UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS INFORMÁTICOS

MÁSTER UNIVERSITARIO EN INGENIERÍA INFORMÁTICA

Intelligent Systems

Tweets NLP

Author:

Jiménez Martín, Carlos - 160190

3 de enero de 2021

1. Introduction

Nowadays social networks are a great source of unstructured data, but thanks to text mining techniques and technologies, a lot of useful information can be obtained for a wide range of applications.

In particular, the prediction of the stock market is a field of great interest and where the analysis and data processing have a great role, since investors have profits or losses, and due to new technologies of data processing and analysis, we can get new strategies for investment in stock market to achieve higher profits.

Stock markets change every moment, and social networks, such as Twitter, upload news as soon as they occur, so in this work it is proposed to use Natural Language Processing techniques to obtain information from tweets and check if the trend of a stock index is affected, positively or negatively, with this information.

2. Problem to solve

To carry out the problem raised in the introduction, it is proposed to use NLP to obtain information from the tweets, since these are written in natural language, and to carry out a Sentimental Analysis of them to know if the news is positive or negative.

The dataset chosen for the analysis is the "Financial Tweets" dataset, which contains more than 16,000 tweets about companies listed on the American stock exchange "S&P 500" (and some crypto-currencies) and which are verified with the company that is tweeting [1].

The software used for the development of the work is the programming language R.

3. Experiment

The scheme of work that has been carried out to solve this problem is the following:

1. Initially, the dataset contains the following columns: "id", "text", "timestamp", "source", "symbols", "company_names", "url" and "verified". As we seek to extract information from the news for each day, we are only interested in the columns "text" and "timestamp", eliminating the rest.

id	text	timestamp	source	symbols	company_names	url	verified
101609	VIDEO: “I was in my office. I was minding my own business...” – David Solomon tells $GS interns how he learned he wa… https://t.co/QClAITywXV	Wed Jul 18 21:33:26 +0000 2018	GoldmanSachs	GS	The Goldman Sachs	https://twitter.com/i/web/status/1019696670777503745	True
101608	How satellites avoid attacks and space junk while circling the Earth https://t.co/aHzIV3Lqp5 #paid @Oracle https://t.co/kacpqZWiDJ	Wed Jul 18 23:00:01 +0000 2018	Forbes	ORCL	Oracle	http://on.forbes.com/6013DqDDU	True

Figure 1 Original Data

2. 2. Rows containing empty, null and "NA" data are deleted.
3. The date format contained in the "timestamp" column is of the type "Wed Jul 18 21:33:26 +0000 2018", so it has to be transformed to achieve the "dd-mm-yyyy" format. To do this:
    - The "timestamp" column is separated into the "WeekDay", "Month", "Day", "Hour", "UTC" and "Year" columns.

WeekDay	Month	Day	Hour	UTC	Year
Wed	Jul	18	21:33:26	+0000	2018

Figure 2 Split Timestamp

- "WeekDay", "Hour" and "UTC" columns are removed.
- Rows containing NA values are removed.
- The months are written in numerical format "mm.
- The date is joined with the format "dd-mm-yyyy".

Wed Jul 18 21:33:26 +0000 2018 → 18-07-2018

Figure 3 Date transformation

4. To be able to perform a Sentimental Analysis of the tweets' text we need to clean it:
    - Text is converted to lowercase.
    - Words containing "@" or any other non-alphabetic character are removed, and the urls.
    - The "stopwords" are deleted, using as a dictionary of "stopwords" the "Stopwords" library of "Snowball". [2]
    - The words are lemmatized using a "stemmer" function from the "corpus" library of "Snowball". [3]

Original	How satellites avoid attacks and space junk while circling the Earth https://t.co/aHzIV3Lqp5 #paid @Oracle https://t.co/kacpqZWiDJ
Transformed	satellit avoid attack space junk circl earth paid

Figure 4 Text transformation

5. Once the text has been prepared, the Sentimental Analysis is performed using the "SentimentalAnalysis" library, classifying the polarity of the tweets in: positive, neutral or negative. [4]
6. The polarity of the tweets is converted to numerical values:
    - Positive = **+1**
    - Neutral = **0**
    - Negative = **-1**
7. Finally, a matrix is generated where for each date, the values corresponding to the polarity of all the tweets of that date are added, so that it is represented if the twitter news of a day have been in their totality positive or negative.

Date	Total
08-07-2018	3
09-07-2018	-87
10-07-2018	-181
11-07-2018	-165
12-07-2018	-281
13-07-2018	-130
14-07-2018	-20
15-07-2018	-78
16-07-2018	-194
17-07-2018	-464
18-07-2018	-1284
19-07-2018	3
23-07-2018	-1

Figure 5 Result matrix

4. Results

As a result, we have obtained a matrix of two columns "Date" and "Total", where for each date we have the sum of the polarization of all the tweets for that date.

This information obtained from the tweets can be useful to carry out a prediction of the trend of a stock index, where if the total value is negative, it could be that the trend that the stock index has taken that day is negative, and in turn, the magnitude of the total value can determine the slope with which the trend change, so that if the value is very high (-4000 or 4000) the slope of growth or decrease in the evolution of the stock price of the stock index is very pronounced, and if the value is low (-4 or 4) the slope is almost constant.

To check if with only this information the behavior of a stock index can be represented, the evolution graph of the resulting matrix is obtained, and the evolution graph of the SP500 stock index.

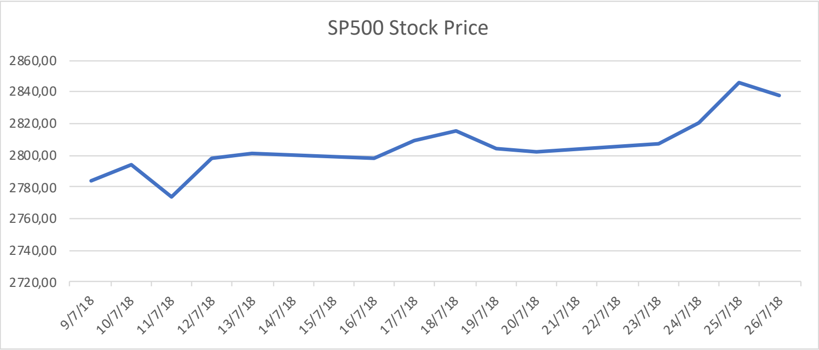

Tabla 6 SP500 Stock price

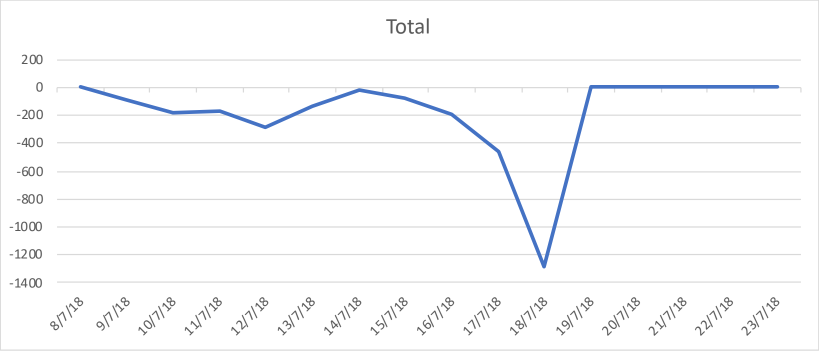

Tabla 7 Total value

As we can see the behavior of the result matrix is not similar to the behavior of the SP500 index, this is because with only this information we are not able to replicate the behavior of the stock market index, so the information obtained can be used as additional information to other types of variables that can be taken into account, such as financial variables.

URL GitHub repository: https://github.com/CarlosJM7/NLP-Tweets

References

[1] D. Wallach, «kaggle,» 2018. [En línea]. Available: https://www.kaggle.com/davidwallach/financial-tweets.

[2] «RDocumentation,» [En línea]. Available: https://www.rdocumentation.org/packages/tm/versions/0.7-8/topics/stopwords.

[3] «Stemming Words,» [En línea]. Available: https://cran.r-project.org/web/packages/corpus/vignettes/stemmer.html.

[4] «SentimentAnalysis Vignette,» [En línea]. Available: https://cran.r-project.org/web/packages/SentimentAnalysis/vignettes/SentimentAnalysis.html.