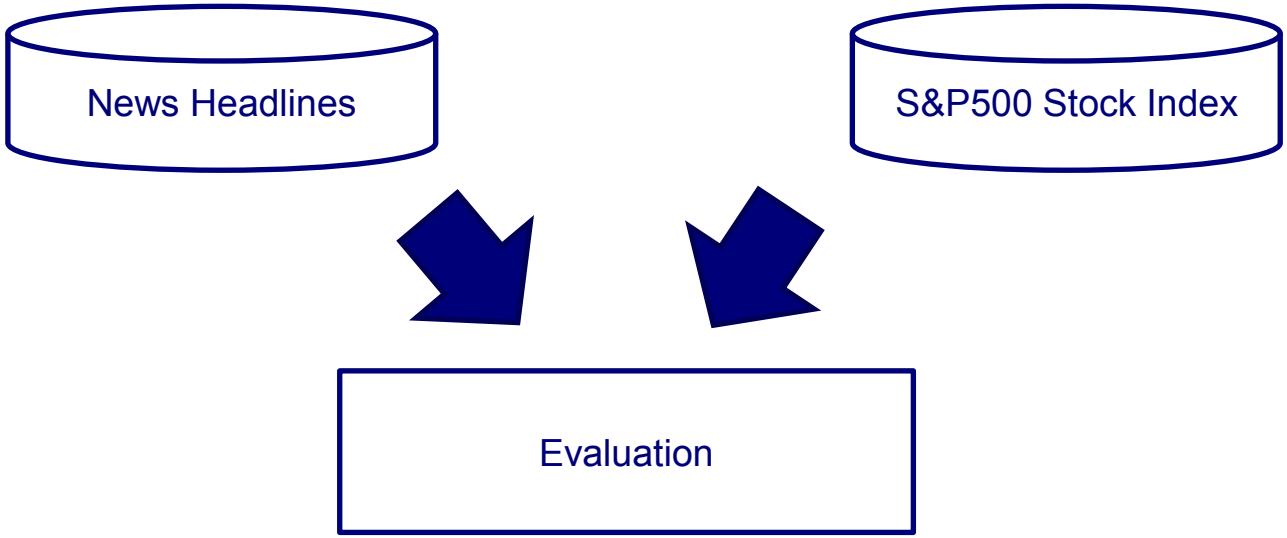
July 2017

Index

Ja

1. Aim of the Document

- The new development in analytics make the data the new soil for stock finance
- The interconection of the information has profound impacts in the economy
- The purpose of this document is to gather news information and see how the stock market behaves and reacts to the news headlines.



Index

2. Introduction

Within the context mentioned above, 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.

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.

Index

3. State of the Art

- Web Scrapping techniques: Exploiting web scraping in a collaborative filtering based approach to web advertising [16] , WebScrapping as a service
- Sentiment Analysis: Sentiment Analysis and Opinion Mining - Bing Liu [15]
- Textmining/ i.e. Polarity measure: Predicting Stock Market Indicators through Twitter “I hope it is not as bad as I fear” [3] , Twitter mood predicts the stock market [4]
- Papers which handle emotion handling in the context of stock markets: Stock market prediction system with modular neural networks [8] ,A Hybrid Machine Learning System for Stock Market Forecasting [9]

Index

4. Methodology

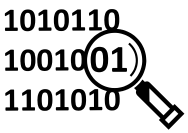
Data Sources

- Different attempts to obtain Data – Web Scrapping
- Two main sources:

Contains the top-25 Headlines for one day – The data was gathered from Reddit and posted on Kaggle

With the package quantmod (R) historical data for the S&P500 index is gathered

4. Methodology



1. Cleaning and selecting relevant variables in both datasets
   - completing cases, deleting duplicate observations,…

2. Analysis of variables of the dataset
   - Analysis of the important features of the variables of the dataframes

3. Analysis of the joint variables
   - Joining dataframes and joint analysis of the variables

4. Conclusions
   - Pointing the conclusions and suggestions

Index

5. Data Analysis

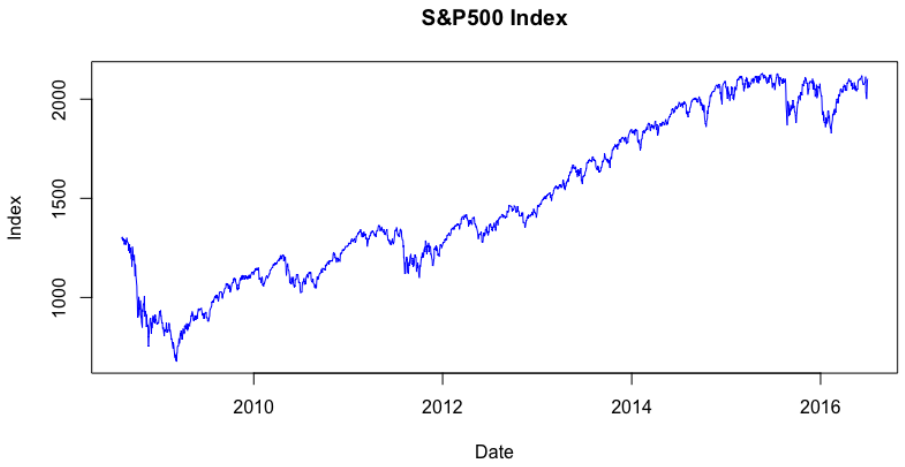
News Headlines Dataframe

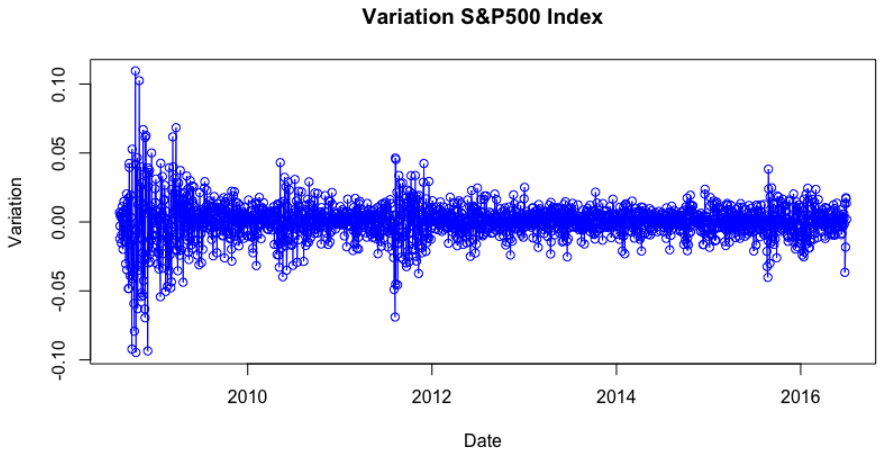
Name	Type	Description
Date	Numeric – Date Format	Date
Top1 – Top25	Character	Headlines – Top 25 of day
Label	Dummy	This variable is not relevant for the document

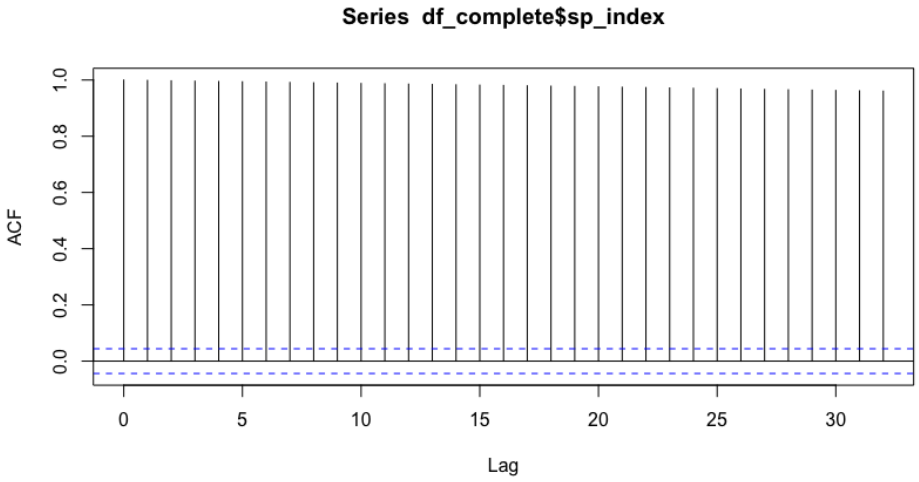
S&P500 Dataframe

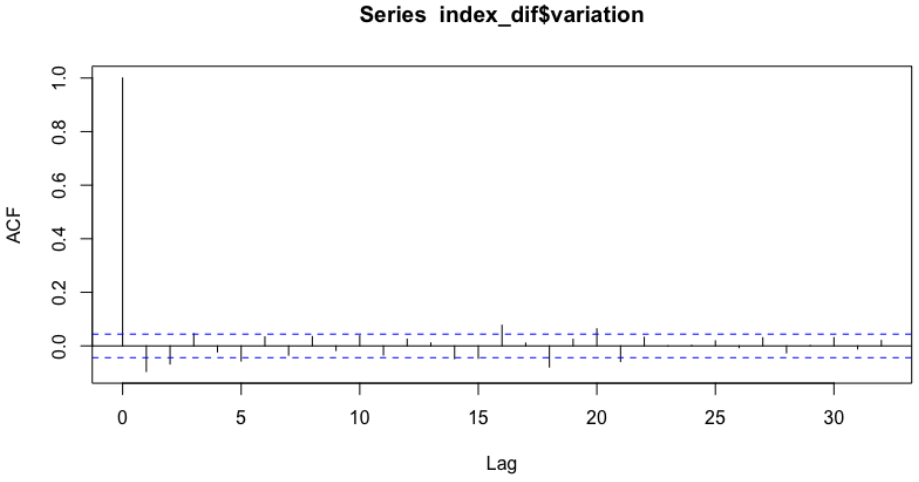
Name	Type	Description
Open	Numeric	Value at opening market
High	Numeric	Highest value at session
Low	Numeric	Lowest value at session
Close	Numeric	Value at closing market
Volume	Numeric	Volume of the S&P market
Adjusted	Numeric	Adjusted index
Date	Numeric – Date	Date of the session

5. Data Analysis



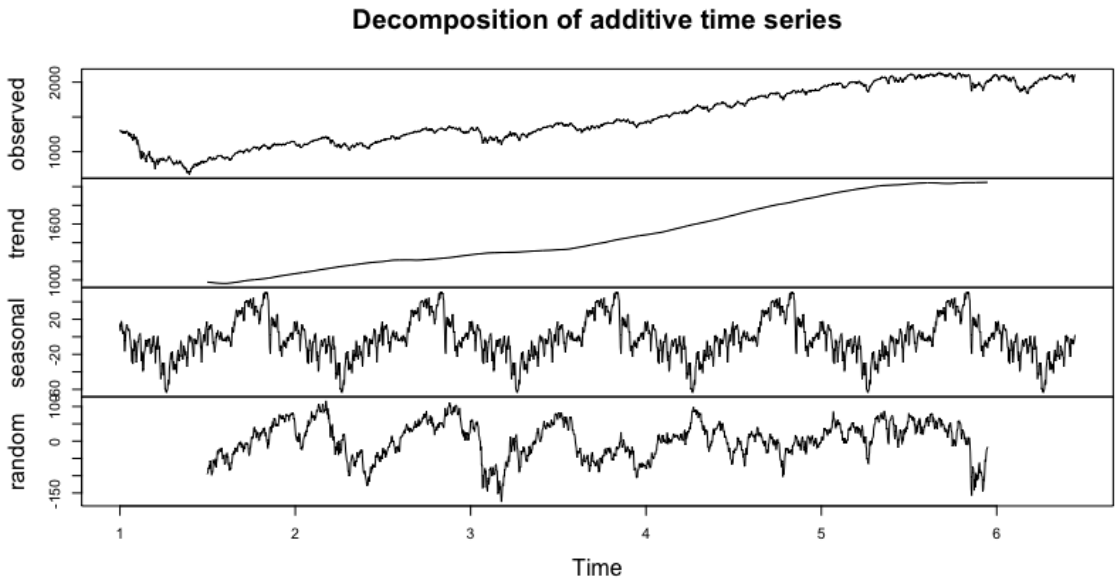




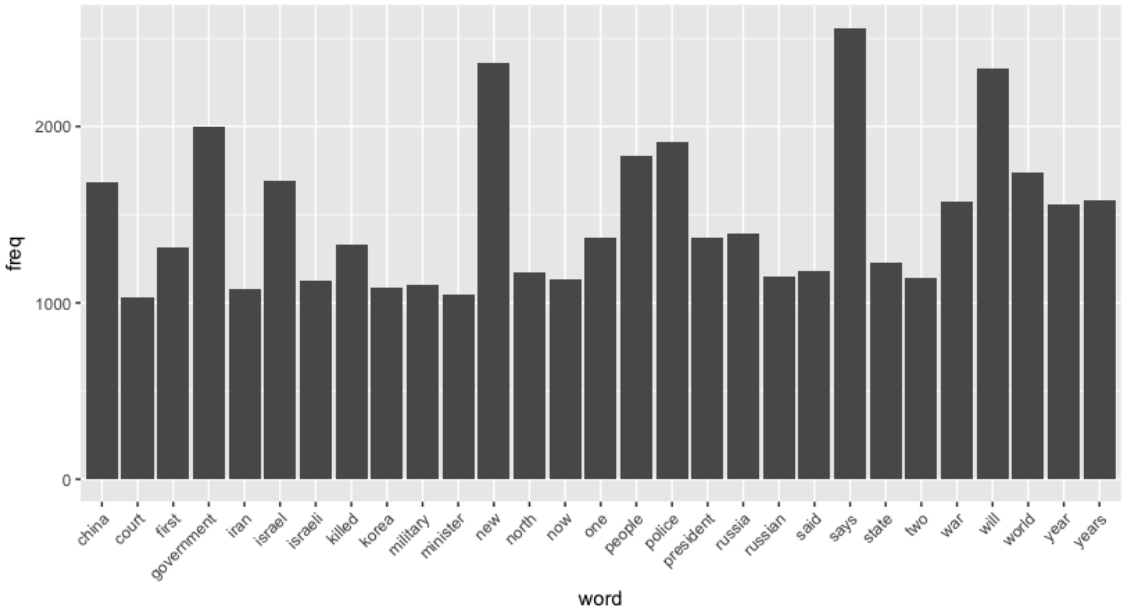


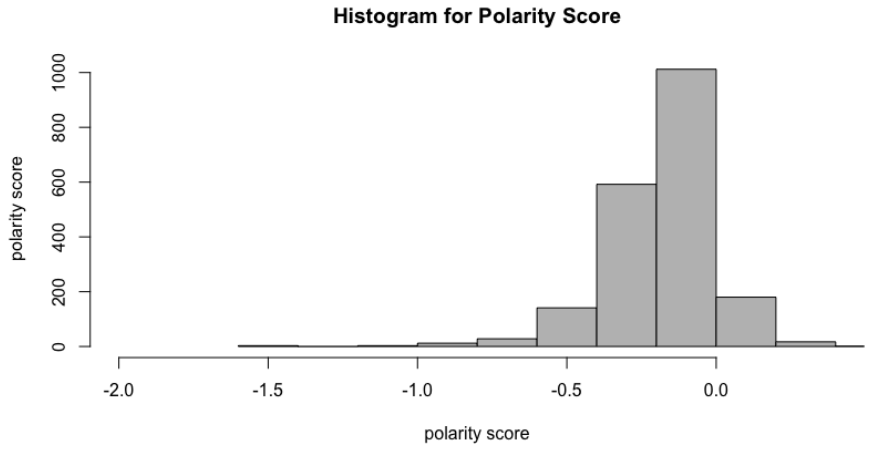
5. Data Analysis

The time series was decomposed…



5. Data Analysis



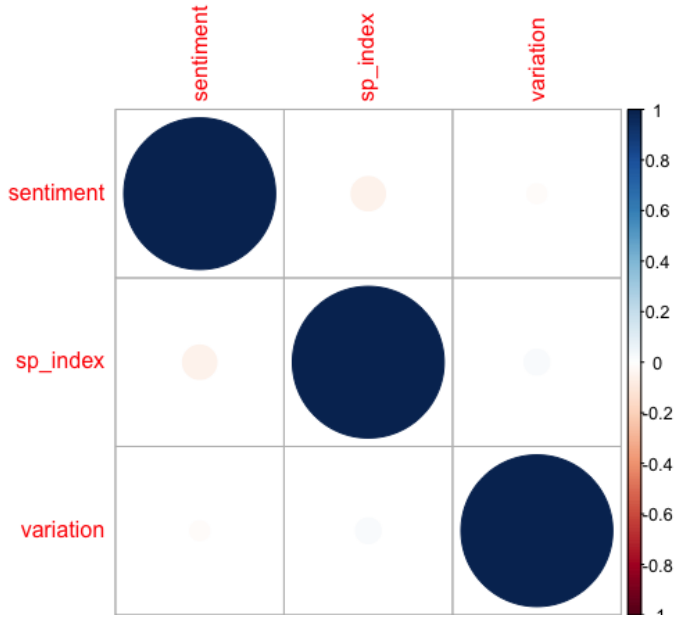


Also the polarity score was calculated for the Headlines

5. Data Analysis

Joining all the variables by Date

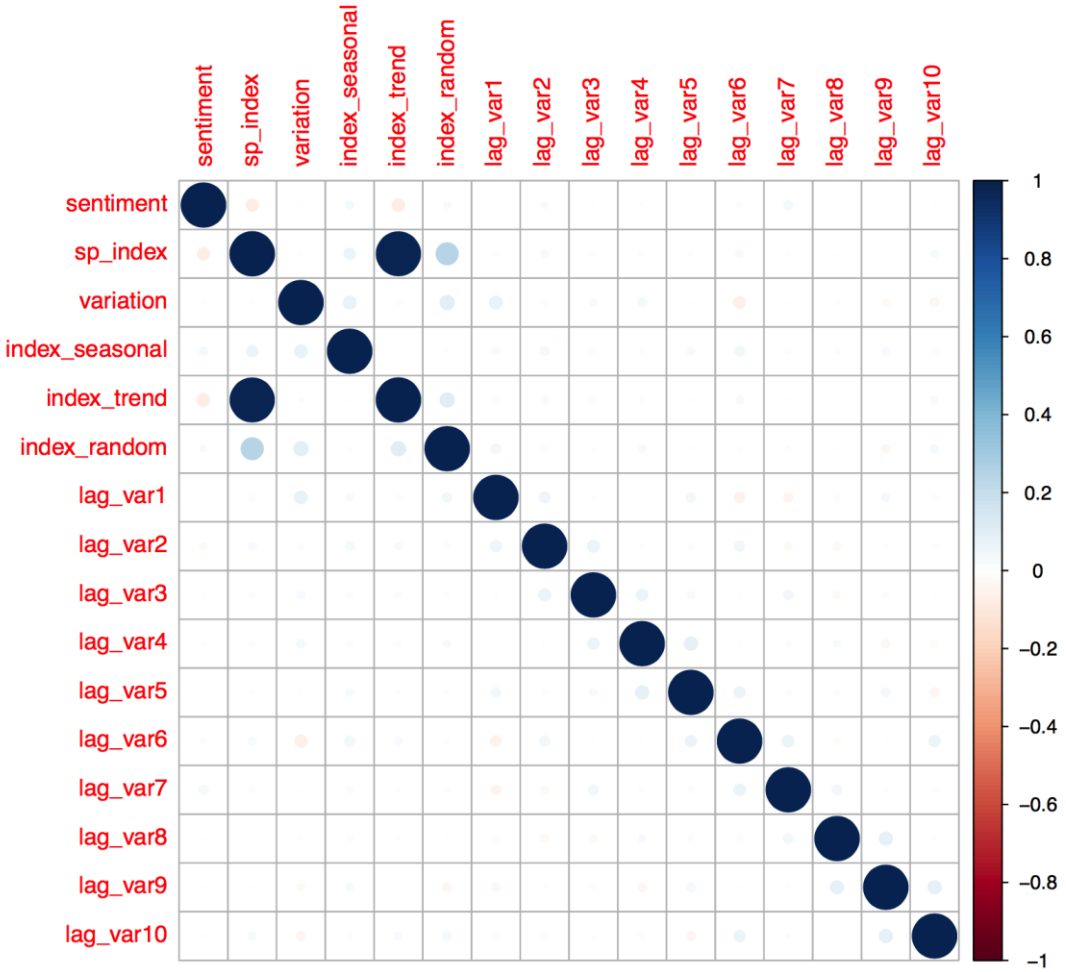
Name	Type	Description
Date	Numeric	Date
All_News	Character	String – All headlines
Sentiment	Numeric	Polarity Score
SP_Index	Numeric	S&P500 Index Adjusted
Variation	Numeric	Variation of S&P500 Index



The variables of the joint dataset were not correlated.

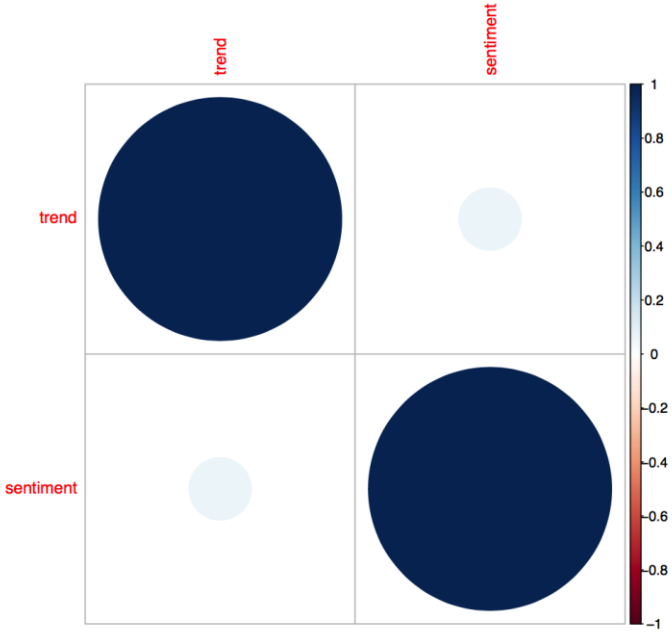
	sentiment	sp_index	variation
sentiment	1.00000000	-0.05022371	-0.01595043
sp_index	-0.05022371	1.00000000	0.02733238
variation	-0.01595043	0.02733238	1.00000000

5. Data Analysis



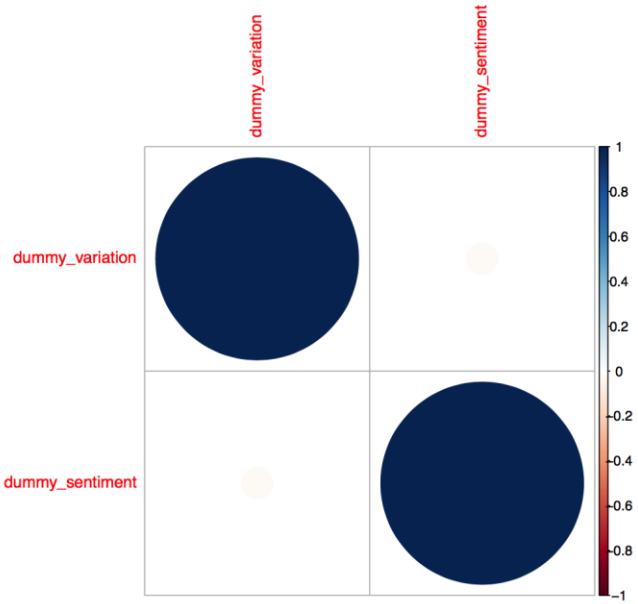
Lags for the variation of the stock index were computed. No lag was correlated with the sentiment.

5. Data Analysis



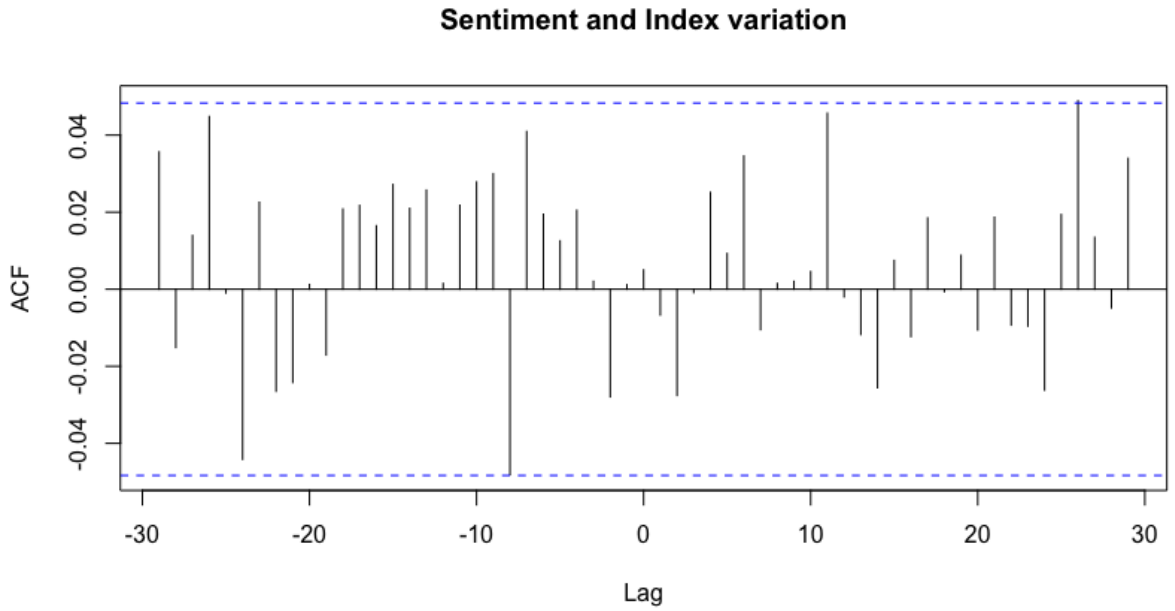
The differences of the trend component is not correlated with the sentiment...

Nor the dummy of variation of and a dummy of the sentiment…



5. Data Analysis

The variables show no joint structure among time...



Index

6. Conclusions

1. There were no strong evidence, neither for correlation not for any tipe of relation between the stocks and the headlines.

2. Seeking some relations between real world and events with help of sentiment analysis, is much harder than it seems.

3. Maybe other kind of news, with help of a proper Web Scrapper, may help to get more insights

4. This document and its conclusions to be a good set up point for further analyses

6. Conclusions

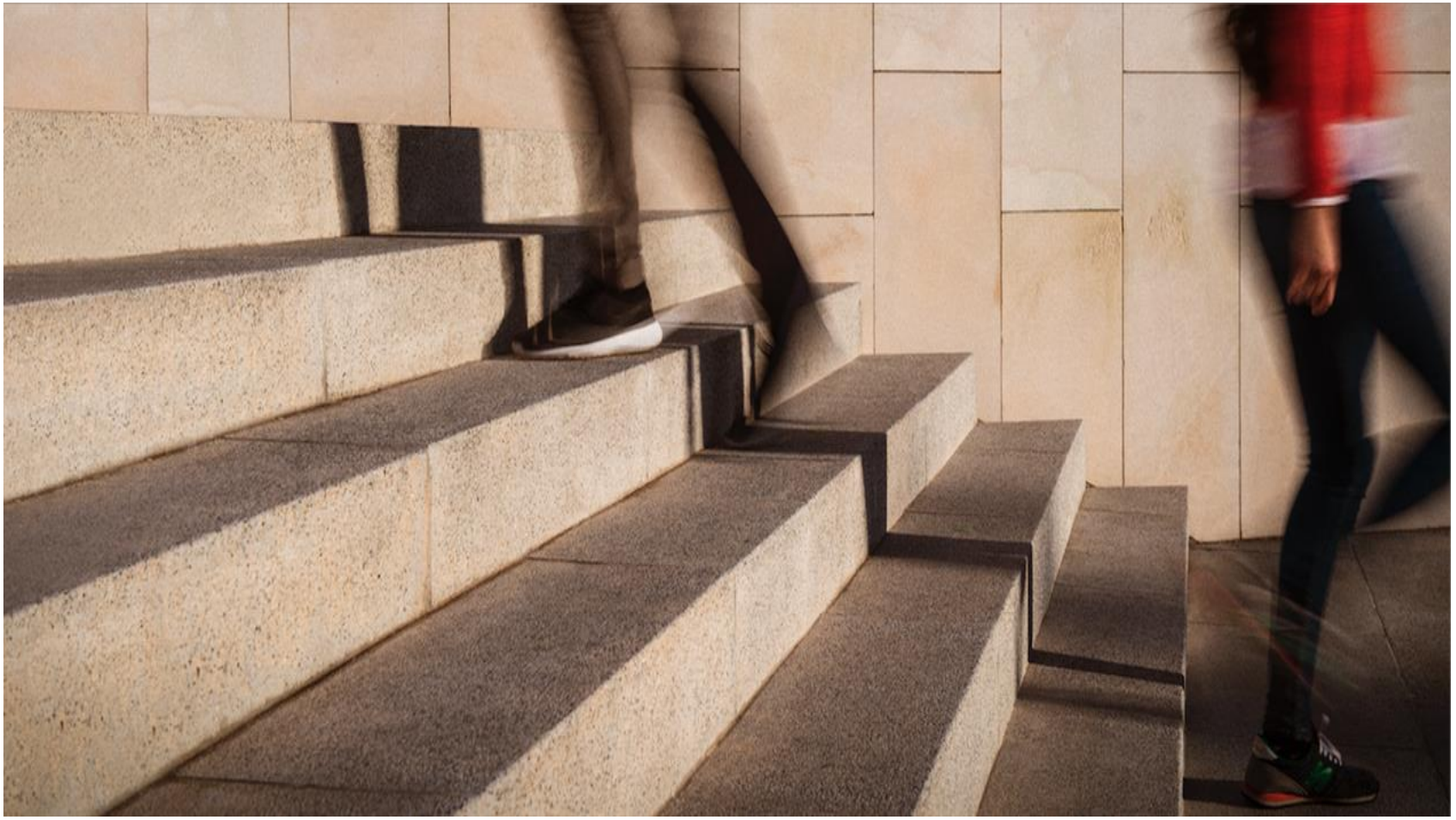
The code was made public in Kaggle under;

https://www.kaggle.com/alvaroanton/headlines-and-s-p500-index

Also, it was uploaded to Github in the following link

https://github.com/aablanco/stock_forecasting

Thank you for watching!



Universitat Oberta
de Catalunya

UOC.universitat
@UOCuniversitat
UOCuniversitat