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

1. Aim of the document
2. Introduction
3. State of the art
4. Methodology
5. Data Analysis
6. Conclusions
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

1. Aim of the document
2. Introduction
3. State of the art
4. Methodology
5. Data Analysis
6. Conclusions
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

1. Aim of the document
2. Introduction
3. State of the art
4. Methodology
5. Data Analysis
6. Conclusions
3. State of the Art

- **Web Scraping techniques:** Exploiting web scraping in a collaborative filtering based approach to web advertising [16], WebScraping 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

1. Aim of the document
2. Introduction
3. State of the art
4. Methodology
5. Data Analysis
6. Conclusions
4. Methodology

Data Sources

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

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

  - S&P500 Stock Index from Yahoo Finance
    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

1. Aim of the document
2. Introduction
3. State of the art
4. Methodology
5. Data Analysis
6. Conclusions
5. Data Analysis

### News Headlines Dataframe

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Numeric – Date Format</td>
<td>Date</td>
</tr>
<tr>
<td>Top1 – Top25</td>
<td>Character</td>
<td>Headlines – Top 25 of day</td>
</tr>
<tr>
<td>Label</td>
<td>Dummy</td>
<td>This variable is not relevant for the document</td>
</tr>
</tbody>
</table>

### S&P500 Dataframe

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>Numeric</td>
<td>Value at opening market</td>
</tr>
<tr>
<td>High</td>
<td>Numeric</td>
<td>Highest value at session</td>
</tr>
<tr>
<td>Low</td>
<td>Numeric</td>
<td>Lowest value at session</td>
</tr>
<tr>
<td>Close</td>
<td>Numeric</td>
<td>Value at closing market</td>
</tr>
<tr>
<td>Volume</td>
<td>Numeric</td>
<td>Volume of the S&amp;P market</td>
</tr>
<tr>
<td>Adjusted</td>
<td>Numeric</td>
<td>Adjusted index</td>
</tr>
<tr>
<td>Date</td>
<td>Numeric – Date</td>
<td>Date of the session</td>
</tr>
</tbody>
</table>
5. Data Analysis

S&P500 Index

Variation S&P500 Index

Series df_complete$sp_index

Series index_diff$variation

ACF

Lag

ACF

Lag
5. Data Analysis

The time series was decomposed…

Decomposition of additive time series
Also the polarity score was calculated for the Headlines.
5. Data Analysis

Joining all the variables by Date

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Numeric</td>
<td>Date</td>
</tr>
<tr>
<td>All_News</td>
<td>Character</td>
<td>String – All headlines</td>
</tr>
<tr>
<td>Sentiment</td>
<td>Numeric</td>
<td>Polarity Score</td>
</tr>
<tr>
<td>SP_Index</td>
<td>Numeric</td>
<td>S&amp;P500 Index Adjusted</td>
</tr>
<tr>
<td>Variation</td>
<td>Numeric</td>
<td>Variation of S&amp;P500 Index</td>
</tr>
</tbody>
</table>

The variables of the joint dataset were not correlated.
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...

Sentiment and Index variation
Index

1. Aim of the document
2. Introduction
3. State of the art
4. Methodology
5. Data Analysis
6. Conclusions
6. Conclusions

1. There were no strong evidence, neither for correlation not for any type 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/aablango/stock_forecasting
Thank you for watching!