NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Analysis of S&P500 using News Headlines Applying Machine Learning Algorithms

Neftali Filipe Nunes Herculano

Supervisor: Mauro Castelli, PhD

Dissertation report presented as partial requirement for obtaining the master's degree in Information Management, Specialization in Business Intelligence

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

BOOK SPINE


2019
Title:
Subtitle:
Student
full name
MEGI


2019
Title:
Subtitle:
Student
full name
MGI

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Analysis of S&P500 using News Headlines Applying Machine Learning Algorithms

by

Neftali Herculano

Dissertation presented as a partial requirement for obtaining a master's degree in Information Management with a specialization in Business Intelligence.

Advisor: Mauro Castelli, PhD

November 2022

DEDICATION

As pessoas que nos são especiais nunca nos deixam. Se houve alguém que me fez ver o mundo de forma differente foi ele e levo-o comigo e as suas memoraveis frases para todo lado. Guardo-o como recordação para a minha vida e em cada decisão que tomo pondero qual seria a sua escolha. Um ser humano incrivel, uma pessoa fantástica e uma figura a seguir na minha vida. Não existem palaavras que possam descrever o misto de emoções que sinto, por uma lado a felicidade e grandeza e de fechar um capito da minha vida que sempre foi tão desajo por ele fosse concluido, por outro lado a tristesa não não poder partilhar este momento. Fica na memória o que as suas palavras e o seu sorriso.

Gostava de dedicar a conclusão desta fase da minha vida ao meu falecido pai em forma de agredecimento pelo exemplo que foi para mim!

Obrigado!

ACKNOWLEDGMENTS

I want to start by expressing my appreciation for my family and friends and being thankful for your unconditional love and support, helping me improve daily. To my mum, especially for being a role model, allowing me to grow every day, and making me a better person.

I would like to also share this special appreciation to my Supervisor, Professor Mauro Castelli, for his contribution and support during this process. I will always be thankful for the time he devoted to helping me, his contribution, and his particularly encouraging words.

Last but not least, I would like to express my deepest gratitude to my girlfriend, Carolina D´Ambrosio, for lifting me in the most challenging moments, sharing hard moments, and never letting me go down on moments where the work was intense.

You make me follow my dreams!

ABSTRACT

Financial risk is in everyone's life now, directly or indirectly impacting people´s daily life, empowering people on their decisions and the consequences of the same. This financial system comprises all the companies that produce and sell, making them an essential factor. This study addresses the impact people can have, by the news headlines written, on companies' stock prices.

S&P 500 is the index that will be studied in this research, compiling the biggest 500 companies in the USA and how the index can be affected by the News Articles written by humans from distinct and powerful Newspapers. Many people worldwide "play the game" of investing in stock prices, winning or losing much money. This study also tries to understand how strongly this news and the Index, previously mentioned, can be correlated. With the increased data available, it is necessary to have some computational power to help process all of this data. There it is when the machine learning methods can have a crucial involvement. For this is necessary to understand how these methods can be applied and influence the final decision of the human that always has the same question:

Can stock prices be predicted? For that is necessary to understand first the correlation between news articles, one of the elements able to impact the stock prices, and the stock prices themselves. This study will focus on the correlation between News and S&P 500.

KEYWORDS

Machine Learning, S&P 500, Sentiment Analysis, Correlation, VADER

List of Contents

List of Figures

List of Equations

List of Graphs

List of Tables

LIST OF ABBREVIATIONS AND ACRONYMS

S&P 500 – The Standard and Poor's 500

EMH – Efficient Market Hypothesis

AMH – Adaptive Market Hypothesis

RW – Random Walk

ANN – Artificial Neural Network

SVR – Support Vector Regression

DT – Decision Trees

LR – Linear Regression

RF – Random Forest

NASA – The National Aeronautics and Space Administration

LSTM – Long Short-Term Memory

MKL – Markel Corp

RCNN – Region Based Convolutional Neural Network

VADER – Valence Aware Dictionary Sentiment Reasoning

API – Application Programming Interface

NLTK - ´Natural Language ToolKit

1. INTRODUCTION

In the financial market, everyone invests money aiming to beat the financial system and make a good return. Financial time series forecasting has been a theme since the 1980s. Several economic, financial, psychological, and political factors impact the behavior of the financial time series. Therefore, predicting them remains one of the most challenging time series forecasting tasks. (Cao & Tay, 2003) "*Due to their inherent noise, non-stationary, and deterministic chaos, they cannot be considered stationary*."

The question that needs to be addressed is: Can stock prices be predicted and profited upon? In efficient markets (where information is disseminated rapidly), it is impossible to predict stock prices and profit from them (Campanella, Mustilli, & D'Angelo, 2016). According to Efficient Market Hypothesis (EMH), that is not possible. However, if this holds in the days we live, investment in the financial field would not exist. Several Finance researchers subscribed to the EMH, which states that the current stock price is merely the result of the rational investors factoring all available information into it (i.e., the market is efficient) and, based on this theory, a way to predict and profit from the future price of stocks does not exist (Manahov & Hudson, 2014) (Malkiel , The Efficient Market Hypothesis and Its Critics, 2003).Random Walk (RW) from Shonkwiler (2013) is the foundation of EMH's belief that stock market prices cannot be predicted, which implies that stock prices are based on a random walk, so investors cannot predict and profit from that stock price movements.

Despite the influence of EMH and RM models in the financial world, there has been a recent proposal for an alternative hypothesis known as the Adaptive Market Hypothesis (AMH) from Lo (2017). Considering each investor as a human being with rational and irrational behaviour, AMH concludes that investors can be adaptive when facing different environments, like the economy or politics (Lo., 2017). As suggested by EMH, this can create moments when the market is efficient and moments when the markets do not follow the Random Walk model and are predictable, and the investors can profit from it (Lo., 2017). To summarize, AMH theory holds that stock market prices can be predicted through some methods; opposing the previous theory, EMH, that assumes that stock prices are not possible to be predicted, identifies the supply and demand as sole driving factors of investor's instincts of buying or selling companies stocks (Thomsett, 2015). Therefore, the goal is not to doubt whether financial time series are predictable but to understand if there is a correlation between News and stock price variations and a model that can do it.

1.1. RESEARCH QUESTION AND HYPOTHESES

Since the beginning of this century, the news is now easily accessible and explored, such as online news services projected to the World Wide Web. The availability of faster machines and the developments in Natural Language Processing and other Machine Learning methods provide many benefits in several areas from the analysis of a large amount of information present in diverse environments.

Text mining techniques can be used to predict the variances in stock prices based on news content. This research has been conducted based on the influence of news on the stock market environment and the reaction of the stock market to the media outlets, as well as how information about a company's report can dramatically affect its share price. Although numeric time series forecasting seems to be the most promising method for predicting stock market movements, the number of text

mining applications using news analysis is somewhat limited due to the enormous complexity of text mining processes in unstructured data.

As a result, this research wishes to answer the following Research question:

(RQ): Can a correlation be defined between the S&P500 index and News Headlines applying Machine Learning Algorithms?

Hence, the research will deal with news articles' impact on Stock prices. There is a necessity for creating a framework that can identify inputs that contain relevant information for the forecasted stock, which can have value for the design and development of the stock prices and News Articles related to that of stock prices. The framework was implemented to run experiments for the S&P 500 stock price movement of the last three years, using machine learning algorithms (ANN, SVR, DT, LR, RF) and News Articles of precisely the same historical timeline transforming this input in a numerical number and understand if a correlation is there, positive or negative, with that movements.

The analysis which will be conducted is explanatory and will confirm if News articles had an impact on the S&P 500 Index over one year and a half and, based on that, if it is possible to forecast the variances of the stock prices. In order to prove this, a set of hypotheses will be tested, such as:

H1: A correlation between the S&P 500 index and the Sentiment Analysis of News Headlines Exists.

H2: A correlation between the S&P 500 index and the Sentiment Analysis of News Headlines Does not exist.

1.2. Study Relevance and Importance

Stock markets have been studied over the years to extract the maximum helpful pattern and models to predict their movements. This prediction of stock markets always has a certain appeal from the perspective of the researchers and investors that try to overcome the market, increasing by this means their profits. If the financial analyst can predict the future behaviour of stock prices, they can make decisions based on that knowledge and use it to maximize the gains.

Stock prices can be profited from more accurately by predicting stock models which are applied to the variance of the stock prices. Usually, the most common way to forecast the trend of the stock price is through technical and fundamental data analysis. However, numeric time series data is limited and only contains one event and not what was the cause of that event and why it happened. Textual data, such as news, have more information about the why. Once they are transformed from textual information into numeric time series data, it increases our quality of information. When the input is structured, the model has a higher quality, and the output also increases the quality, making the predictions more accurate.

The mass media, or even more specifically, news articles, are one of the most significant factors affecting human decisions and behaviour. However, it is also possible to observe movements in financial markets resulting from investors' perceptions of what is happening around the markets and how they perform specific actions based on that perspective. It is possible to affirm that, indirectly,

news articles can influence the stock markets since they highly impact human decision-making. Stock prices are also highly affected by the decisions of every investor.

It is becoming increasingly common to find valuable and vital real-time news articles on the internet related to companies and financial markets impacting stock prices. Identifying how to extract this valuable information and relate it to the financial market is one of the most critical aspects of any financial analysis, as it helps every financial analyst predict future stock market behaviour, leading to more and more profits. By providing their customers with more profitable trading rules, stockbrokers can increase the satisfaction of their customers.

1.3. STRUCTURE AND ORGANIZATION

This study was organized as follows: introduction, literature review, methodology, results, discussion and conclusion. Chapter 1, the current section, constitutes the introductory chapter, where the topic has been presented, the structure of the research, and the research questions. Chapter 2 will summarize the relevant literature on the topic and provide a critical overview of some theories. Chapter 3 will present the research design, which will approach topics such as the collection of the datasets, data transformations, presentation of the methods selected to analyze the sentiment of the headlines, and application of the correlations algorithms to achieve the final results. Chapter 4 will present the results and the analysis of the same by multivariate charts and tables, as well as the discussion of the findings, which will relate the results with the relevant theory and literature. The final chapter will be the conclusive one, which will provide the answer to the initial question and the main limitations of the main argument.

2. LITERATURE REVIEW

The decision to invest in stocks is one of the most exciting and trending but also very risky that almost any active individual can make. A trader´s decision-making depends on the information he has access to, which can lead to enormous profits or partial or complete losses on his account. Human error must be minimized in decision-making scenarios to maximize profits (Asur & Huberman, 2010). Nonetheless, human beings cannot access and consider all the vital information available when making decisions. As a result, machine learning and predictive modelling have become one of the most popular research areas in the financial market today.

According to *Malkiel's* random walk theory (1999), Stock prices are purely random and cannot be forecasted because they are defined as purely random events. However, with the advancements of machine learning and the increase in information available to investors, it is now possible to forecast stock prices and trends predictions than using a random approach (Bollen & Mao, 2011) (Bijari & Khashei, 2011) (Nassirtoussi, Aghabozorgi, & Wah, 2015).

The literature review will be split in three different sections: Section 1, is related to the Stock Prices prediction. Since the research analysis S&P 500 Index, it is relevant to have a deeper knowledge about the S&P 500, specifically, the theories and models are searched to in use to predict the S&P 500. The second section concerns the sentiment analysis subject. It will be necessary to have further information about the prototypes already in use to interpretate, in a large scale, the sentiment of the text since it will be the major functionality to interpret the correlation between the headlines and the Index. The third and last section of this literature review represents the combination of both topics mentioned in the previous sections. Once a greater knowledge about the S&P 500 index is reached, and more information regarding Sentiment Analysis models is provided, it will be possible to interpret the correlation between them. The third section represents the connection between the News Articles and the S&P 500 index.

2.1. S&P 500 PREDICTIONS

"S&P 500 Stock Price Index is a capitalization-weighted index comprising of the 500 largest and most actively traded domestic industrial stocks." - (Chicago Mercantile Exchange, 1988)

Using a Takagi-Sugeno (TS) technique, Sheta (2013) developed fuzzy models for two nonlinear processes. They were used to predict the following week of the S&P 500 index by a NASA software project. First, using the model input data was possible to determine the membership functions in the rule antecedents; Second, estimating the consequence parameter using the least squares method to estimate these parameters. Excellent results were achieved.

M.H. FazelZarandi et al. (2009) has also developed a system for stock price based on a type-2 fuzzy rule-based expert system. This model allowed us to understand that the interval of every element was the membership value. An automobile manufactory in Asia was used as an example to test the type-2 fuzzy model, which was based on technical and fundamental indexes as input variables. The model

successfully forecasted the variation of the stocks along different sectors after very complex tests. Those results were such a success that they were used in a real-time trade system to predict stock prices.

In order to predict the S&P 500 index and stock prices, several studies have been conducted. From 2012 to 2019, Jiang et al. (Jiang, Liu, Zhang, & Liu, 2020) estimated the direction of the stock prices after one month based on 24 technical variables and ten different models. As a result of reconstructing the Long Short-Term Memory (LSTM) series to provide multiple outputs, Yu and Yan (2020) were able to predict tomorrow´s closing prices using an LSTM model based on daily S&P 500 index closing prices coupes with wavelet smoothing.

LSTM models were used in Jiang et al. and Yu et al. research, but the ability of these models to fit the data was limited by their lack of observations, affecting the effectiveness of the model, having a maximum of 4000 observations, as compared to over more than 10,000 necessary parameters.

Ding et al. (2015) predicted daily S&P500 closings for ten months using text embedding and convolution. His research found that news information slowly began to fade over time, and the news more than weekly or monthly data predictions better aid daily data predictions. However, the news still influences weekly and monthly predictions.

Gorenc Novak and Veluš˘cek (2014) analysed three hundred and seventy companies from the S&P 500 between 2004 and 2013, using technical indicators to predict their daily high. Grid searches were conducted to determine the model that would fit better to achieve higher results, using five different options starting with five and doubling by that number. They also chose 500 days for training to compromise between a smaller and larger stationary set. Different pre-processing steps, period choices, and dependent variables make it difficult to compare different methods.

2.2. TEXT MINING IN OTHER AREAS

As Nagar and Hahsler's research (2012) describes, a news corpus can be constructed by aggregating news stories from various sources using an automated text-mining algorithm. Natural Language Processing (NLP) techniques analyse the Corpus and filter out the relevant sentences. Using positive and negative polarity word counts, NewsSentiment, a sentiment metric, measures the sentiment of the overall news corpus by counting the positive and negative polarity words. They claim that the NewsSentiment shows a robust correlation with actual stock price movement due to the use of various open-source tools and packages.

Using a text mining approach, *Yu et al.* (2007) developed a framework to quantify the sentiment of news articles illustrating how it impacts energy demand by being shown as a time series and compared with fluctuations in energy price and demand. J. Breen (2011) uses keyword tags to score tweets to access polarity and sentiment to determine the sentiment prevailing about airlines and their customer satisfaction ratings.

Furthermore, Sentiment analysis is also used in other studies by companies such as Google and Amazon, that make use of this analysis to understand the satisfaction and engagement of the employees. A survey was used to collect the negative sentiment in a few different categories like

work/life balance, career perspective, and future opportunities meaning that the sentiment models are critical to show public opinion (Alamsyah & Ginting, Analyzing Employee Voice Using Real-Time Feedback, 2018).

Similarly, in the tourism field, this kind of study can be used to, for example, checking the sentiment of the reviews of hotels. In his study, Xiang et al. (2017) confirmed a significant difference between platforms in three applications (Trip Advisor, Expedia, and Yelp), using a hotel in New York and Manhattan as a research topic. Reviews about landmarks and attractions, experiences, essential services, core products, and value are among the topics reviewed by consumers. They have different connotations from different platforms, such as basic service has a higher connotation in Yelp, contrasting with value, which has a lower one. Even in the movie field it is possible to apply sentiment analysis as Nair et al. (2015) did, by analysing Malayalam movies determining the reviews' polarity and understanding their sentiments, giving them a positive or negative value.

2.3. STOCK MARKET CORRELATION WITH NEWS ARTICLES

According to *Nofsinger* (2001) , sometimes investors buy after good news, causing demand to rise. From what we were already able to see from the study of S.G. Chowdhury, S. Routh, and S. Chakrabarti (2014), nowadays, stock prices are predicted based on sentiment analysis. It aimed to investigate the correlation between the sentiment from news and stock prices and test the efficient market hypothesis based on the public sentiment of fifteen companies for one month. Around 70% of the headlines predicting positive sentiment were accurate. The study also showed that the company´s profit or loss could influence the stock price volatility. However, the information from the news available also influences Stock Prices (Kirange & Deshmukh, 2016). Khedr et al. (2017) used data mining in a study to also predict market behavior by collecting news data. The results showed an accuracy of 86% on the Naïve Bayes algorithm.

A study on thirty companies was conducted by Alanyali et al. (2013), to understand the correlation between financial news and the stock market. The data were taken from Dow Jones Industrial Average, and Spearman Rank correlations were used to achieve the results. Those results showed that the most frequent the name of that company was called, the more positive the effect on its transaction volume. However, there is no evidence that the association between the company's name called and the company's transaction volume results in good results, an important subject missing from this study. In an efficient market, information is readily available at any given time, demonstrating market conformity (Chowdhury, Routh, & Chakrabarti, 2014).

Using news articles with varying levels of relevance to the target stock in conjunction with financial forecasting can improve results, according to T.M. McGinnity (2015). Multi-kernel learning techniques were used for portioning data extracted from different sectors, sub-sectors, and industries based on the five categories of news articles. In order to analyse each category of news articles, separate kernels are applied to each subcategory, industry, and sector of the targeted stock.

Compared with methods based on fewer news categories, the simultaneous use of five news categories improves prediction performance. Also, according to the results, using all five categories of news with two kernels of the polynomial and Gaussian types for each of the five categories of news, MKL achieved the highest prediction accuracy and return per trade.

Sentiment analysis uses public feeling to classify texts based on their expression, as seen in this study from Andry Alamsyah (2019), where stock prices were predicted using data mining. This research shows that headline news directly correlates with stock returns and monitors public sentiment about the companies via Twitter, using the same sentiment analysis. In his study, Gupta et al. (2020) used the same method to understand the sentiment of the Stock Twits to predict Stock Prices.

Mohan et al. (2019) improved the stock price prediction by analysing News articles with deep learning models by having a more significant amount of time-series data. At the same time, Pagolu et al. (2017) showed that public tweets sentiments and stock sentiment markets movements are highly correlated.

A study by Chen & Bahsoon (2013) applied machine learning techniques to various fields like historical price data, volume, and average variance to achieve a model that could predict stock prices more accurately as have been done by other studies before (Huang, Nakamori, & Wang, 2005) (Kouloumpis, Wilson, & Moore, 2011). The viability of the readings is improved by combining deep neural networks, applied to the previous variables, and sentiment analysis models achieving better results (Shynkevich, McGinnity, & Colema, 2015). Allen (1999) proposed evolutionary computation through genetic algorithms to find technical trading rules, and Kim (Kim, 2003) proposed statistical learning through support vector machines to achieve the same rules.Finally, Melo (2012) discussed the analysis of text in news data and the modelling of the data, as Kim suggested.

Some studies have also used deep learning-based models for deep belief networks, and analysed textual data, such as the one by Batres Estrada (2015), which consists of Restricted Boltzmann Machines stacked on top of a Multi-Layer Perceptron. Another well-developed neural network of this type is Long Short-Term Memory (LSTM). The vanishing gradient issue that recurrent networks suffer from has been addressed by Hochreiter (1997) with this algorithm. The majority of the studies involving textual news data have used LSTM in order to improve their results.

The results of a study reported by Ding (2014) show that daily forecasting outperforms weekly and monthly forecasts. Different approaches have been used to map the textual information to an embedding space. (Bengio, Ducharme, Vincent, & Jauvin, 2003) (Ding, Zhang, Liu, & Duan, 2015) Bengio's and Ding´s researches have also used event-embeddings and word-embeddings as representation techniques. These are superior to the previous use of textual information because lower-dimensional dense vectors could show developed characteristics of words. An RCNN model has been proposed (Vargas, Lima, & Evsukoff, 2017) for forecasting the daily directional moves of the S&P500. In order to reflect the actual effect of the news sentiment, they combined seven technical indicators extracted from the target series with financial news headlines. Zhang et al. (2019) describe a new study of different financial indices, such as the S&P 500 and NYSE, which is predicted based on the Generative Adversarial Networks model.

According to Cambria and White (2014), market participants are likely to consider any information revealed about the potential future path of monetary policy. In this study, the sentiment score of the information is adjusted on grammatical and syntactic hints. This is done by applying the VADER algorithm based on a word dictionary containing both positive and negative words. The sentiment analysis dictionary Shah (2018) used was developed based on the sentiment analysis model dictionary for the financial industry.

3. METHODOLOGY

3.1. RESEARCH DESIGN

The purpose of a descriptive research study is to provide initial insights and guide any more necessary research, indicating any specific objectives to achieve the main goal and data collection that might be required. An individual who wishes to gain a deeper understanding of a situation, or identify alternatives to a particular decision, will find this research most beneficial. Quantitative and qualitative research methods are two different approaches that can be undertaken (Schindler & Cooper, 2003). Research in qualitative areas commonly employs an unstructured design on small samples to provide some inputs, whereas research in the quantitative field usually uses statistical analysis to compute data.

Primary data is typically used for exploration, though sometimes, enough existence of data allows us to conduct data mining and apply machine learning methods precisely to investigate relationships between every single dimension. Researchers collect primary data for specific purposes of a specific question, and secondary data were already collected for other objectives. Malhotra and Birks (2007) classified this secondary data in two different ways: Internal, when the data comes from specific fields, like finance, operations, and marketing, and there is internal access, and External, typically data that was published or gathered from the internet. To conclude, this research uses two types of data: the S&P 500 index, which is assembled data collected from the stock markets exchange, and financial news articles, focusing on the headlines of every article, collected using an API from two different external sources. The following hierarchy graphs depict the approach described above.

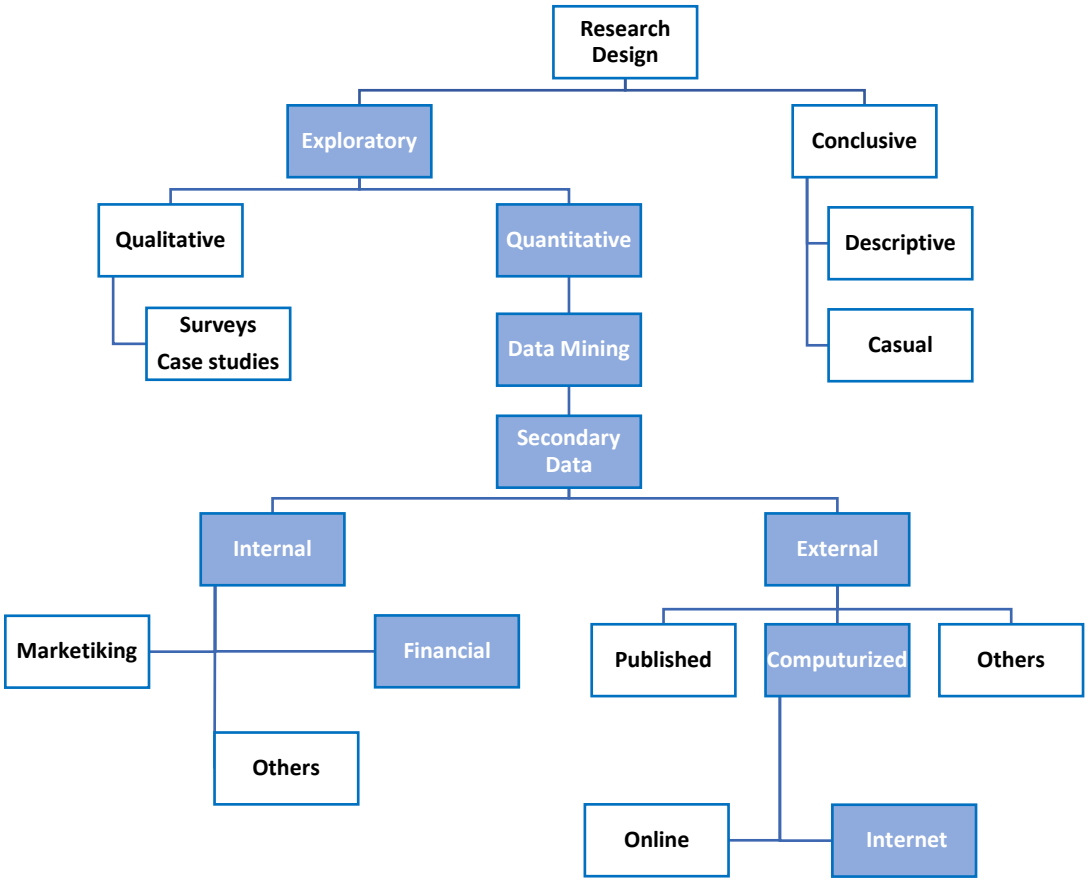


Figure 1 - Research Design Schema

3.2. RESEARCH PROCESS

A review and evaluation of different studies of the prediction of stock price movements using financial news articles were conducted based on the global process outlined by Fung (Fung, Lu, & Yu, 2005). As part of this process, data are collected, followed by a pre-processed procedure of that data text and numerical data, features are selected, machine learning algorithms are used, correlations are done, and evaluation is performed as is described in the following picture:

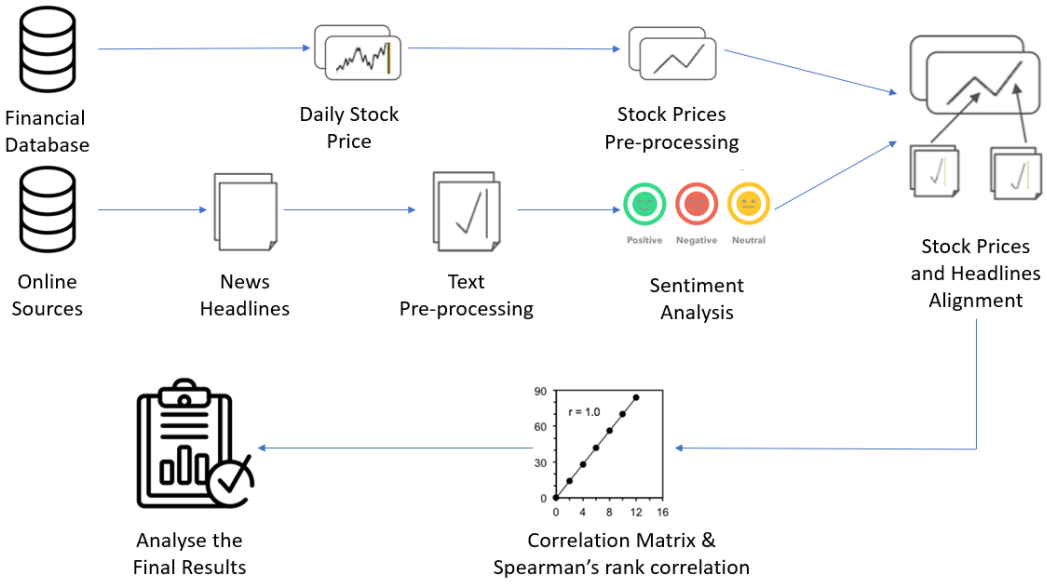


Figure 2 - Methodology Design Schema

A programming language such as Python or R language is required to accomplish the overall research process. A diverse number of processes are also necessary to achieve the right results as linear and non-linear models, analysis of time series, classification, application of correlation methods, and finishing with a visualization report that allows an evaluation and analysis of the results. "R is an environment within which statistical techniques are implemented and can be extended via packages" (Gentleman & Ihaka, 2003), and "Python is a dynamic object-oriented programming language that can be used for many kinds of software development" (Rossum, 1990).

3.2.1. Research Process - First Phase

To summarize the research processes described, the following Figure has been provided:

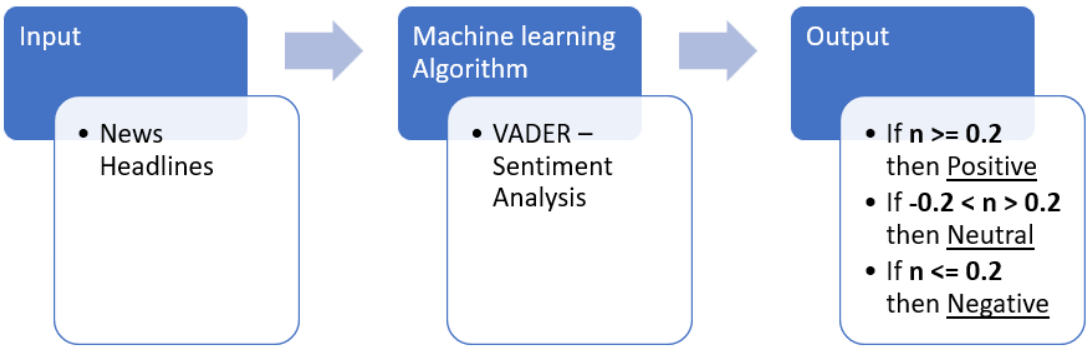


Figure 3 – Research Process, First Phase

In the first phase, news Headlines are used as input and are inducted in the phase of predicting and analysing those headlines' sentiment using the VADER algorithm. This analysis aims to understand whether the news headlines are positive, negative, or neutral based on a conjecture of the words included in every headline; hence, once those headlines are included on the VADER, they feed the algorithm as an input replicating an output of the system that could be one of those three options: above or equal to 0.2 meaning that would be a positive evaluation of the sentiment, between -0.2 and 0.2 representing a neutral sentiment or less or equal to -0.2 expressing that the headline as a negative connotation.

3.2.2. Research Process - Second Phase

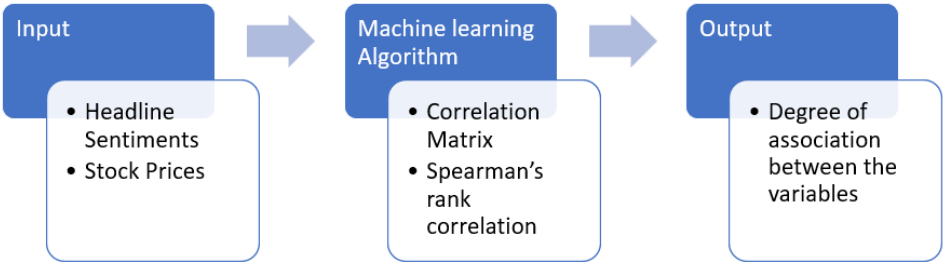


Figure 4 - Research Processes, Second Phase

As a second phase of the research process, the sentiment of the headlines and the stock prices will be used as input. Both will suffer tests of correlation using the correlation matrix algorithm and spearman's rank correlation to achieve the final goal of the research as an output. The correlation between those two variables will be defined by the degree of association between the two variables used. With this information, investors can make a conscious decision based on a correlation between the two variables.

3.3. DATA COLLECTION

To conduct this research is necessary to have access to two different types of data: first, the historical data from the intraday stock price of the S&P 500 as numerical data, and second, the news headlines from different sources - media outlets - related to the index itself or to the companies that are composing the index.

As the name indicates, Daily Stock Price is the daily data corresponding to the daily stock price. In order to correlate this stock price to the news headlines, a sample of news items was chosen between a specific range of times. After investigating the most reliable data source, the dataset was chosen between December 1st of 2019 and April 15^{th,} 2021.

For this step of data collection, Python is the tool that presented better resources to achieve the goal planned. Using different Application Programming Interface (API) to collect both necessary datasets,

it is possible to create a database big enough to achieve the results that are expected. To gather data about the daily stock prices of the S&P 500 between the dates previously mentioned, the API that presents solid and trustful results is the Yahoo Finance API. This API allows everyone to collect data from any stock price worldwide. There exist different ways of collecting the data for the second dataset regarding the News headlines. The one chosen for this research was the API from New York Times, which gives the News headlines from the New York Times newspaper, and the City Falcon News API, which collects news from different sources about different companies.

The Stock Prices gathered from Yahoo Finance API are considered internal secondary data from a financial institution, and the News Headlines are the secondary data from an online source.

3.4. DATA PRE-PROCESSING

Data, when gathered, usually presents a lot of "noise," meaning that it is not well structured or ready to be worked; it is possible to have a lot of different sources, different ways of structure, and even different expressions for the same result (Data Preprocessing in Data Mining, 2014).

Data preprocessing can be split into three steps: Data Cleaning, Data Transformation, and Data Reduction. In the Data Cleaning phase, we check for missing data, errors, inconsistencies, and outliers. Data transformation is correlated to the fact that most models are "picky." Hence, the data must fit specific properties such as normal distribution, equal values range, etc. Typically, it is a mistake to think that all the data is relevant and that as more data is available, the results will be better; however, not all data is equally important; some features bring the same information, what is called multicollinearity and normally more features, that might have been created during the analysis can lead to a loss of interpretability so is necessary to understand what is the vital information that will bring value to the final output.

For this reason, it is mandatory to analyze the data to understand its essential concepts. This process of data pre-processing is needed for both datasets, the Stock Prices one and the News Headlines one.

3.4.1. Stock Price Data Pre-Processing

For the pre-processing of the Stock Prices dataset it is necessary to:

- Check for missing values by counting the null values of the dataset and having the information about it in every column;
- Delete the Columns not necessary from the dataset;
- Use the Date as an index;
- Apply the percentage change between the current and prior elements;
- Apply the cumulative return sum of the column elements to understand the returns.

3.4.2. News Headlines Data Pre-Processing

For the pre-processing of the News headlines dataset it is necessary to:

- Check for missing values by counting the null values of the dataset and having the information about it in every column;

- Delete the Columns not necessary from the dataset;
- Use the Date as an index;
- Text pre-processing:
    - Remove all the punctuation using the “string punctuation” function
    - Lower all the letters in capital of every sentence “lower” function
    - Tokenization of every sentence, meaning that the lines will be split into smaller lines splitting the sentences using the “re.split” function
    - Remove the personal words such as “I,” “me,” and “my,” avoiding having personal sentiment in the sentences using the “stopwords” function from the NLTK library
    - Normalize the text by reducing the inflection in words and removing the words’ affixes using the “Stemming” function from the NLTK library (E.g., *Friends* becomes *friend*)
    - Normalize the text by reducing the inflected words and transforming the words into a canonical and dictionary form using the “Lemmatization” function from the NLTK library (E.g., *Running* becomes *run*).

3.5. SENTIMENT ANALYSIS

“Sentiment analysis, also called opinion mining, is the field of study that analysis people’s opinions, sentiments, appraisals, attitudes and emotions toward entities and their attributes expressed in written text” ***(Sentiment Analysis – Mining Opinions, Sentiments, and Emotions, 2020).***

When finishing the pre-processing of the News Headlines, it is time to give them meaning by giving an attribute to every single headline, a representation of the sentiment that the headline possibly gives to the reader. This headline evaluation should be given as Positive, Negative, or Neutral. As the names indicate, a positive evaluation indicates that the sentiment taken from the headline is positive, a negative represents a negative impact on the readers, and a neutral sentiment shows that the headlines did not affect most of the readers.

Nowadays, with the amount of information available, it would be impossible to do it without recurring the technology and the methods currently available. During the research for the best fit between the Machine Learning Algorithm to apply the sentiment analysis and the headlines, it was necessary to understand what kind of data the algorithms produce better results with. The algorithm chosen was the Valence Aware Dictionary for Sentiment Reasoning, known as VADER. This algorithm is known to have better results in classifying financial news sentiment and retrieving the best output possible, among all the others.

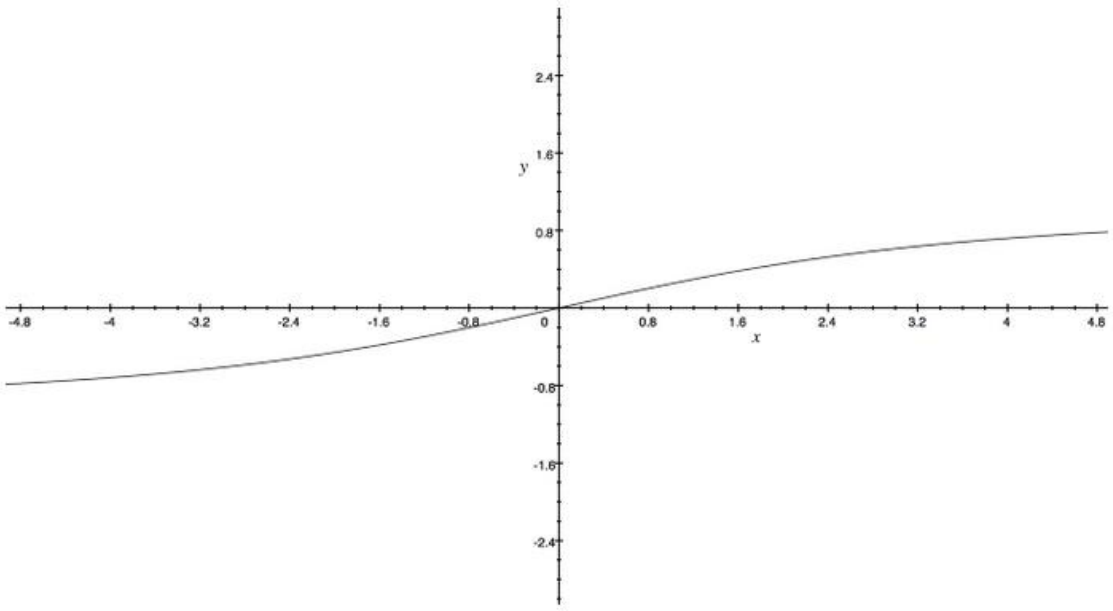
VADER is a model that analyzes the text from two different approaches: First, the sensitivity of the text polarity, Positive and Negative sentiments; Second, the strength of the emotion applied in every text. This model is based on a dictionary that contains lexical elements and a sentiment score mapping, a conjugation of those two approaches simultaneously. For example, “Happy” will be classified as a positive word and “Unhappy” as a negative word. However, the algorithm is intelligent enough to comprehend that when “Not happy” is written, it is a negative text, classifying it properly , in this case, as a negative. VADER also understands when the meaning is more substantial, for example, when capital letters express some opinion much more vigorously.

The sentiment score from VADER ranges between -4 and 4, being -4 the most negative and four the best possible for every word in the VADER dictionary. However, the final sentiment score is given back in a range between -1 and 1, representing the sum of the sentiment of those words in the same sentence. This happens because VADER also does a normalization in the middle of the process, represented by:

$$\frac{x}{\sqrt{x^2 + a}}$$

Equation 1 - VADER Normalization

The sum of the sentiment score of every word is represented by "x," and the "a" is representative of the normalization parameter, that by default is 15. The following graph represents the normalization:



Graph 1 - VADER Normalization

There are four different characteristics that VADER evaluates:

1. Punctuation – This takes into consideration the type of punctuation that gives a more considerable empathy to the sentiment that is explicit in the text, adding or subtracting, depending on the sentiment score of the word, 0.292 and 0.18 points per exclamation point and interrogative mark respectively;
2. Capitalization – Assume that capitals in words have a higher sentiment represented, increasing, or decreasing to the sentiment analysis 0.733 points, depending on the word sentiment;

3. Degree modifiers – VADER has on the dictionary a booster that contains words that can boost or damp the words used to give a higher or smaller sentiment. One modifier adds or takes 0.293 of the sentiment; for every modifier, it counts 5% less than the 0.293. E.g., A second modifier would add or take 95% of 0.293; a third modifier would add or take 90% of 0.293, and so on for each word;
4. Shift in polarity due to "but" – When a but is used to connect two parts of the sentence, it typically means that a constraint of sentiment that generally represents a more dominant sentiment, and VADER on these situations removes or adds 50% to the words before the "but" and gives or takes 150% more importance to the words after the "but."

The following table is representative of every one of the four characteristics:

Characteristic	Text 1	Text 2
Punctuation	I love it.	I love it!!!!
Capitalization	Great help.	GREAT help.
Degree Modifier	Very good.	Sort of cute.
Shift Polarity due to "but"	Was a great Idea, but was not very helpful	

Table 1 - VADER Characteristics

The following table demonstrates how the VADER algorithm's evaluation changed along with the sentence we are evaluating.

Text	Negative	Neutral	Positive	Compound
"This is a good song."	0.0	0.508	0.492	0.4404
"This is the best song ever made, is AMAZING!!!"	0	0.425	0.575	0.8877

Table 2 - VADER Evaluation

For all those reasons, VADER is an excellent algorithm to be applied to the dataset on the headlines, since it allows to take not only the sensitivity, but also the strength of the words, apply the characteristics that define VADER, and give a trustful output.

3.6. Stock Price and Headlines alignment

Once the sentiment of every sentence is given, it is necessary to match the dates of the headlines with the Stock price returns. It is necessary to work on the headlines.

There is the possibility of having more than one headline per day because it is also necessary to consider the weekends when the Stock Exchange is closed, so there will not be returns at that time.

The best way to consider the news headlines per day is by grouping all the headlines by date and applying a mean feature. This means that a mean of the sentiment analysis of all the headlines of the same day will be calculated, grouping them by the day they were released. Regarding the headlines released during the weekends, Saturdays and Sundays will be grouped with the following Monday after that weekend, applying the same mean but for three days instead of one day.

Once that is at the same level, both Stock Prices returns and Headlines Sentiment Analysis aligned on the same date, we are ready to apply the next steps.

3.7. CORRELATION MATRIX AND SPEARMAN´S RANK CORRELATION

Once the sentiment analysis of the headlines and the return of the stock prices are aligned, is time to move to the next step: the correlation between those two variables. In order to have the coefficient of that correlation, two different approaches will be taken. The Correlation Matrix and the Spearman's Rank Correlation.

3.7.1. Correlation matrix

It will use a vast dataset, even if with only two different variables, with a few hundred rows. The correlation matrix is an algorithm that allows understanding the correlation between those two variables in a speedy and summarised way.

A correlation matrix is as simple as a table that displays the coefficient between different variables and gives an idea about the presented dataset. As an example, in the case that one intents to predict a plane ticket based on the fuel, size of the plane, distance of the travel etc, the correlation matrix allows to understand how correlated is all of that information with the prices of the tickets helping to predict the price of the plane ticket. This correlation is given the following formula:

$$R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

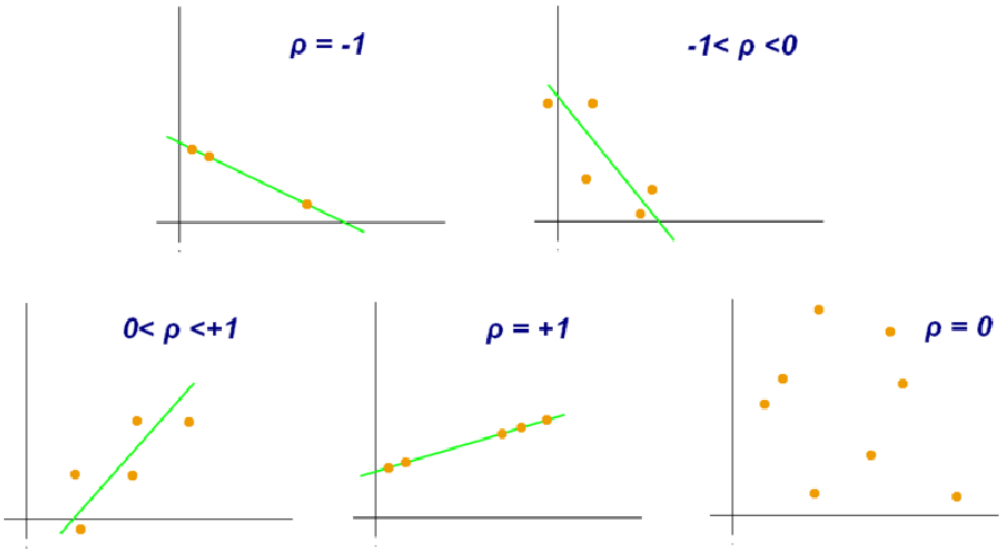
Equation 2 - Correlation Matrix

Where:

- x_i represents the value of the x dataset;
- x represents the mean values of the x dataset;
- y_i represents the value of the y dataset;
- y represents the mean values of the y dataset.

The results will vary between -1 and 1, and as close as they are from the 1, a positive correlation as the opposite also happens to the -1 representing a negative correlation. Close to 0 shows a neutral relationship.

The graphs of this correlation can be analysed as the following figure shows.



Graph 2 – General Correlation

For this research, the correlation matrix will help to understand if and how correlated are the headlines' sentiments with the variances of the S&P 500.

However, Spearman´s Rank Correlation will also be computed to complement the correlation study.

3.7.2. Spearman´s rank correlation

Spearman's Rank Correlation is an alternative to Pearson´s Correlation and it is used for curvilinear, monotonic, and ordinal data. To talk about Spearman's Rank Correlation it is necessary to understand the categories that the data might require.

3.7.2.1. Ordinal Data

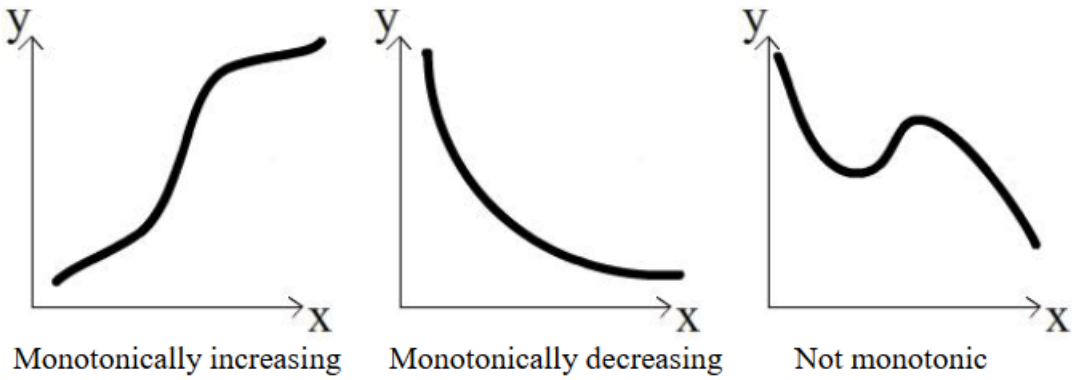
The first step will be to order/rank all the rows from the dataset from both variables in descending order. This means that the highest headline sentiment will have number one on the ranking as the highest value of the S&P 500 will also have ranking number one on the list, as the following table can demonstrate with another example.

Students	Math	Rank	Science	Rank
A	35	3	24	5
B	20	5	35	4
C	49	1	39	3
D	44	2	48	1
E	30	4	45	2

Table 3 - Ordinal Data Example

3.7.2.2. Monotonic Function

One of the characteristics of Spearman´s Correlation is the Monotony of the data, which means that the data is only increasing or only decreasing. If this condition is present in the data, then a monotonic pattern exists, otherwise it does not exist, as can be understood in the following charts.



Graph 3 - Monotonic Functions

On the Monotonically Increasing chart it is possible to understand that, while the x is increasing, the y does not decrease, while the complete opposite happens on the monotonically decreasing chart. In the last chart, non-Monotonic, the x variable increases, and the y has positive and negative variances, making the data not monotonic.

3.7.2.3. Spearman´s rank correlation formula

Once it is confirmed that this Monotonic relationship exists, the Spearman´s rank correlation will measure the direction of the two ranked variables that are given in the following formula:

$$p = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Equation 3 - Spearman´s Rank Correlation

Where:

- d_i is the difference between the two ranks of each observation;
- n is the number of observations.

The results' interpretation is very similar to the correlation matrix, since closer to 1 it corresponds to a positive correlation, and closer to -1 it represents a negative correlation. Close to 0 it shows that a correlation does not exist.

3.8. ANALYSE THE FINAL RESULTS

The last topic of the methodology can be approached once every step previously mentioned is done, and it will be understanding the results. Those are directly correlated to the Correlation Matrix and Spearman's Rank Correlation.

Once those two methods are finalized, it is possible to read them and conclude, based on the results explained in the previous point, if there is or is not a correlation between the general headline sentiments and the S&P 500 index.

4. RESULTS AND DISCUSSION

4.1. DAILY STOCK PRICES

Before diving into the correlation analysis, creating the stock prices dataset and doing a descriptive study of the same was necessary. Using Python to generate the dataset, an API from Yahoo Finance has been used, that provided all the daily stock prices from the S&P 500 between December, 1st of 2019 and April, 15th of 2021, by using the symbol "GSPC" that corresponds to the representative of the S&P 500 index in the Yahoo Finance database. Table 4 below displays the entire dataset for the daily stock prices. The dataset counts seven different variables: "Date", "Open", "High", "Low", "Close", "Adj Close", and "Volume", and contains 344 observations. These observations represent the days between the dates previously chosen in which the market exchange was open, not counting weekends and holidays. The description of the six variables is the following:

- Date – The date of the observation;
- Open – The value of the index when the market exchange opened;
- High – The highest value that the index reached on that day;
- Low – The lowest value that the index reached on that day;
- Close – The value of the index when the market exchange closed;
- Adj Close – The value of the index when the market exchange closed, considering other factors, for example, the dividends, when distributed, must be subtracted from the share price;
- Volume – The number of shares of the index traded on that day.

Date	Open	High	Low	Close	Adj Close	Volume
2019-12-02	3143.850098	3144.310059	3110.780029	3113.870117	3113.870117	3285750000
2019-12-03	3087.409912	3094.969971	3070.330078	3093.199951	3093.199951	3671580000
2019-12-04	3103.500000	3119.379883	3102.530029	3112.760010	3112.760010	3702980000
2019-12-05	3119.209961	3119.449951	3103.760010	3117.429932	3117.429932	3360480000
2019-12-06	3134.620117	3150.600098	3134.620117	3145.909912	3145.909912	3483310000
...	...	...	...	...	...	...
2021-04-08	4089.949951	4098.189941	4082.540039	4097.169922	4097.169922	3907100000
2021-04-09	4096.109863	4129.479980	4095.510010	4128.799805	4128.799805	3640390000
2021-04-12	4124.709961	4131.759766	4114.819824	4127.990234	4127.990234	3588900000
2021-04-13	4130.100098	4148.000000	4124.430176	4141.589844	4141.589844	3734720000
2021-04-14	4141.580078	4151.689941	4120.870117	4124.660156	4124.660156	3985350000

344 rows × 6 columns

Table 4 - S&P 500 Index: Daily Stock Price

4.1.1. Stock prices pre-processing

The Yahoo Finance API's data needed to be cleaned, transformed and prepared to be used to fit the models that are supposed to be applied.

This preparation was done in Python. When checking if there was any column with less observation or any missing value, it was possible to understand that every observation was completed and there was no value missing on the dataset, as shown in Table 5. The table also showed that the date type required, float64, was already implemented.

```
Data columns (total 6 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Open       344 non-null    float64
 1   High       344 non-null    float64
 2   Low        344 non-null    float64
 3   Close      344 non-null    float64
 4   Adj Close  344 non-null    float64
 5   Volume     344 non-null    int64
dtypes: float64(5), int64(1)
```

Table 5 - S&P 500 Index: Information

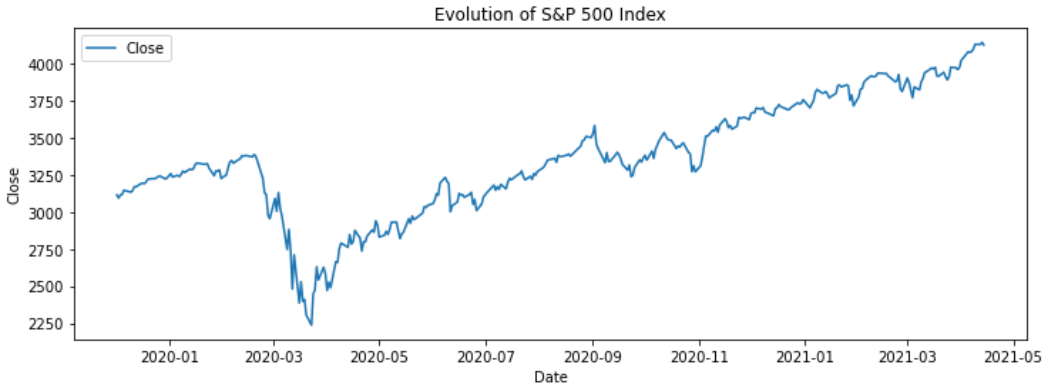
The dataset had more information than needed, so a data selection process followed. On this data, a selection was necessary to understand what would be the observations that were required to achieve the correlation. In order to achieve the final returns of the stock prices, it was needed to calculate those returns based on the price of the share when the market exchange closed. Also, to correlate with the headlines of different days, one column in common was necessary; the best fit for that requirement was the date. To conclude, the information required to keep with the research was the date and the close value of the index deleting all the other columns of the dataset, as observed in the following table.

Date	Close
2019-12-02	3113.870117
2019-12-03	3093.199951
2019-12-04	3112.760010
2019-12-05	3117.429932
2019-12-06	3145.909912
...	...
2021-04-08	4097.169922
2021-04-09	4128.799805
2021-04-12	4127.990234
2021-04-13	4141.589844
2021-04-14	4124.660156

344 rows × 1 columns

Table 6 - S&P 500 Index: Close Values

Following the analysis and pre-processing of that S&P 500 Index Dataset, a line chart was created to understand the evolution of the Close Value between December 2019 and April 2021, as the Graph 4 can show. It was possible to realise that the S&P 500 generally grew during this time. However, it significantly decreased between February and April of 2020, going from almost 3500 dollars to the lowest value of the index between the dates used for this research, reaching around 2300 dollars. After April 2020, the index increased the value constantly, month after month, reaching a peak of more than 4100 dollars.



Graph 4 - S&P 500 Evolution

The closing values of the S&P 500 Index were used to calculate the returns. Those returns were given by the percentage difference of the closing value of one observation minus the closing value of the previous observation. It was allocated to a new dataset called sp500returns, as shown in Table 7. There was a doubt if the difference in percentage day by day would make sense and if it would be the best

fit for the model, or if it would make more sense to have their percentage difference in a cumulative way. Doing that was taken into consideration daily differences but in incremental order. For that, a second Dataset, called sp500cumreturns, was created to calculate the cumulative returns of the Close Value of the S&P 500 Index as it is possible to see in Table 8.

Date	Close
2019-12-02	NaN
2019-12-03	-0.006638
2019-12-04	0.006324
2019-12-05	0.001500
2019-12-06	0.009136
...	...
2021-04-08	0.004221
2021-04-09	0.007720
2021-04-12	-0.000196
2021-04-13	0.003294
2021-04-14	-0.004088

344 rows × 1 columns

Table 7 - S&P 500 Index: sp500returns dataset

Date	Close
2019-12-02	NaN
2019-12-03	-0.006638
2019-12-04	-0.000357
2019-12-05	0.001143
2019-12-06	0.010289
...	...
2021-04-08	0.315781
2021-04-09	0.325938
2021-04-12	0.325678
2021-04-13	0.330046
2021-04-14	0.324609

344 rows × 1 columns

Table 8 - S&P 500 Index: cumreturns dataset

The pre-processing and analysis of the S&P 500 Index were concluded. It was also supposed that would be analysed two different correlations instead of one as previously defined—the Sentiment Analysis score correlating with S&P 500 returns and with the S&P 500 cumulative returns.

4.2. NEWS HEADLINES

After cleaning and preparing the dataset for the S&P 500 Index, it was time to create the dataset for the Media Outlet headlines. This dataset had the necessity of having a considerable number of observations that would allow the correlation to be significant. The main goal of this number should be at least 10 thousand observations during the time predefined.

After considerable research on how to create this dataset, it has been defined that the APIs, Application Programming Interface, would be used . Those APIs are an interface that connects two different applications, making them communicate between them and providing the information required from one application to the other. The first API used was the API from the New York Times (NYT) that is displayed on the official website of NYT. The NYT allows external users to access their API to download other news, headlines as a column called abstract, the URL of the news, descriptions, first paragraph and dates. Using their developer's network portal, a dataset called news could be created with 165 240 observations. Those observations were all the article news that NYT published between October 2018 and September 2021, as displayed in the table 9. Exporting all that news made working the dataset necessary since it had more headline dates than the time required and Headlines that were not correlated to the S&P 500 Index or the companies that compose that index.

	Column1.abstract	Column1.web_url	Column1.snippet	Column1.lead_paragraph	Column1.pub_date
0	The company is still struggling to produce and...	https://www.nytimes.com/2018/09/30/business/el...	The company is still struggling to produce and...	Elon Musk was chastened by federal regulators ...	2018-10-01T00:05:13+0000
1	The signs seemed to point to the Giants being ...	https://www.nytimes.com/2018/09/30/sports/gian...	The signs seemed to point to the Giants being ...	EAST RUTHERFORD, N.J. Given that the opponen...	2018-10-01T00:23:39+0000
2	Goodbye, #saddesklunch.	https://www.nytimes.com/2018/09/30/smarter-liv...	Goodbye, #saddesklunch.	Welcome to the Smarter Living newsletter, whic...	2018-10-01T00:44:58+0000
3	On Monday, the Cubs will host the Brewers and ...	https://www.nytimes.com/2018/09/30/sports/base...	On Monday, the Cubs will host the Brewers and ...	Major League Baseball will stage two games on ...	2018-10-01T00:45:38+0000
4	Quotation of the Day for Monday, October 1, 2018.	https://www.nytimes.com/2018/09/30/todayspaper...	Quotation of the Day for Monday, October 1, 2018.	Now books are becoming like drugs. You have t...	2018-10-01T01:02:50+0000
...	...	...	...	...	...
165235	Only nine of 52 African countries met the W.H....	https://www.nytimes.com/2021/09/30/world/afric...	Only nine of 52 African countries met the W.H....	JOHANNESBURG Only nine African countries hav...	2021-09-30T16:32:40+0000
165236	Those online eyeglasses and mattress start-ups...	https://www.nytimes.com/2021/09/30/technology/...	Those online eyeglasses and mattress start-ups...	This article is part of the On Tech newsletter...	2021-09-30T16:38:52+0000
165237	The Department of Justice filed a court brief ...	https://www.nytimes.com/2021/09/30/us/texas-ma...	The Department of Justice filed a court brief ...	The Justice Department signaled its support on...	2021-09-30T16:45:24+0000
165238	A recent indictment suggested that researchers...	https://www.nytimes.com/2021/09/30/us/politics...	A recent indictment suggested that researchers...	WASHINGTON The charge was narrow: John H. Du...	2021-09-30T16:53:50+0000
165239	The subject is one of the gravest topics in ar...	https://www.nytimes.com/2021/09/30/arts/design...	The subject is one of the gravest topics in ar...	Some headlines from the last few months. March...	2021-09-30T16:54:19+0000

Table 9 - New York Times Headlines Dataset

4.2.1. Headlines pre-processing

As mentioned before, the dataset must suffer some transformations to be used. Starting with the missing values, there were 653 observations on the column abstract, the column that displays the headlines, 4 444 observations on the column snippet and 1 799 observations on the column lead_paragraph. To deal with those missing values, the observations missing were filled with the word

Null making the observations not empty anymore. The decision to change the value to Null was based on the necessity of only using the headlines' column and dates for the correlation. With this, the other three columns, url, snippet and lead paragraph, were unnecessary to continue with the study being immediately dropped. At that moment, the dataset was composed of only two columns, date and headlines. Those columns counted 165 240 observations; however, the number of observations was inflated by the fact that the dataset contained all the Headlines published by the NYT between 2018 and 2021. This number includes even the news that was not correlated to the S&P 500 Index. To solve this issue and to only have the headlines associated with the S&P 500, a search was done to examine the word S&P 500 on every observation headline, deleting all the other observations that were not containing it. That drove the study to a final result of 86 observations between November 2018 and September 2021, a number far from what was decided to have been the goal of observations for the headlines dataset, as observed in the table 10. Those results dictated that it was necessary to use other APIs to achieve more observations.

Column1.pub_date	Column1.abstract
2018-11-07T09:00:30+0000	The S&P 500 has risen in four of the first fiv...
2018-12-03T20:12:18+0000	Wynn Resorts is among the best-performing stoc...
2018-12-17T10:00:42+0000	Stocks slumped through the afternoon, with the...
2018-12-18T10:06:25+0000	The S&P 500 hit its lowest point in 2018 on Mo...
2018-12-26T10:02:20+0000	Shares had their best day since 2009, but the ...
...	...
2021-02-26T12:57:56+0000	The bond markets are getting fussy.
2021-03-24T11:55:23+0000	The bull market is one year old. Can it last?
2021-07-25T11:00:06+0000	A new spike in coronavirus cases, driven by th...
2021-09-10T14:02:44+0000	Apple led the losses, but companies relying on...
2021-09-21T12:16:53+0000	The jitters over a potential collapse of the C...

68 rows × 1 columns

Table 10 - New York Time Final Dataset

A new research for an API was conducted to increase the number of observations of the dataset that was being used to store the headlines to reach the minimum of 10 thousand observations. The aim of this research for an API changed not being based on Media Outlet directly, newspapers or magazines anymore, but in companies that would have as a core business providing financial and business news as well twitter publications and other types of Media Outlet information to other companies. The best company to provide this kind of data is CityFALCON. Their API allows connecting to their database and exporting headlines from different sources and media outlets, from a range of dates and choosing the topic being searched. For this study, the dates between December 1st 2019 and April 15th 2021 and the subject correlated to the S&P 500 Index, including news from all the companies that compose that index. Using the CityFALCON API, it was possible to create a new dataset for headlines called NewsData, which contained 1 038 533 observations. This number of observations was way higher than

what initially predefined, making the dataset suitable for this study. The table 11 represents the dataset and has six different columns. The PublishTime, the CF UUID that represents the ID of the headline; the Title is the headline itself; the description; the URL of the headline; and the Source Name is the media Outlet where that published those headlines.

	ï»¿PublishTime	CF UUID	Title	Description	URL	Source Name
0	2019-12-01	87711ff6-3dfb-4499-b50b-7c24b4dcbf0e	Did Hedge Funds Drop The Ball On Centene Corp ...	How do we determine whether Centene Corp (NYSE...	https://finance.yahoo.com/news/did-hedge-funds...	Yahoo Finance
1	2019-12-01	69278f27-8996-425c-a301-69d067907d58	Is Danaher Corporation (DHR) A Good Stock To B...	The 700+ hedge funds and famous money managers...	https://finance.yahoo.com/news/danaher-corpora...	Yahoo Finance
2	2019-12-01	cfd8576b-68ad-45c1-95b9-35c2e825f3c0	Here is What Hedge Funds Think About Lennar Co...	We are still in an overall bull market and man...	https://finance.yahoo.com/news/hedge-funds-thi...	Yahoo Finance
3	2019-12-01	cedf7e14-5521-4ec5-84f8-2ef410f3fd34	Hedge Funds Have Never Been This Bullish On Oc...	It seems that the masses and most of the finan...	https://finance.yahoo.com/news/hedge-funds-nev...	Yahoo Finance
4	2019-12-01	f96227b2-6e7b-4c8e-986b-b499637ff683	Is PepsiCo, Inc. (PEP) Going to Burn These Hed...	We are still in an overall bull market and man...	https://finance.yahoo.com/news/pepsico-inc-pep...	Yahoo Finance
...	...	...	...	...	...	...
1038528	2021-04-15	94f6e380-3d86-4121-b781-922bd318746c	PC Server Power Management Software Market Fut...	April 15, 2021 (Reports and Markets) â PC Se...	https://jumbonews.co.uk/uncategorised/2847138/...	Jumbo News
1038529	2021-04-15	c29ef02c-1dc4-441f-96d3-18c4a5261379	Global Enterprise NAS Market 2025: Dell EMC, H...	Global Enterprise NAS Market: IntroductionThe ...	https://jumbonews.co.uk/politics/2847254/globa...	Jumbo News
1038530	2021-04-15	3a2fee45-fa21-4f97-8f8c-7b0179b0ce85	Global Dvr And Nvr For Use In Cctv Surveillanc...	The latest report on âDvr And Nvr For Use In...	https://jumbonews.co.uk/uncategorised/2846172/...	Jumbo News
1038531	2021-04-15	70deedd8-57fa-4634-9e42-434fb044fa9c	Citigroup profit triples on reserve release, t...	Citigroup Inc trounced market estimates for fi...	https://www.devdiscourse.com/article/business/...	Devdiscourse
1038532	2021-04-15	2ac4f7d4-f3a7-4ff5-876d-5ee0cffd6cd8	Global MetrologyInspectionand Process Control ...	Global Metrology,Inspection,and Process Contro...	https://jumbonews.co.uk/politics/2847314/globa...	Jumbo News

1038533 rows × 6 columns

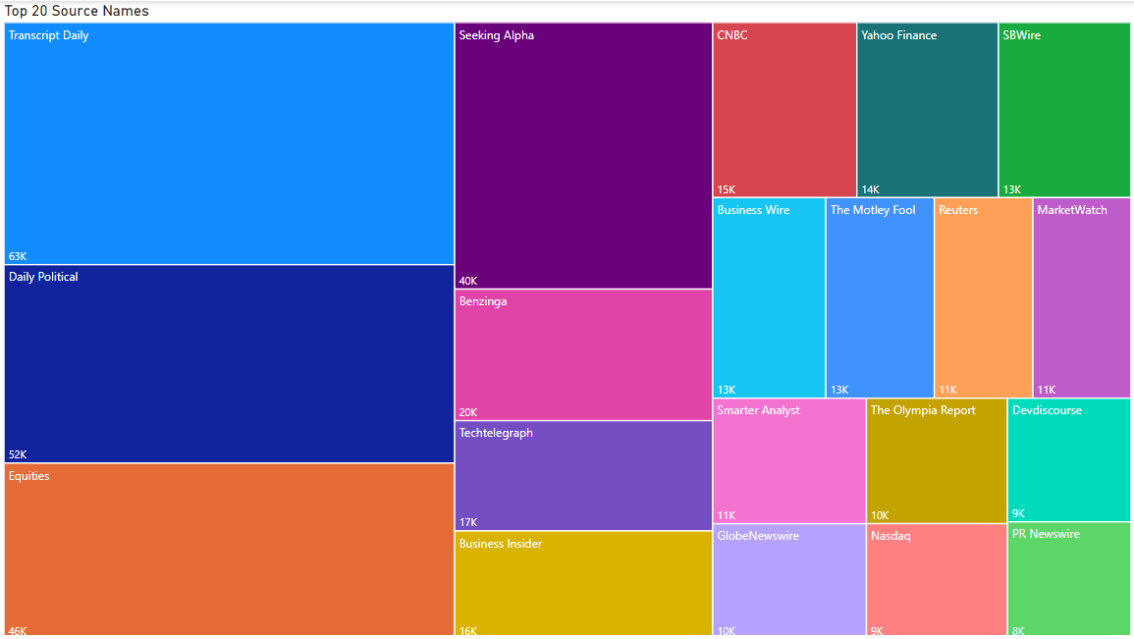
Table 11 - CityFALCON Headlines Dataset

As it was done on the News dataset from the NYT API, also this dataset needed to be pre-processed and analysed. Starting once again by checking the missing values was possible to see that were 26 observations missing in the column Source Name, 23 observations in the URL column and 56 094 observations in the description column. Since only the columns PublishDate and Title were the ones that would be used for the correlation was possible to fill all the missing values with a "Null" value, allowing the dataset not to have any missing value at any observation.

Before continuing with the pre-processing, a statistical analysis was done on the dataset to gain more knowledge about it. The table 12 represents the count of different Titles, the length of various dates and the length of the variable source names. It was possible to understand from this analysis that the dataset contained 502 different dates, 502 days, 4 853 different data sources and 1 009 732 different Headlines meaning that the dataset contained 28 801 duplicate headlines. Eliminating the duplicates, the length of different data sources changed to 4 801 distinct sources.

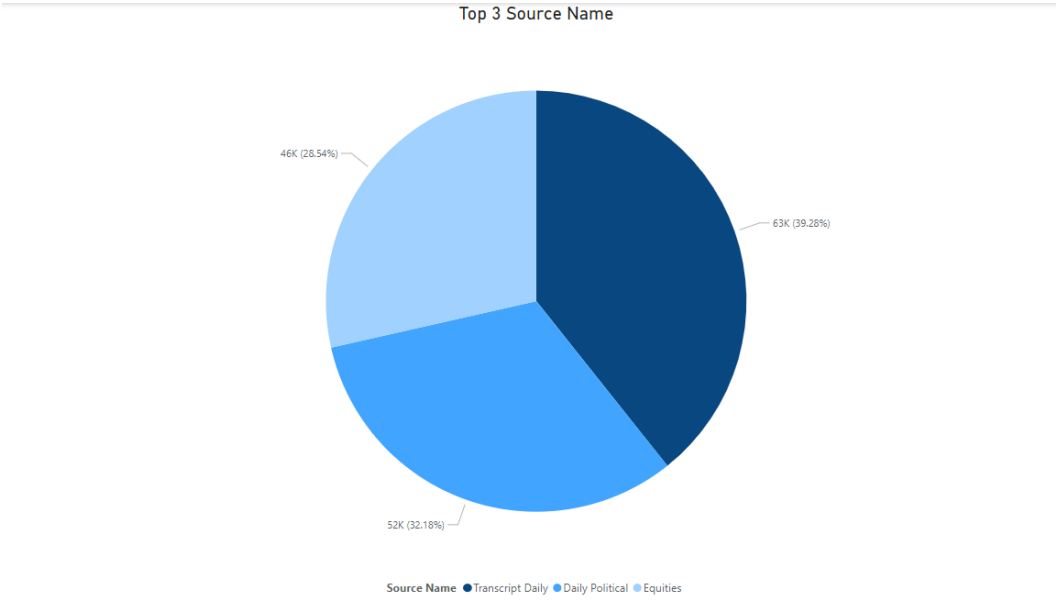
	Titles	Dates	Source Name with Duplicates	Source Name without Duplicates
Length	1009732	502	4853	4801

Table 12 - CityFALCON Headlines Dataset Info



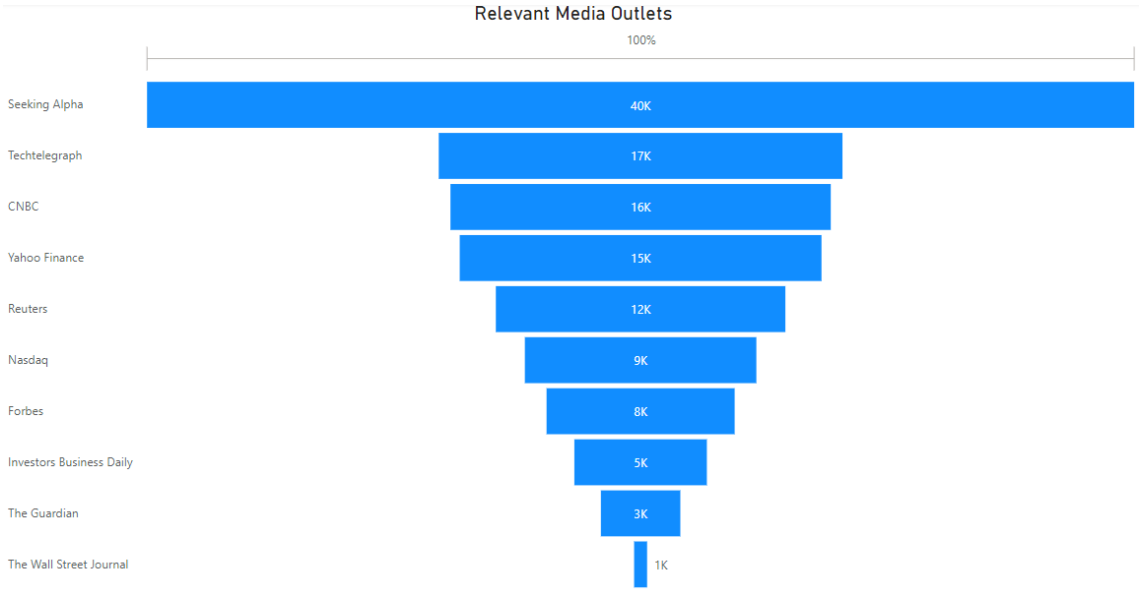
Graph 5 - Top 20 Headlines Data Sources Heatmap

It was also relevant to understand the dataset when grouped by the data source to recognise if it existed data sources with more relevance than others and the number of observations from distinct and relevant media outlets. In order to do so, a temporary dataset called "Sourcecount" was created that corresponds to the initial dataset but is grouped by the Source name column summing all the number of times that we have the same data source. The chart 5 illustrates the top 20 Source names with more observations. It was possible to understand that observations with the source name Transcript daily, Daily Political and Equities are the top 3 with 63 000, 52 000 and 46 000 observations, respectively, as is possible to see in Chart 6.



Graph 6 - Top 3 Headlines Data Sources

It was also relevant to understand if, on the data source, any critical media outlet with an impact on the financial world existed so that it could complement the study with headlines from impactful media outlets. Research on the dataset was conducted to realise the existence of these relevant media outlets, and it was possible to build a top 10 data source with a significant number of observations. These top 10 data Sources are given by Graph 7, containing media outlets such as Seeking Alpha, TechTelegraph, CNBC, Yahoo Finance, Reuters, Nasdaq, Forbes, Investor's Business Daily, The Wall Street Journal and the Guardian.



Graph 7 - Most Relevant Headlines Data Sources

After finishing this small dataset analysis, the pre-processing continued, this time text pre-processing on the headlines. For every step of this pre-processing, one column was created and added to the initial dataset with the results for every observation.

Removing the punctuation – The punctuation can be "noisy" to sentiment analysis and, in general, is irrelevant for the sentiment analysis, except when it is an exclamation or interrogation point because they can be useful for the sentiment analysis since they show a different degree of emotions that the VADER algorithm also identifies and takes into consideration. A list was created with the following punctuation and symbols: *"# $ % & \' ( ) * + , - . / : ; < = > @ [\ \ ] ^ _ ` { | } ~*. Using the loop function allowed Python to go observation by observation and, if located, to remove any punctuation or symbols on the previous list from the headline. The dataset now contained a new column called *clean_msg* that corresponds to the headlines without the punctuation or symbols.

Lowering the letters – Normally, any sentiment analysis attempts at removing the capital letters because it can bring different sentiment analyses to the algorithms as it does to VADER sentiment. So, using a function, Python went through, once again, every single observation from the column created in the previous step, named *clean_msg*, and lowered all the capital letters that were founded on a new column called *msg_lower*. However, even if this step is usually required for text pre-processing since the algorithm used to do the sentiment analysis would be VADER, and it considers the sentiment given by the capital letters, this column was not used for the rest of the study.

Tokenisation – Sometimes, the headlines contain too much information and are composed of more than one sentence, which can drive the sentiment analysis to get the info wrongly and do a wrong sentiment analysis because of containing too much information. For this purpose, a function called tokenisation was created, which splits the headline into smaller lines, breaking the sentence every time it finds a comma (,). With this function, a new column, called *msg_tokenied*, was created and corresponded to the column created on the first step, *clean_msg*, split into different lines in case there was a capital letter in the middle of the sentence.

StopWords – This sentiment analysis will not be relevant personal words such as "I", "me", and others, so a list called stopwords was created that is present on the library NLTK and contains not only pronouns, but also other irrelevant words for this study. A new column was built on the dataset called *no_stopwords* and corresponded to all the observations from the previous column made, *clean_msg*, and removed those words from the list that could be found on the observation. Table 13 corresponds to the words that were removed from the observations.

Stopwords list								
I	himself	that	a	through	here	own	re	ma
me	she	thatll	an	during	there	same	ve	mightn
my	shes	these	the	before	when	so	y	mightnt
myself	her	those	and	after	where	than	ain	mustn
we	hers	am	but	above	why	too	aren	mustnt
our	herself	is	if	below	how	very	arent	needn
ours	it	are	or	to	all	s	couldn	neednt
ourselves	its	was	because	from	any	t	couldnt	shan
you	its	were	as	up	both	can	didn	shant
youre	itself	be	until	down	each	will	didnt	shouldn
youve	they	been	while	in	few	just	doesn	shouldnt
youll	them	being	of	out	more	don	doesnt	wasn
youd	their	have	at	on	most	dont	hadn	wasnt
your	theirs	has	by	off	other	should	hadnt	weren
yours	themselves	had	for	over	some	shouldve	hasn	werent
yourself	what	having	with	under	such	now	hasnt	won
yourselves	which	do	about	again	no	d	haven	wont
he	who	does	against	further	nor	ll	havent	wouldn
him	whom	did	between	then	not	m	is	wouldn't
his	this	doing	into	once	only	o	isnt	

Table 13 – List of Words on StopWords

Stemming – The following process was removing the affixes of the words by using the function stemming from the NLTK library in Python. For that, a new column, *msg_stemmed*, was created based on the previous column, *no_stopwords*. For every observation, if the words had an affix, it was removed from the sentences. For example, the words friends became friend. These affixes sometimes can confuse the sentiment analysis algorithms by giving different meanings to the words making it harder for the algorithm to identify the real meaning of the sentences.

Lemmatizing - If in the previous step the affixes were taken out, on this step, the function lemmatiser from NLTK removed the canonical and dictionary forms; for example, the word running became run. A new column called *msg_lemmatized* was created and applied to every observation of the column *msg_stemming*, the function lemmatiser removing all the canonical forms of the words. For the same

reason that the words' affixes were released, their canonical forms were removed so the sentiment algorithm could not be "fooled" by these different ways of saying the same word.

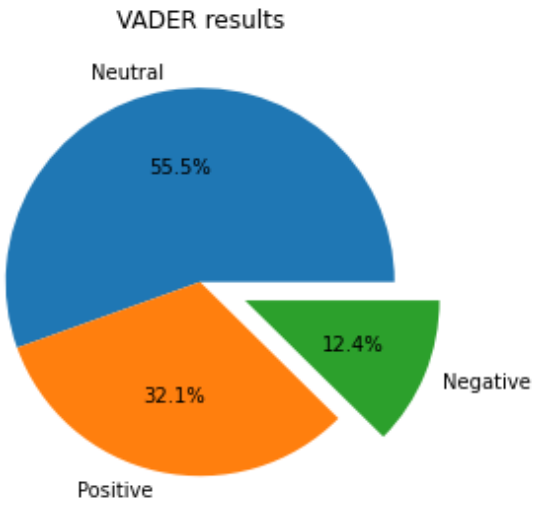
Those were all the steps of the data pre-processing of the headlines and the following topic will be related to the sentiment analysis.

4.3. Sentiment Analysis

The pre-processing phase of the headline was done, and there were only two relevant columns to continue the study, the last column of the text pre-processing, msg_lemmatized and the date. After dropping all the other columns, the dataset was ready for analyse. The algorithm chosen to analyse the sentiment of the data was the VADER. A function called vadersentimentanalysis was created and gave the score of the VADER sentiment as explained in the methodology that was allocated to a new column for each observation. Those scores were between -1 and 1. However, another column giving a status to the observations was created to visualise the sentiment analysis better. These statuses were defined by a function that determined if the sentiment score was higher or equal to 0.2, then the observation should be classified with the status Positive; if it would be shorter or equal to -0.2, then the observation would be organised with a Negative status, all the others should be classified with the status Neutral. This column was called Vader Analysis. Table 14 and Chart 8 show the results of the VADER sentiment. The table allows us to understand that we had a much higher number of observations with a score on the sentiment analysis between -2 and 2 560 432 observations, more precisely, that corresponds to 55.5% of the total of observations. The Positive group corresponds to 32.1% of the total observations, 324 073 different observations had a classification higher than 0.2, and only 12.4% of observations had a score below -0.2, corresponding to 125 227 observations.

Classification	Neutral	Positive	Negative
Total Number of Observations	560,432	324,073	125,227

Table 14 - Headlines Classification

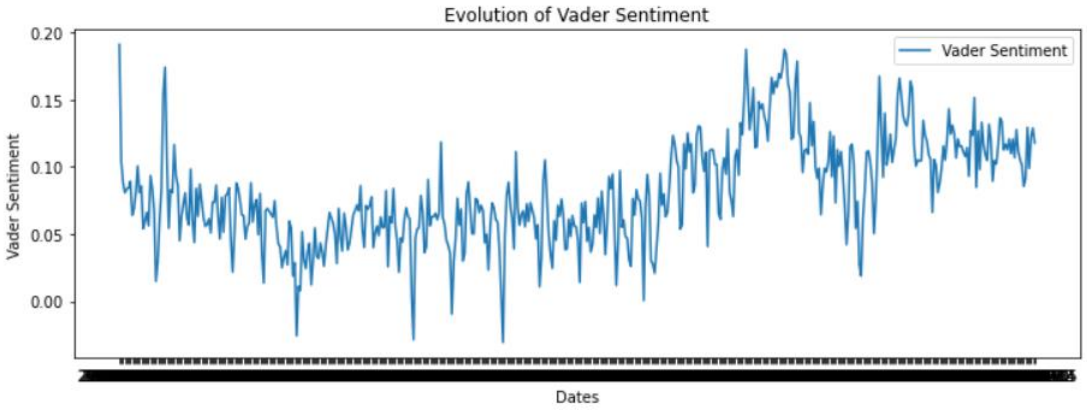


Graph 8 - Headlines Classification

The dataset contains 1 009 732 observations because we have more than one observation per day; however, to analyse the correlation between the sentiment score and the returns of the stock prices it was necessary to have the same number of observations on both datasets. In order to achieve the correct number of observations, and because the dataset contains more than one observation per day, it was necessary to group the dataset by dates, applying a mean calculation on the Vader Sentiment Score. Furthermore, the dataset contained only one observation per day, calculating the mean for every observation from the same day. The dataset at this point had 502 observations from 502 different dates. The statistical analysis was done and presented in Table 15. Those new 502 observations saw their maximum and minimum changes from 0.9983 to 0.1908 on the maximum value and from -0.9607 to -0.0306, meaning that the scores are much more concentrated now. This can be explained by the fact that every day there was news that could be considered positive and others negative. However, since half of the total observations was neutral and the neutral sentiment score was between -0.2 and 0.2, it concentrated all the score close to zero. The new score was given by the mean of the score per day, close to the Neutral Classification. The evolution of the Vader sentiment score is inconstant, having high and low values depending on the days as is possible to appreciate in Graph 9.

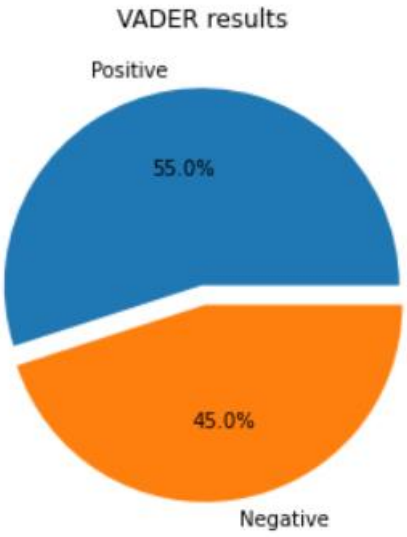
count	mean	std	min	0.25%	0.50%	0.75%	max
502	0.080822	0.038177	-0.030641	0.05497	0.074754	0.108547	0.190849

Table 15 - Headlines Descriptive Statistics



Graph 9 - Evolution of Headlines Score

Since the new dataset has a lower margin between the maximum and minimum, the news headlines were classified only as positive or negative. The number chosen to split the Headlines into positive and negative was 0.07, meaning that, when the sentiment score is higher or equal to 0.07, the headline would be classified as positive, and if lower than 0.07, the headline would be classified as negative. The choice of 0.07 was based on the quantiles, and the 50% quantile was 0.074. The approximation of the Vader score between all days regarding the group by function based on the mean for each day previously done made the results utterly different in terms of classification. The dataset contained at that moment 276 positive observations; this means that the score was higher than 0.07 and around 55% of the observations. On the other hand, there were 226 observations where the sentiment score was lower than 0.07, then classified as negative, representing 45% of the observations as is possible to understand from Graph 10.

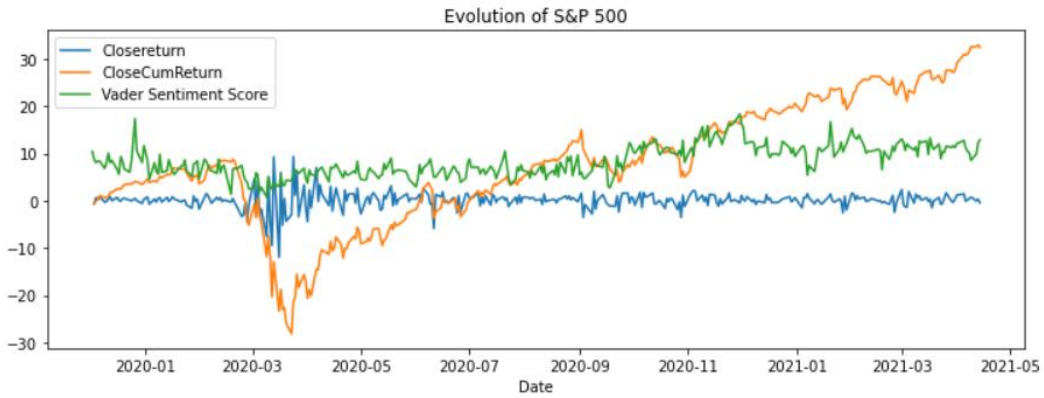


Graph 10 - Headlines Final Classification

The last step of the text pre-processing phase of the dataset was reducing the number of observations to the same number of observations obtained from the S&P 500 Index market exchange. As mentioned before, the Stock Exchange is not open every day; however, media outlets publish headlines every day, and between the dates of December 1st of, 2019 and March, 15th of 2021, there exist 502 days, and the dataset contains exactly 502 observations; however, the dataset with the S&P 500 Index close values only contains 344 observations, meaning that between the exact dates the Stock Market Exchange was open 344 days and not 502 days. It was necessary to deal with the other days. Since there were not available studies regarding the impact of those news on the next day of the market exchange, and this would be out of the aim of this research, it was decided to delete those days. Based on the index of the dataset, that was the date it was checked which days the dataset of the S&P 500 index didn't contain and deleted those days on the headlines dataset. When finished, both datasets had 344 observations, and the study was ready to move to the next step.

4.4. Stock price and Headline alignment

Since the two datasets, the headlines and the S&P 500 index, already suffered the necessary pre-processing and were aligned in terms of observations, containing precisely the same number of observations, it was possible to merge the two datasets. Using the formula join in Python it was possible to create the two new datasets. The first dataset was called Finalcumreturns and the other Finalreturns. The finalcumreturns dataset contained the date as an index that was the same for both datasets, the Vader sentiment score and the cumulative returns previously calculated on the pre-processing of the S&P 500 Index from the Close value. The dataset Finalreturns contains the same index, the dates, the Vader Sentiment score, however, instead of the cumulative values of the returns, it included only the returns of close values. Using those two datasets, it was viable to understand better the correlation between the sentiment score and the S&P 500 index and understand if there was any difference between the two calculations of the Close Values. The following chart gives the evolution of those two datasets. Through Chart 11 it is possible to evaluate the movement of the CloseReturns values, CloseCumReturns values and Vader sentiment Score. Those columns were added to one single dataset, which was the final dataset. There were similarities between the Closereturns shape and the Vader Sentiment Score, more than the CloseCumreturns on the line representative of the evolution of the S&P 500 over time. However, there was not possible to evaluate the correlation based on this chart, so the next step was applying correlation models to those variables.



Graph 11 - Evolution of S&P 500 and Headlines Score

4.5. CORRELATION MATRIX AND SPEARMAN'S RANK CORRELATION

The correlation was the last step to achieving the conclusion to the initial question: "Can a correlation be defined between the S&P500 index and News Headlines applying Machine Learning Algorithms?"

After all the previous processes, data manipulation, pre-processing phases, and data preparation, the correlation phase started with the final dataset containing three different variables. Those variables were the VaderSentimentScore, which represents the final score of the correlation, the Closereturn, which means the percentage of the difference between the Close values of the S&P 500 index of one day with the previous day, and the CloseCumReturn variable that represents the same as the Closereturn however, as a cumulative result.

To answer those questions, a correlation was made between the variables related to the S&P 500 Index (Closereturn and CloseCumReturns) with the variable from the headline analysis (VaderSentimentScore). Python started to evaluate a Pearson correlation between the variables, as the Correlation Matrix in Table 16 can demonstrate. The colours used on this chart allowed a straightforward evaluation of the correlation coefficients. As it was closer to dark red, the coefficient was more prominent, meaning the higher the positive or negative relationship between the two variables. A strong relationship indicates that both variables move in the same direction, positively or negatively. For this correlation it was possible to understand that, when correlating the Closereturrn with the VaderSentimentScore, the cell was blue, meaning that the coefficient is low. The coefficient for this result was 0.17, meaning that it was a very low correlation between those two variables. On the other hand, when checking the correlation between the CloseCumReturn variable and VaderSentimentScore, the cell was light red, meaning that a correlation exists but it was not the strongest possible. The correlation coefficient was 0.68, suggesting a moderate correlation exists between those two variables. The results show a correlation between the S&P 500 Index cumulative returns and the VaderSentimentScore from the headlines. As an initial guess, it would be possible to answer positively to the question done at the beginning of the study and say that a correlation exists between S&P 500 Index and the Headlines, on a preliminary basis. However, it would be a premature conclusion.

	Closereturn	CloseCumReturn	VaderSentimentScore
Closereturn	1.000000	0.083438	0.173792
CloseCumReturn	0.083438	1.000000	0.681678
VaderSentimentScore	0.173792	0.681678	1.000000

Table 16 - Pearson Correlation Coefficients

To answer more precisely the questions, it was necessary to apply another correlation, called Spearman's Rank Correlation, to double-check the results as demonstrated on the Correlation Matrix in Table 17. The Spearman's Correlation chart works the same way as the previous one about Pearson's Correlation; the colours represent the same. Also, the variables used were the same, Closereturn,

CloseCumReturn and VaderSentimentScore, to this correlation. Spearman's correlation concluded the same as the Pearson correlation with similar results. It was possible to comprehend that for Spearman's correlations, as it also had already happened with the Pearson's correlation, there was no correlation between the variables VaderSentimentScore and Closereturn with a coefficient of 0.16 being very similar to the 0.17 given by Pearson's Correlation. When checking the correlation between the CloseCumReturn and the VaderSentimentScore, it was possible to understand that a high correlation exists, and the coefficient was 0.72, slightly higher than the one given on Pearson's correlation of 0.68.

	Closereturn	CloseCumReturn	VaderSentimentScore
Closereturn	1.000000	0.031661	0.167673
CloseCumReturn	0.031661	1.000000	0.725121
VaderSentimentScore	0.167673	0.725121	1.000000

Table 17 - Spearman's Correlations Coefficients

4.6. Statistical Significance Test

The coefficients are typically not enough to admit if there exists or not a correlation between variables. On both correlation methods, Pearson's and Spearman's, it has been necessary to run a statistical significance test, since the coefficients don't give all the required information to affirm the correlation between them. Those statistical significance tests allowed to quantify the relationship between the variables and say if the coefficient received is or is not representative for the entire dataset. For this statistical test it was necessary to apply hypothesis testing.

H0: Is a null hypothesis, meaning that there is not a significant correlation between the variables and where p = 0

H1: Alternative hypothesis, meaning that the correlation can be significant between the variables and p != 0

For this statistical test it was necessary to calculate the P-Value. The P-Value represents the probability that the null hypothesis, H0, is true. The lower the score, the higher the probability of H0 being actual.

The correlations done for this study had a very low P-Value, meaning that the correlations between the variables were significant. When applying the Pearson correlation statistical test, the variables Closereturn and VaderSentimentScore and the correlation of the variables CloseCumReturn and VaderSentimentScore had a P-Value of 0.001 and 3.3E-48, respectively. Those P-Values allowed it to be concluded that both the correlations were significant and even more relevant, the correlation between CloseCumReturn and VaderSentimentScore. When applying the same statistical test to the Spearman's correlations it was possible to understand that there was also a significant correlation between Closereturn and VaderSentimentScore variables given by a P-Value of 0.002 and an even more robust and significant correlation between the variables CloseCumReturn and VaderSentimentScore, provided by a P-Value of 3.4E-57.

Table 18 constitutes a summary of the two different correlations in one single table to make it easier to analyse the final results.

	Pearson				Spearman			
	Closereturn		CloseCumReturn		Closereturn		CloseCumReturn	
	Coeficient	P-value	Coeficient	P-value	Coeficient	P-value	Coeficient	P-value
VaderSentimentScore	0.17379237	0.00123118	0.68167843	3.37E-48	0.16767298	0.0018326	0.72512142	3.35E-57

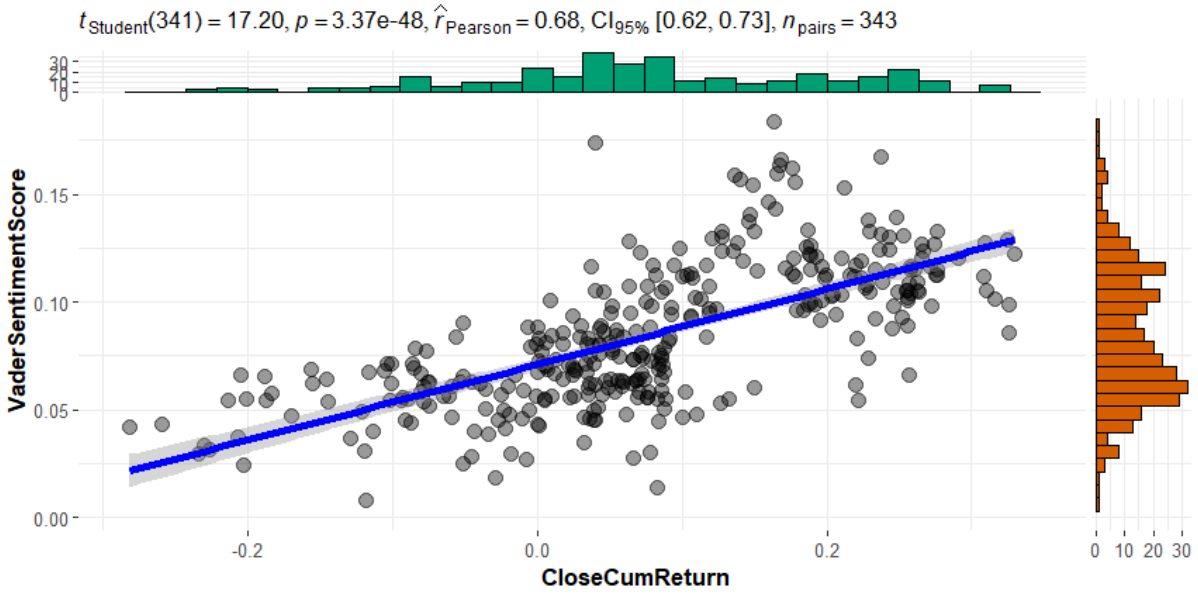
Table 18 - Pearson and Spearman's Correlation Resume

It was possible to understand from the previous table that a strong correlation between the variables CloseCumReturn and the VaderSentimentScore exists in both of the correlations calculated, Pearson's and Spearman, with a positive relationship. This relationship was given by high coefficients of 0.68 and 0.72 and also had a vital significance given by the P-Value of e.eE-48 and 4.4E-57 from the statistical tests. The correlation between the CloseReturn and VaderSentimentScore variables had a significant correlation; however, it did not contain a strong correlation between the variables as it was also possible to align by reading the coefficients and P-Value of the correlation.

4.7. Robustness Check

To test the correlation in Python with the two different algorithms, Pearson and Spearman's correlation, it was necessary to perform a robustness check on the results using another software, R, to increase the outcome's reliability. This robustness check was performed using different methods to calculate the Pearson and Spearman's correlation in R. The difference in the software ensured the soundness of the results.

By confronting the results from each software, it was possible to check that they were the same for Python and R as it is possible to see in Chart 12, where the Pearson correlation in R between CloseCumReturn and VaderSentimentScore variables was seen. The P-Value was 3.3E-37, maintaining a status of highly significant correlation, and the coefficient was 0.68, the same coefficient obtained in Python. The same robustness check gave the same results of coefficients and P-Values as it is possible to see in the Appendix1, 2 and 3.



Graph 12 - Pearson Correlation Coefficients and P-Value in R

In light of the presented results, the initially stated research question – "Can a correlation be defined between the S&P500 index and News Headlines applying Machine Learning Algorithms?" can be confirmed after thoroughly checking through a robustness analysis. In the following paragraph, the discussion will provide future research suggestions and a complete investigation of the main weaknesses and open questions this research has yet to encompass.

4.8. Discussion

The research started with one question and hypothesis to be debated and studied. The hypotheses were contradictory, and only one could be correct at the end of the study.

H1: A correlation between the S&P 500 index and the Sentiment Analysis of News Headlines exists.

H2: A correlation between the S&P 500 index and the Sentiment Analysis of News Headlines does not exist.

In the rational view of the outcome provided by the results presented above, it is possible to ensure that this research achieved the conclusion that H1 can be confirmed. The results are aligned with the theories presented in the second chapter.

For instance, the correlation between the percentage difference of the Close values of the S&P 500 and the VADER Sentiment score cannot be considered regarding the low coefficient. However, the correlation between the same VADER Score and the cumulative returns of the Close Values of the S&P 500 is a moderated/strong correlation, confirmed by a high coefficient. The commitment to reinsure that the correlation could be trustful drove the research to implement a second correlation algorithm. Spearman´s rank correlation demonstrated similarly on the correlation coefficients between both variables. Two correlation methods confirm a correlation between the cumulative returns of the S&P 500 and the VADER Sentiment Score.

The statistical tests proved that the correlation of every correlation was significant. Both variables, the returns and the cumulative returns, had a very low P-Value when correlated with the VADER Score. Based on these statistical tests it is possible to affirm that both variables are significantly correlated with the VADER Sentiment Score. However, it would have been necessary to confront the results obtained with a more robust check. The robust check came to confirm and ensure that the results in Python were trustworthy, so the correlation was also done in R. By achieving the same results in coefficients and P-Values, received in Python when correlating the variables. The robustness check successfully validated the strong positive and significant correlation between the cumulative returns of the S&P 500 and the VADER Score.

Considering the results previously shown and opposing them with the two hypotheses defined at the beginning of the study, it was possible to ensure that a correlation exists between the S&P 500 index and the Sentiment Analysis of News Headlines, accepting H1.

Consequently, H2 must be refused since it was possible to achieve a correlation between the S&P 500 index and the news headlines.

5. CONCLUSIONS

In summary, this thesis aimed at analysing the impact of media outlet headlines on the Market Exchange, specifically on the S&P 500 Index, by answering the following research question:

(RQ): Can a correlation be defined between the S&P500 index and News Headlines applying Machine Learning Algorithms?

In hindsight, it is possible to affirm that it is possible to define a correlation between the S&P 500 index and News Headlines by applying Machine Learning Algorithms. Specifically, applying machine learning algorithms that allow us to Score the Sentiment Analysis from the headlines and confront them with the S&P 500 index.

Indeed, the statistical analysis conducted by the thesis, managed to confirm 50% of the initially stated hypothesis, reported here below, namely H1:

H1: Exists a correlation between the S&P 500 index and the Sentiment Analysis of News Headlines.

H2: Does not exist correlation between the S&P 500 index and the Sentiment Analysis of News Headlines.

Since no previous study had linked the correlation with such a high number of observations from the Media Outlet headlines - at least to the author's best knowledge - this thesis has proven to be unique. This thesis also demonstrates to bring value to the financial and data science community by reaching a possible correlation between the fields. The future perspectives resulting from the establishment of the correlation between the S&P 500 index and Sentiment Score aim to possibly bring different strategies to the investment market and take into account more integrated decisions with the headlines analysis.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

The main limitations of this analysis stem from the methodology, which constitutes the main strength of the research but also presents a restraint: the number of observations gathered from the media outlet, the headlines, is a very considerable number however lacks a higher number of observations from more impactful and known media outlet companies. This aspect, nonetheless, opens the possibility for future research to be done by taking the option of using the correlation here defined and apply to a more robust dataset.

A second limitation of the results presented in this research is related to the need for theories on the correlation between those two fields: financial and sentiment analysis field. Specifically, the need for theory on the sentiment analysis of financial information. Sentiment analysis is a topic in a growing phase; however, the procedures of sentiment analysis are mainly the same no matter what the subject of the text is. Text can be related to many different fields, as it was possible to read in the literature review, aeronautics, tourism, finance and even client and employee satisfaction. The process of pre-processing is very similar to all the fields previously mentioned, yet the sentiment for the reader is not taken into consideration the fields of the text. Does a different sentiment for the same word exist when considering different fields? Can a word have a double meaning depending on the field? The word "growing" in the financial field can have a positive but negative status in the aeronautics field. Those are all questions that might conduct a new research in the text mining field: the sentiment of the words in different fields.

Consequently, from what was stated previously, future research should be foreseen to battle the limitations mentioned. A more detailed choice of observations can be made, and a future correlation between the cumulative returns of the S&P 500 index and the Sentiment Score of the more impactful media outlet diminishing the number of observations to only those media outlet companies.

Furthermore, a much-needed future prospect is related to the limitation of sentiment analysis for different fields. A more precise dictionary of words, with more words related to the financial field and stock markets, can be built and added to the VADER dictionary. This future study makes the sentiment algorithm more robust, specifically for the financial area. This more precise dictionary could drive the correlation results to a higher level of correlation, achieving higher results by preparing the observations on the sentiment score more accurately.

7. BIBLIOGRAPHY

Alamsyah, A., & Ginting, D. M. (2018). Analyzing Employee Voice Using Real-Time Feedback. *4th International Conference on Science and Technology (ICST).* Yogyakarta, Indonesia : IEEE.

Alamsyah, A., Rikumahu, B., & Ayu, S. R. (2019). Exploring Relationship between Headline News Sentiment and Stock Return. *7th International Conference on Information and Communication Technology (ICoICT).* Kuala Lumpur, Malaysia : IEEE.

Alanyali, M., Preis, T., & Moat, H. S. (2013). *Quantifying the Relationship Between Financial News and the Stock Market.* Scientific Reports.

Allen, F., & Karjalainen, R. (1999). Using genetic algorithms to find technical trading rules. *Journal of Financial Economics 51*, 245-271.

Asur, S., & Huberman, B. A. (2010). Predicting the Future with Social Media. *International Conference on Web Intelligence and Intelligent Agent Technology.* Toronto, ON, Canada : IEEE.

Batres-Estrada, B. (2015). Deep learning for multivariate financial time series.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research 3*, 1137–1155.

Bijari, M., & Khashei, M. (2011). Mehdi Khashei. *ScienceDirect*, 2664-2675.

Bollen, J., & Mao, H. (2011). Twitter mood predicts the stock market.

Breen, J. (2011, Jul 04). R by example: mining Twitter for consumer attitudes towards airlines. Boston: Predictive Analytics MeetUp.

Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*, 48 - 57.

Campanella, C., Mustilli, M., & D'Angelo, E. (2016). Efficient Market Hypothesis and Fundamental Analysis: An Empirical Test in the European Securities Market. *Review of Economics & Finance*, 27-42.

Cao, L. J., & Tay, F. (2003). Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Transactions on Neural Networks*, 1506 - 1518.

Chicago Mercantile Exchange. (1988). *S&P 500: Using S&P 500 Stock Index Futures and Options.* Chicago: Chicago Mercantile Exchange.

Chowdhury, S. G., Routh, S., & Chakrabarti, S. (2014). News Analytics and Sentiment Analysis to Predict. *International Journal of Computer Science and Information Technologies*, 3595-3604.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep Learning for Event-Driven Stock Prediction. *Twenty-Fourth International Joint Conference on Artificial Intelligence.* Buenos Aires, Argentina: IJCAI.

Duan, J., Liu, T., Zhang, Y., & Ding, X. (2014). Using Structured Events to Predict Stock Price Movement: An Empirical Investigation. *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1415–1425). Doha, Qatar: Association for Computational Linguistics.

Fung, G., Lu, H., & Yu, J. (2005). The Predicting Power of Textual Information on Finantial Markets. *IEEE Intell. Informatics Bull*, 1-10.

Gentleman, R., & Ihaka, R. (2003). *What is R?* Retrieved from The R Project for Statistical Computing: https://www.r-project.org/about.html

Gupta, R., & Chen, M. (2020). Sentiment Analysis for Stock Price Prediction. *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR).* Shenzhen, China: IEEE.

Huang, W., Nakamori, Y., & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 2513-2522.

Jiang, M., Liu, J., Zhang, L., & Liu, C. (2020). An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. *ScienceDirect*.

Khedr, A. E., Salama, S., & Yaseen, N. (2017). Predicting Stock Market Behavior using Data Technique and News Sentiment Anal. *Modern Education and Computer Science*.

Kim, K.-j. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, 307-319.

Kirange, D., & Deshmukh, R. (2016). Sentiment Analysis of News Headlines for Stock. *International journal of advanced computer technolog*.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! *International AAAI Conference on Web and Social Media* (pp. 538-541). Palo Alto, California, USA: AAAI Press.

Lo., A. (2017). *Adaptive Markets: Financial Evolution at the Speed of Thought.* New Jersey, United States: Princeton Univers. Press.

Malhotra, N., & Birks, D. (2007). *Marketing Research: An Applied Approach.* New Jersey, United States: Prentice Hall/Financial Times.

Malkiel , B. G. (2003). The Efficient Market Hypothesis and Its Critics. *Journal of Economic Perspectives*, 59-82.

Malkiel, B. G. (1999). *A Random Walk Down Wall Street; Including a Life-Cycle Guide to Personal Investing.* W. W. Norton & Company.

Manahov, V., & Hudson, R. (2014). A note on the relationship between market efficiency and adaptability – New evidence from artificial stock markets. *Expert Systems with Applications*, Expert Systems with Applications.

Melo, B. (2012). Considerações cognitivas nas técnicas de previsão no mercado financeiro. Universidade Estadual de Campinas.

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. (2019). Stock Price Prediction Using News Sentiment Analysis. *IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService).* Newark, CA, USA : IEEE.

Nagar, A., & Hahsler, M. (2012). Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live Streams. *International Conference on Computer Technology and Science.* New Delhi, India: IACIT Press, Singapore.

Nair, D. S., Jayan, J. P., R.R, R., & Sherly, E. (2015). Sentiment Analysis of Malayalam film review using machine learning techniques. *International Conference on Advances in Computing, Communications and Informatics (ICACCI).* Kochi, India : IEEE.

Nassirtoussi, A. K., Aghabozorgi, S., & Wah, T. Y. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *ScienceDirect*, 306-324.

Nofsinger, J. R. (2001). The impact of public information on investors. *Journal of Banking & Finance*, 1339-1366.

Novak, M. G., & Velušček, D. (2014). Prediction of stock price movement based on daily high prices. *Quantitative Finance*, 793-826.

Pagolu, V., Reddy, K., Panda, G., & Majhi, B. (2017). Sentiment analysis of Twitter data for predicting stock market movements. *International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES).* Paralakhemundi, India : IEEE.

Rossum, G. v. (1990). *What is Python? Executive Summary*. Retrieved from python: https://www.python.org/doc/essays/blurb/

Santhanam, P., & Devarasan, E. (2013). An analysis on Stock Market Prediction using Data Mining Techniques. *International Journal of Computer Science & Engineering Technology (IJCSET)*, 49-51.

Schindler, P. S., & Cooper, D. R. (2003). *Business research methods.* New York: McGraw Hill.

Schmidhuber, J., & Hochreiter, S. (1997). Long Short-Term Memory. *Neural Computation*, 1735–1780.

Shah, D., Isah, H., & Zulkernine, F. (2018). Predicting the Effects of News Sentiments on the Stock Market. *IEEE International Conference on Big Data (Big Data).* Seattle, WA, USA: IEEE International Conference on Big Data (Big Data).

Shonkwiler, C., & Cantarella, J. (2013). The Symplectic Geometry of Closed Equilateral Random Walks in 3-Space. arXiv.

Shynkevich, S., McGinnity, T., & Colema, S. (2015). Predicting Stock Price Movements Based on Different Categories of News Articles. *IEEE Symposium Series on Computational Intelligence.* Cape Town, South Africa : IEEE.

T., C., & R., B. (2013). Self-adaptive and sensitivity-aware QoS modeling for the cloud. *8th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)* (pp. 43-52). Melbourne, Australia: IEEE.

Thomsett, M. (2015). *Technical Approach To Trend Analysis: Practical Trade Timing for Enhanced Profits.* London, UK: Financial Times Prent.

Vargas, M., Lima, B., & Evsukoff, A. (2017). Deep learning for stock market prediction from financial news articles. *International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA).* Annecy, France : IEEE.

Xiang, Z., Du, D., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, 51-65.

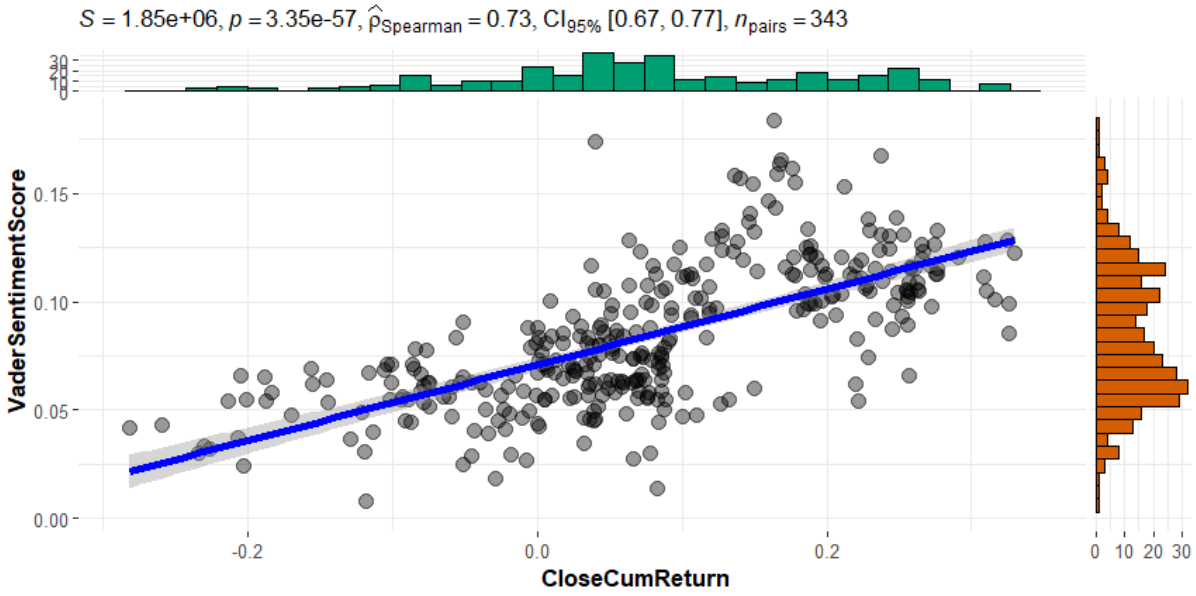
Yu, P., & Yan, X. (2020). Stock price prediction based on deep neural networks. *Springer Link*, 1609-1628.

Yu, W.-B., Guruswamy, B., & Lea, B.-R. (2007). A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting. *International Journal of Electronic Business*, 211-224.

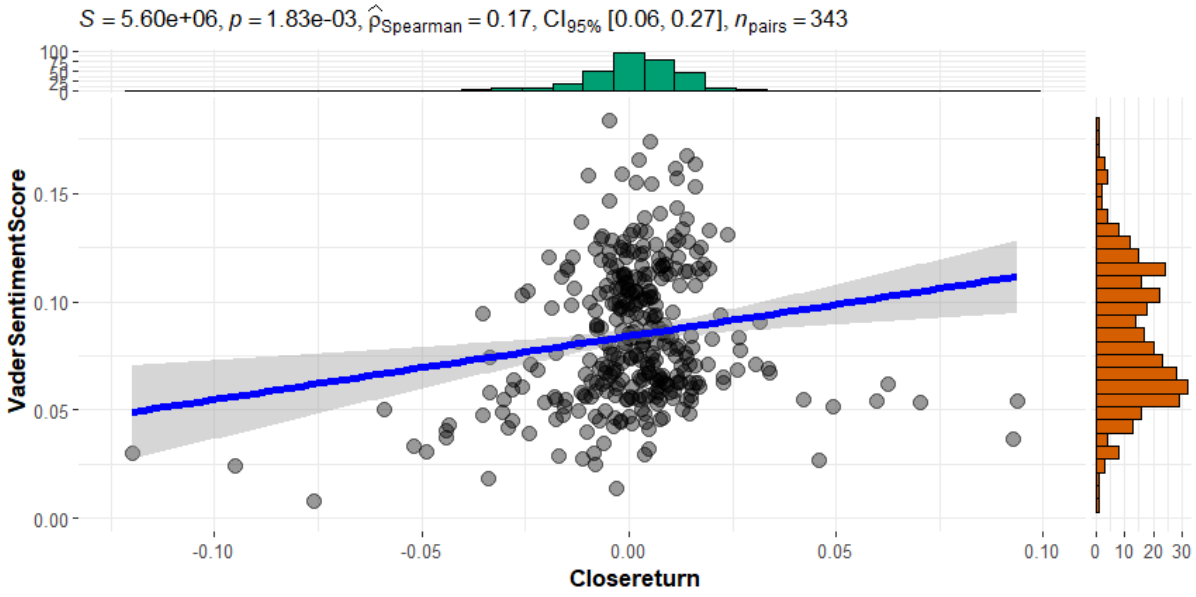
Zarandi, M. F., Rezaee, B., Turksen, I., & Neshat, E. (2009). A type-2 fuzzy rule-based expert system model for stock price analysis. *ScienceDirect*, 139-154.

Zhang, K., Zhong, G., Dong, J., Wang, S., & Wang, Y. (2019). Stock Market Prediction Based on Generative Adversarial Network. *Procedia Computer Science*, 400-406.

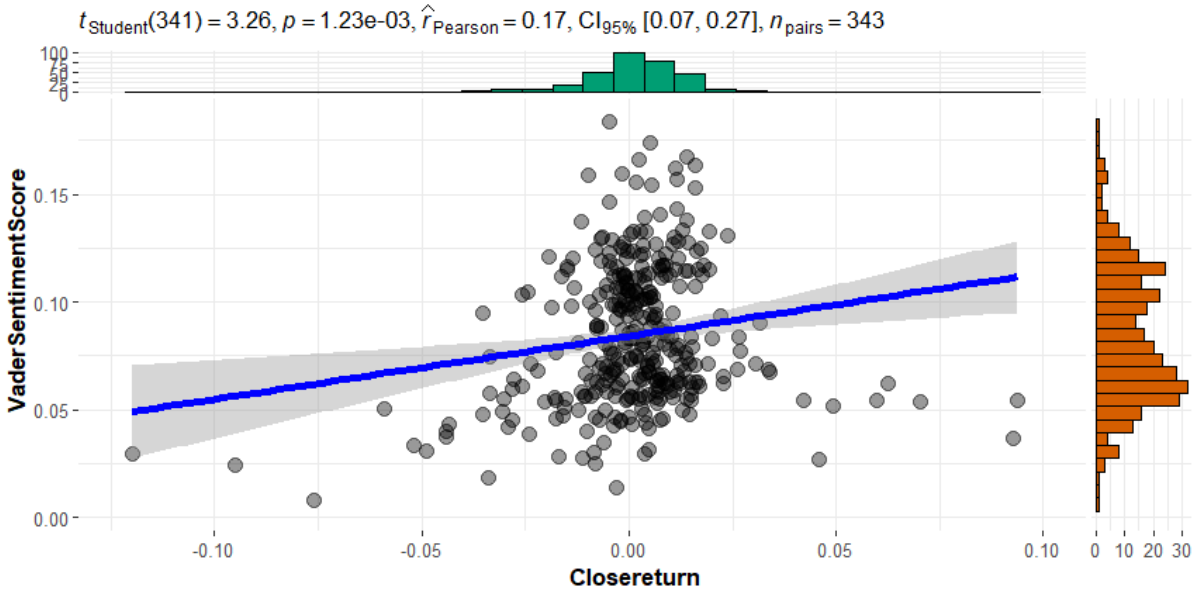
8. APPENDIX



Graph 13 - Appendix 1



Graph 14 - Appendix 2



Graph 15 - Appendix 3