Instituto Tecnológico y de Estudios Superiores de Monterrey

EGADE Business School

México City, México

A comprehensive analysis of behavioural economics applied to social media using automated methods and asymmetric modelling

A dissertation presented by

Román Alejandro Mendoza Urdiales

Submitted to the
EGADE Business School
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

In

Financial Sciences

México City, Sept 1, 2022

Declaration

I, Román Alejandro Mendoza Urdiales, declare that this dissertation titled, *A comprehensive analysis of behavioural economics applied to social media using automated methods and asymmetric modelling* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this dissertation has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this dissertation is entirely my own work.
- I have acknowledged all main sources of help.
- Where the dissertation is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

_____

Román Alejandro Mendoza Urdiales
September 1st, 2022

Acknowledgements

"Intelligence is a gift, not a privilege and you use it for the good of mankind."

Otto Octavius, Dr.

To the research Project FOSEC SEP-INVESTIGACIÓN BÁSICA through Consejo Nacional de Ciencia y Tecnología, México, A1-S-43514 (https://conacyt.mx/ciencia-de-frontera/, AGM). which supported the publication of the paper *"Measuring information flux between social media and stock prices with Transfer Entropy"*

To Dr. Andrés García who took the time to supervise and collaborate with this project.

To my wife Miriam, who took this journey with me and encourages me every day to be the best version of me.

To my father, my mother and my brothers, the ones who patiently heard me all those never-ending nights talking about school.

To Dr. Juan Sampieri who landed me that interview at the Firm.

…And finally, to my high school math teacher, who told me that I would never be anybody in life. He can now call me, Dr. nobody.

A comprehensive analysis of behavioural economics applied to social media using automated methods and asymmetric modelling

By

Román Alejandro Mendoza Urdiales

Abstract

Financial economic research has extensively documented the fact that the impact of the arrival of negative news on stock prices is more intense than that of the arrival of positive news. The authors of the present study followed an innovative approach based on the utilization of two artificial intelligence algorithms to test that asymmetric response effect. Methods: The first algorithm was used to web-scrape the social network Twitter to download the top tweets of the 24 largest market-capitalized publicly traded companies in the world during the last decade. A second algorithm was then used to analyze the contents of the tweets, converting that information into social sentiment indexes and building a time series for each considered company. After comparing the social sentiment indexes' movements with the daily closing stock price of individual companies using transfer entropy, our estimations confirmed that the intensity of the impact of negative and positive news on the daily stock prices is statistically different, as well as that the intensity with which negative news affects stock prices is greater than that of positive news. The results support the idea of the asymmetric effect that negative sentiment has a greater effect than positive sentiment, and these results were confirmed with the EGARCH model.

Table of contents

List of figures

List of tables

Chapter 1: Introduction

Over 200 years ago, Napoleon Bonaparte was defeated by the British in Waterloo ending with his Empire. The relevant part for this study is what happened afterwards. A British businessman and banker profited greatly from knowing this fact well before it became common knowledge, or at least, this is the story as told. This businessman name was, indeed, Nathan Rothschild. How this gentleman son of jew immigrants performed such an incredible task is what matters to us. It is virtually part of England's history that the Rothschild courier and communications network had gained a justifiable reputation for speed and reliability and that the Government at the time had already failed to establish a similar network of its own and had been let down by other more established London firms. It confirms that the Rothschild couriers brought news of victory at Waterloo "*a full 48 hours before the government's own riders brought the news to Downing Street*". With this time window advantage, Nathan was able to play the market to his personal pleasure.

Now a days the history repeats itself using the tools that now are at hand. There are countless studies that aim to measure how the news impact the performances of the markets and how the news is delivered, by primary or secondary hand. More important is to understand how deep the effect of the news in the market performance is. The interest in measuring the signal from news to the stock market can be tracked back over a century and it has become an evolving discipline. Even when the methodologies of analysis and sampling techniques have improved over time the opinions still are divided between two branches. In one hand, there are studies that conclude that its not possible to measure quantitatively the impact of news in the stock market. The other branch has stated that there is definitely flow from the news into the stock market.

This study makes an historical comprehensive literature review of finance focusing on demonstrating the impact of the news in the stock market, comparing methodologies and sampling techniques. We concluded that more important than the methodology used, is how the sampling is performed and structured. Since we are taking qualitative variables into analysis, how to consider their participation in the modelling sample introduces a new factor into consideration for the model construction process.

The basic general framework that is common through all the studies can be divided in three steps:

1. Data mining. - Extraction, what data will be collected and what methodology is applied for the collection.
2. Data structuring. - When the dataset is put together, how the qualitative variable is transformed into a quantitative measurable useful for modelling variable is crucial.
3. Modelling. - When the data is collected, transformed an structured, running a model to measure statistical significance is the only part left before reporting results.

This framework was followed basically in each chapter of this study. Variations were performed in the structuring of the data and different models were tested in each chapter. We are glad to report that in all the different methods we concluded similar positive results.

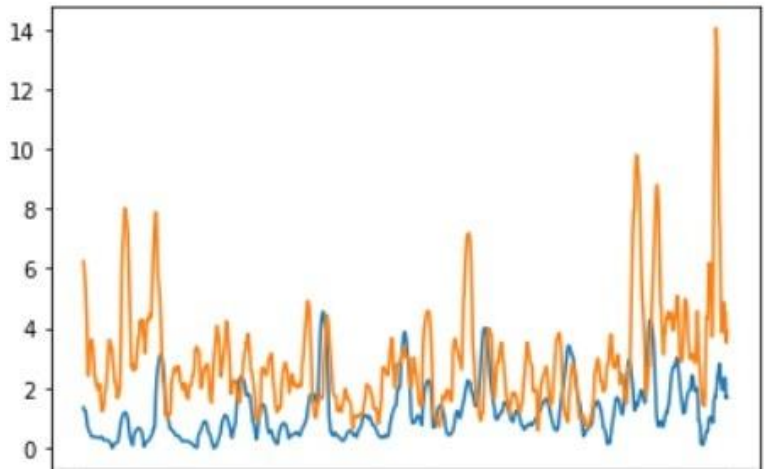
Some considerations to take are for the framework mentioned are that the data mining can be done very differently for each study, and this has proven different results for all the studies

mentioned in this report. The data can be extracted both manually and automatically, each method has its own risks, for example, extracting the data manually is slow and humans make mistakes transcribing data from one document to another, omissions can be done, and subjective criteria applied differently throughout the extraction process. In the other hand, automating the task can improve speed and mining accuracy, yet still the human behind the programming can make mistakes, which will be transferred to the code.

The data structuring is a science by itself, a full branch of research focuses only on understanding text subjectively and measure its objective. This branch is known as Natural Language Processing, and it's considered a very complex study field. Since language is created by humans it can held different meanings the same word base on tone, expression, context, and volume in which its mentioned. How to teach computers to understand human language has been a never-ending improving area.

The last step is a different challenge, in most of the studies analysed and mentioned in this document, classic statistics are not enough for measuring the signal between the qualitative variables and the stock market. As an example, we present Figure 1.1, In which we are comparing the 35 day rolling window historical volatility of the performance of a certain stock compared to the 35 rolling window of the sentiment mentioning the stock and even when it can be appreciated visually a certain synchronization of the movement of both time series the correlation is 24%, meaning that the causality between both time series could be discarded from the beginning if no further analysis is performed.

Figure 1.1 Historical volatility of stock performance of a company and volatility of the sentiment of the same company, using a 35-day rolling window for both cases with a correlation of 24%.



Going further into the causality analysis if a regression analysis is performed, generally we would have a good statistical result from the sentiment, considered as an independent variable versus the stock performance as dependent variable. Let's take for example Figure 1.2, we have that for the results of the Ordinary Least Squares regression (OLS) for Amazon an R squared of 95% which would be a great fit, the independent variable analysed returned great statistical values and it could be considered a successful analysis. Yet, the Durbin-Watson for the residuals came back with negative results, meaning that the model should be discarded. Even more, when

plotting the residuals, it would be apparent that normality is being kept in the model but applying a Shapiro Wilks test confirms the diagnose that the model does not work. This has been a challenge found in a large part of the literature reviewed here.

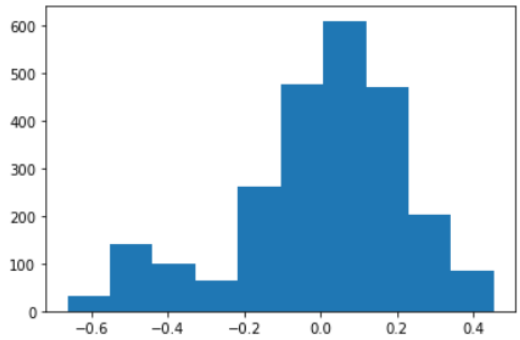
Figure 1.2 OLS Regression of Amazon company for period 2009-2018. Considering its cumulative sentiment as the independent variable and the cumulative performance as the dependent variable

OLS Regression Results

Dep. Variable:	amazon_perf_cum	**R-squared:**	0.944
Model:	OLS	**Adj. R-squared:**	0.944
Method:	Least Squares	**F-statistic:**	2658.
Date:	Mon, 20 Jun 2022	**Prob (F-statistic):**	0.00
Time:	21:53:16	**Log-Likelihood:**	223.32
No. Observations:	2445	**AIC:**	-442.6
Df Residuals:	2443	**BIC:**	-431.0
Df Model:	1		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
amazon_pol_cum	0.0042	8.24e-05	51.552	0.000	0.004	0.004
constant	1.4206	0.026	55.031	0.000	1.370	1.471

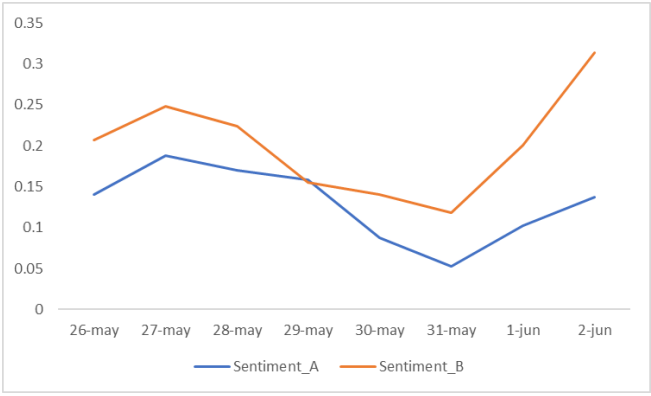
Omnibus:	199.935	**Durbin-Watson:**	0.008
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	249.310
Skew:	-0.766	**Prob(JB):**	7.29e-55
Kurtosis:	3.312	**Cond. No.**	516.



When the time series are standardized, the statistical results remain for the fit model, the normality in the residuals is improved but sometimes not fulfilled and finally the R squared can be dropped to nearly zero, discarding the model again.

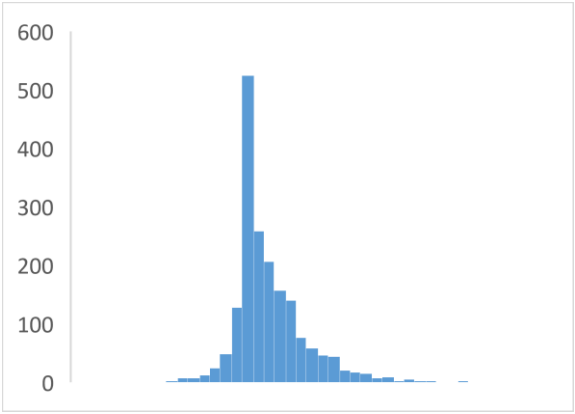
Before feeling fully discouraged from continue reading this study, we present you with little hope. When transforming the qualitative sentiment variable to quantitative using Natural Language Processing, we find some challenges, some were mentioned earlier, and some will be addressed and resolved in this section. Starting with the sentiment measurement of text, there are several libraries or packages used with software that work similarly, and the broad adoption of libraries have generalized the interpretation and measurement of sentiment. For example, in Figure 1.3 we present the average sentiment for the day of the tweets mentioning Facebook ticker *$FB*, the sentiment was calculated using 2 different libraries (TextBlob and Vader) and the results were quite similar, in fact the results were over 60% correlated. Meaning that the results would not vary much if analysis would be performed if one library was used or the other.

Figure 1.3 Average sentiment of the tweets mentioning $FB ticker for the period May 26, 2022, to June 2, 2022 (TextBlob for Sentiment A and Vader for Sentiment B)



Additionally, the distribution of the added daily sentiment of the companies are closely related to the Normal distribution. Allowing adaptations of different methods without manipulating much of the processed data and making easier the interpretation of the results.

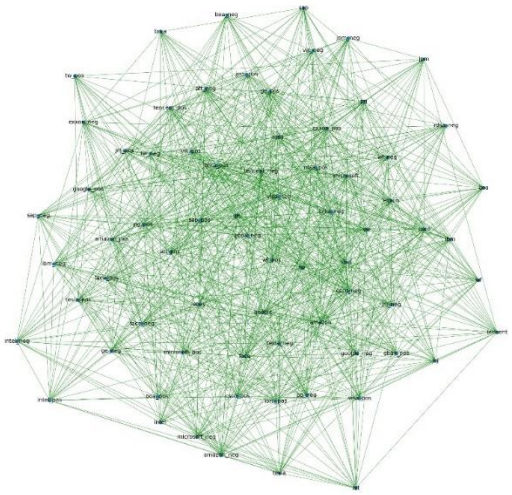
Figure 1.4 Sentiment distribution of Facebook company for the daily added polarity of the tweets mentioning $FB ticker.



So, if we have already stated that traditional statistical methods are not very useful for measuring the relationship of the sentiment in social media with the stock market, What else is left? There are recent models that are designed for situations in which classic models don't work. There is Causality Granger younger brother named Transfer Entropy, developed in 2001 by Tomas Schreiber in which signals are measured using bootstraps simulations and measure the signal flow as bits of information. Working with Transfer Entropy has demonstrated without worrying for symmetry in the behaviour of the variables the effective signal transfer from social media to stock markets and vice versa.

In Figure 1.5 we introduce the graphs chart of the Effective Transfer Entropy of 24 stock daily performance and their corresponding Sentiment Index calculated with the daily addition of the polarity of the tweets mentioning the stock. Even when the correlations are calculated and Granger causality and the results are non-conclusive, transfer entropy captures non asymmetric effects between the time series. It is by this method in which signals that are not captured by the classical methods can be measured.

Figure 1.5 Graphs representation of 24 stocks and their corresponding sentiment indexes, in which there is relationship with the sentiment indexes with more than their corresponding stock.



The rest of the study is structured as follows, in the second chapter we present the simplest form in which we measured signal transfers between social media and the stock price performance using transfer entropy. In the third chapter we present an initial attempt to splitting the signal from social media in negative and positive and measured the marginal effects of the sentiment in the stock price using a Tobit regression. In the fourth and final chapter, we present a standardized framework to normalize both stock performances and sentiment indexes (negative and positive) and measure by both transfer entropy and EGARCH modelling the influence of social media in the stock performance. The same dataset was used in the 3 approaches, in order to be able to compare different results.

The general conclusion of this work is that there is definitely signal flowing form social media to stock performance, and this was demonstrated by different methodologies using contrasting models.

Chapter 2: Measuring signals from social network to stock prices[1]

2.1 Introduction

There is a large strand of the finance literature interested in demonstrating how the news related to a publicly-traded company affect its stock price performance, and how there is a direct correlation between the direction and magnitude of the stock price response and the nature of the published news [1].

Several studies conclude that the news' impact is related to the media coverage of the company [2]. Furthermore, still more studies propose that whether the information is public or private is not relevant and that what matters is that traders have access to it [3]. Eugene Fama stated that the information available about traded companies is fully reflected in the prices, naming this theory the Efficient Market Hypothesis [4].

The Efficient Market Hypothesis (EMH) has been challenged by evidence in diverse studies in different markets and periods of time. Different phenomena not explained by the EMH can be justified from Adaptive Market Hypothesis (AMH) and the Fractal Market Hypothesis (FMH). Studies like Kim, Lim and Shamsuddin [5] for the US data, Shi, Jiang, Zhou [6] for the Chinese market, Árendáš and Chovancová [7] for the very known group of Brazil, Russia, India and China, studied predictability and concluded consistency with the AMH. In the case of the Nigerian stock market, Adaramola and Obisesan [8] found out linear and non-linear predictability and unpredictability periods, i.e., they establish that this market is not efficient and follows the concept of Adaptive Market Hypothesis. For the Vietnamese stock market, D Phan Tran Trung et. al [9] founded that the empirical evidence supports the AMH. A similar result is discovered for the Dhaka Stock Exchange studying seasonal anomalies of the market in Akhter and Yong [10]. For the case of investment, we can study the market for example in periods of turbulence, see for example Moradi et. al [11] where the FMH is confirmed for the Tehran stock exchange and not confirmed for the London stock exchange. In Kristoufek L. [12] the FMH is accepted because of the short term investments dominance over long term investments in the case of financial turmoil. This study is developed for developed markets NASDAQ, FTSE 100, DAX,CAC,HIS and NIKKEI. In the same line Dar et. al [13] studied a long period of time (since mid-eighties) including important events like Black Monday (1987), subprime crisis (2008) and dotcom (2000) and again as in Kristoufek L. [12] the FMH is confirmed.

Farag and Cressy [14] studied how price limits, aimed to prevent speculation amongst traders when new information is released in the market. In their study, Farag and Cressy, extend Fama's EMH in two trends. The first one (Mixture of Distribution Hypothesis; MDH), assumes that the information in the market is available to all traders simultaneously and in consequence, the market achieves equilibrium immediately. The second hypothesis (Sequential Information Arrival Hypothesis SIAH), states that the investors receive information sequentially and in a random

[1] Mendoza Urdiales RA, García-Medina A, Nuñez Mora JA (2021) Measuring information flux between social media and stock prices with Transfer Entropy. PLoS ONE 16(9): e0257686. https://doi.org/10.1371/journal.pone.0257686

manner and the market is adjusted by the traders according to when the information is received over time. Farag and Cressy concluded that there is inefficiency in the information dispersion and that volatility of the stock market is increased instead of reduced when price limits are regulating the market.

There are studies that compare trading regulations between countries [15] and present the differences between laws and penalties regarding similar crimes in different countries. These studies trace back the regulation of securities markets as far as the beginning of the century. Other studies [16, 17], remark gaps in securities regulation considering that interconnected institutions are affected by the same issues, even more perceived during international financial crises. And how regulators are making an effort to address the differences between countries and standardize them, especially ones considering market transparency and data consolidation.

There are some documented cases, in which a negative false announcement referred to a given company affected its stock prices heavily during a short period of time, but when the news was revealed to be fake, the stock price did not recover entirely. An example happened in September 2008, when United Airlines stock dropped more than 75% due to a six-year-old article that resurfaced on the Internet about the 2002 bankruptcy of the United's parent company, mistakenly believed to be reporting a new bankruptcy filing. However, once the news was cleared, the stock price still ended 11.2% below its prior valuation [18]. The United Airlines case illustrates the effect of high volatility after a news release called "drift". The drift is usually present after the news release and its amplitude depends on the nature of the news. Evidence suggests that companies with negative news releases have longer drifts than companies with positive news [19].

Numerous publications aim to build models to predict stock prices, considering the traditional models have not been fully successful in doing so. In contemporary markets, stockholders' opinions are considered faithful indicators of the future value of their investment holding [20]. With the common use of social networks, the opinion of the stockholders has been more present than ever before. Social networks flux of opinions, in combination with the traditional prediction models, have improved the rate of success of prediction methods significantly.

There are several published studies that report new models to predict stock prices using mostly social media opinions [21]. While some studies achieve some degree of prediction capability [22, 23] some others conclude that social sentiment is not useful for stock price prediction [24, 25]. Sentiment analysis of social media has also been used to study the effect of news releases on the price of cryptocurrencies [26]. It can even be stated that cryptocurrency prices are more susceptible to volatility because this category of assets has not yet gained the complete trust of investors.

There is a study that aims to measure how publicly available macroeconomic news influence stock returns by applying Auto-Regressive Vectors (VAR) [27]. The study concludes that there definitely are market movements that coincide with major economic events. The author states that there is also information not considered in the study that affects the stock price performance. Another work, while using Ordinary Least Squares (OLS) regressions, present how the percentage of negative words of news mentioning the situation of a particular firm has a direct impact on the stock return [28]. The study states that the negative words in firms' news precede

low earnings and that in consequence a delayed impact on the stock price. The volume of information analysed varies greatly between both studies, largely to the technology available at the time of each research, and thus the software and methodology applied to analyse the information. Both authors mention this issue, the first study, released in 1988, remarks that in their analysis they failed to consider more information sources to correctly account for the price volatility whether in the second study the authors emphasize the importance of considering a complete set of events is crucial in identifying patterns in firm responses and market reactions to the events. A different study [2], presents a linear regression model to describe the impact of information flux from news on trading activity in Nokia stock price. The results present a dependency on both volatility of the stock price and news regarding the company. The authors remark the relationship of news sentiment in Nokia stock returns.

In this study we compare how the information flux has been evolving through time. Since social media phenomena are recent, it was not considered in early studies regarding the news impact on stock prices [27]. Two conclusions can be drawn comparing our model with early work; first, there is definitely an information flux between the stock market and news media. Second, depending on the window of time analysed, the information signal could be still in the stock market or in the news media, meaning that there is a delay between the signal emission, the reception in the stock price, and when the signal can be measured in its largest magnitude. A major concern here is understanding how the signal flux is structured. Tetlock [28] states in his research paper that investors mostly get their information second hand. Investors initially put more attention on the news/social media than directly in the company's reports or activities. We can infer that there is an initial signal sent from the stock market towards social media yet there is also influence from the news (directly and indirectly) regarding the firms' activity in the stock market in social media and, in consequence there is an impact on the stock market.

There has been well established that news mentioning public traded companies are considered a factor in stock returns performance [29]. But the question that current academia is trying to solve is, How can we translate this impact in a quantitative method and in a statistically measurable fashion that can be scientifically replicated?

Transfer Entropy (TE) is a recent method that measures the statistical coherence between systems through time. This method proposed by Thomas Schreiber [30] considers the exclusion of information that could affect the experiment results by irrelevant characteristics such as common history or input signals. This method usually applied for natural sciences has been lately applied in finance to understand better the causal relationships of exogenous variables in public traded companies [28, 31].

In this study, Transfer Entropy and other advanced computational techniques, web-scraping, and text mining with Natural Language Processing (NLP) to build an index that measures the information flux from social media to prices and vice versa. In this way our Ex-ante expectations, which are proved to be consistent with the existing literature on the subject, are that there will be a positive information flux from social media to stock prices.

We analyse the general public opinion on Twitter (www.Twitter.com) and its impact in the behaviour of publicly traded companies' stocks has been analysed. The simulations show evidence that the information flux exists from public opinion towards the stock market. By

combining two computing methods, web scraping and Natural Language Processing, and by constructing a time series we propose a method for measuring the impact of the news in the stock market behaviour giving room for predicting stock price returns monitoring the news of public traded companies.

The article is structured as follows. In the first section the variables used, and the construction of the data indexes are declared. Section two explains the methodology to structure and pre-process the information. Section three presents the main results and compares them with existing methodology in prior literature. The fourth section presents the discussion under different perspectives. Finally, the last section states the conclusions and proposes further work.

2.2 Materials and Methods

For the construction of the sets, two main data sources were considered. The daily closing prices for the largest publicly traded companies in the world and operating in multiple markets (Nasdaq, NYSE, BCBA, BMV and OTC for the case of Tencent). This condition was followed under the premise that the larger the company the larger the public opinion information that would be available in social media. The second data set was obtained with help of a web scraping software that selected specific mentions in a determined period of time and language. The ticker for each company was searched for the tweets in the English language for the same period as the stock prices time series, from period 2013–2018. The filter selected for reading the information during the search was the top tweets. Each company index was constructed individually, and the data sample obtained from Twitter varied widely, the company with the smallest sample of tweets was Royal Dutch (ticker $RDSA), with 459 mentions and a market capitalization of 227.61 billion U.S.D. as of August 23, 2019. The largest sample retrieved corresponded to Microsoft, with a current market capitalization of 1.018 trillion U.S.D. The full list of companies analysed in this study can be found in Table 1

2.2.1 Methodology

In Figure 2.1 the three-phase framework applied for our model is presented. In which two A.I. robots were used. Robot 1 used for automated web scraping and text mining, Robot 2 was used for Natural Language Processing in which the tweets were filtered, prepossessed and the polarity was calculated. The analysis was performed after the data was analysed and structured, pairing the index with the corresponding daily closing stock price.

Figure 2.1. Framework model in which we present the 3 processes developed in order to obtain, structure and analyse the information.

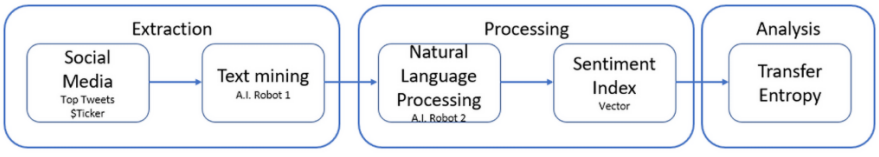


Table 2.1. Characteristics of the analyzed companies

Companies Profile		
Company	Country of Origin	Ticker
Amazon	U.S.A.	$AMZN
Facebook	U.S.A.	$FB
J.P. Morgan	U.S.A.	$JPM
Tesla	U.S.A.	$TSLA
IBM	U.S.A.	$IBM
Berkshire	U.SA.	$BRKA
Exxon	U.S.A.	$XOM
Visa	U.S.A.	$V
Wells Fargo	U.S.A.	$WFCC
Royal Dutch	U.K. Dutch	$RDSA
Ten Cent	China	$TCEHY
Volkswagen	Germany	$VW
AT&T	U.S.A.	$ATT
Intel	U.S.A.	$INTEL
Johnson & Johnson	U.S.A.	$JNJ
General Electric	U.S.A.	$GE
SAP	Germany	$SAP
Microsoft	U.S.A.	$MSFT
eBay	U.S.A.	$EBAY
Google	U.S.A.	$GOOG
Bank of America	U.S.A.	$BOA
Procter & Gamble	U.S.A.	$PG
Cisco	U.S.A.	$CSCO

2.2.2 Phase 1 Extraction with text mining

The technique used for retrieving the comments for the tickers of each company is known as web scraping. Web scraping is a data mining method in which information is retrieved from selected web pages to create large pools of information that may then be analysed to find new patterns [32]. It is considered our first step for analysing information.

For the text mining step, the social network was scanned with a JSON format written in Java, it is from open source and can be found online (https://webscraper.io/). The advantage of this robot (Robot 1) is that the information is extracted and structured in two columns, date and tweet. The text mining technique considered the following variables:

1. (a). Date: The period of time that was selected covered from February 1st 2013 until December 21st of 2018, which intended to cover most part of a full global economic cycle. The aftermath of the 2008 crisis until the preview of the economic deceleration of 2019.
2. (b). Language: The English was language selected for analysing the information since it is the language chosen for business and most of the stock exchange is done in the US.
3. (c). Keywords: The only word that needed to be mentioned in each tweet was the company ticker(abbreviation used for trading preceded with the $ symbol)
4. (d). Top tweets: The page allows the search engine to classify the results in top tweets, a sample of 1% of the most commented and shared tweets.

The criteria were applied for the 23 companies, were each company was mined individually, meaning that there could possibly exist tweets that mention two or more companies. In that case, if Robot 1 detected the same tweet with two different tickers, this opinion was used in both

Sentiment Indexes. In the final part, each unprocessed database was ordered chronologically. This gave room to compare data sizes and mention frequency since some analysed companies began operating on stock market well after 2009 (Facebook IPO was in 2012).

2.2.3 Phase 2 Processing

Natural Language Processing was applied for calculating the polarity in our unprocessed data vector for each comment extracted. For us to be able to construct our polarity vector each tweet was analysed individually, and the vector constructed posthumously.

Python was the language in which a Natural Language Processing algorithm was coded, and the specific library was TextBlob [33]. This library calculates Polarity by breaking each text analysed individually into the words that compose the text. Single-letter words are ignored and for the rest of the text, a numeric value is given for each word that is already assigned inside the library, a value for polarity, subjectivity and intensity. When composed expressions are used (i.e. 'very great') the library considers interpretation rules that improve the analysis for structured sentences. While the total polarity is calculated with the simple addition of each individual polarity, the rules follow some structure, some of the rules are explained further:

- One letter words are ignored
- Negation words add the negative sign to the posterior word
- Multipliers are words that emphasize the meaning of the next word.

In addition to the existing algorithm, the software was improved for the cleaning of each tweet phrase. The technique was changing abbreviations to the full extent of the words (i.e. 'ive', to 'I have', 'im' to 'I am', etc); this step was essential since the abbreviation of words is very common on Twitter given the limited space for each tweet (280 characters). By performing the additional cleaning for each tweet, the margin error for results with no value calculated was reduced considerably.

As an example, a real tweet from January 26, 2014:

```
When you have to set up an "ethics board" you know you don't have any.
Maybe time for @google to stop saying "don't be evil" $goog.
```

We can understand that the user is stating that the company Google lacks ethics in its directive board since it is setting up one and that the connotation of the sentence is not positive. The issue at hand is how to help the Robot to interpret the negativity as such. The first step is to clean the tweet from characters other than letters and from abbreviations. The algorithm returned the clean sentence:

```
when you have to set up an ethics board you know you do not have any.
maybe time for google to stop saying do not be evil goog.
```

We can observe that all the characters that were not letters disappeared and the abbreviation *don't* extended to *do not*. The word *goog* since is not in the English dictionary was not cleaned with the library, unless you add it to its dictionary. For this particular case we let the word untreated.

In the second step the algorithm breaks the tweet into sentences and words:

- `[Sentence("When you have to set up an ethics board you know you do not have any."), Sentence("Maybe time for @google to stop saying do not be evil $goog")]`
- `WordList(['When', 'you', 'have', 'to', 'set', 'up', 'an', 'ethics', 'board', 'you', 'know', 'you', 'do', 'not', 'have', 'any', 'Maybe', 'time', 'for', 'google', 'to', 'stop', 'saying', 'do', 'not', 'be', 'evil', 'goog'])`

Finally, the algorithm analyses the polarity and subjectivity for the sentence adding the individual score for each word. The individual score is also pre-recorded in the library. Since the words *google* and *goog* are not in the library, they do not have a polarity value, meaning that the value assigned in the calculations will be zero not affecting the outcome of the analysis. The result for this example was a polarity $P = -1$ meaning that it has the highest negative score available.

It is important to mention that accuracy issues emerge when the words are scrambled (consider that this will definitely give a sentence with no sense) the calculated polarity will be different. For further reading review the source code [33].

2.2.3.1 ***Sentiment Index vector construction.***

Once the data sets have been processed and the polarity P has been calculated for each tweet, the results were added for each observed day, it does not matter if the stock market was operating or not (including weekends and days off). For the day t the Polarity value Y was calculated with the simple addition of the Polarity P of the i tweets in the same day t:

$$Y_t = \sum_{n=1}^{i} P_n \qquad (2.1)$$

$$Y = [Y_t, Y_{t+1}, Y_{t+2}, \dots] \qquad (2.2)$$

The variation of the stock X for the day t, was calculated by subtracting the closing price from the day before (S_{t-1}) to the closing price of the Stock of the day (S_t). In this manner vector X is the stationary values of the prices:

$$X_t = S_t - S_{t-1} \qquad (3.3)$$

$$X = [X_t, X_{t+1}, X_{t+2}, \dots] \qquad (3.4)$$

2.2.4 Phase 3 Transfer Entropy

In this phase, the analysis of TE was applied to measure the flow of information between Sentiment Indexes vectors and stock market returns. The theory behind this methodology is explained further.

Let Y and X denote two discrete random variables with marginal probability distributions $p(x)$ and $p(y)$ with joint probability distributions $p(x, y)$, whose dynamical structures

correspond to stationary Markov processes of order k and l for systems Y and X respectively. The Markov property considers the probability to observe X at the time $t + 1$ in state i conditional on the k previous observations is

$$p(x_{t+1}|x_t, \dots, x_{t-k+1}) = p(x_{t+1}|x_t, \dots, x_{t-k}), x_i \in X \quad (3.5)$$

Having $p(A|B)$ representing the conditional probability of A given B, $p(A|B) = p(A, B)/p(B)$, the TE from Y to X can be defined as the average information included in Y excluding the information reflected by the past state of X for the next state information X. Hence, TE measure is defined as [30]

$$T_{Y\to X}(k,l) = \sum_{i,j} p\left(x_{t+1}, x_t^{(k)}, y_t^{(l)}\right) log \frac{p\left(x_{t+1}, x_t^{(k)}, y_t^{(l)}\right)}{p\left(x_{t+1}, x_t^{(k)}\right)}, \quad (3.6)$$

where $xt+1$ of X is affected by k previous states of X, in other words, the lagged values affecting the current value of X. In addition, X is affected by l previous states of Y, in other words, the lagged values affecting the current value of Y.

TE attempts to incorporate time dependence into account by relating previous observations xi and yi in order to predict the next value $xi+1$. Then, it quantifies the deviation from the generalized Markov property, $p(xi+1|xi, yi) = p(xi+1|xi)$, where p denotes the transition probability density to state $xi+1$ given xi and even yi. If there is no deviation from the generalized Markov property, Y has no influence on X. Then, TE quantifies the incorrectness of this assumption and is formulated as the Kullback-Leibler entropy between $p(xi+ 1|xi, yi)$ and $p(xi+1|xi)$ is explicitly nonsymmetric with respect to the exchange of xi and yi.

An important property of TE is that under specific conditions it can be formulated as a nonlinear generalization of Granger causality. The last quantity plays an important role in the parameter estimation of a vector autoregressive (VAR) model in econometrics. There exists a series of results [34–36] that state an exact equivalence between the Granger causality and TE statistics for various approaches and assumptions on the data generating processes. It makes it possible to construct TE as a nonparametric approach of the Granger causality test. This relation can be regarded as a bridge between the information-theoretic approach and the causal inference under autoregressive models. It is important to mention that for highly nonlinear and non-Gaussian data, as is the case for most financial instruments, it is more adequate to model causality by using the TE information method instead of the traditional Granger causality test [37, 38].

On the other hand, the transfer entropy measure in Eq 6 is derived for discrete data. However, in any economic application the observed time series are continuous. There exist several techniques for estimating TE from observed data in order to apply it to real-world data problems. However, most of them require a large amount of data, and consequently, their results are commonly biased due to small-sample effects, which limits the use of TE in practical data. A straightforward approach to estimate TE is to partition the data into discretized values, for this particular case the data sampling was structured in. Thus, a time series $x(t)$ is partitioned to obtain the symbolically encoded sequence $S(t)$. This sequence replaces the value in the observed time

series by the discrete states {1, 2, …, *n* − 1, *n*}. The model allows to discretize continue data by partitioning the data by choosing the quantiles of the empirical distribution of the time series. By denoting the bounds of the pre-selected number of bins by *q*1, *q*2, *q*3,…*qn*, where *q*1 < *q*2 < *q*3, … < *qn*. By performing this data partitioning, each value in the original time series is replaced by an integer.

Moreover, the expression of *TE* (Eq 6) is likely to be biased due to several factors as finite sample effects and the not strictly stationarity of financial data. Also, time series with higher entropy naturally transfer more entropy to the others. To reduce this bias, the Effective Transfer Entropy (ETE) [39] has been proposed where

$$ETE_{J\to I}^{shuffled}(k,l) : T_{J\to I}(k,l) - T_{J\ shuffled\to I}(k,l), \qquad (3.7)$$

Where $T_{JShuffled\to I}$ indicates the transfer entropy from *J* to *I* with randomly shuffled time series *J*. Thus, all statistical dependencies between the *two-time series are destroyed. An important characteristic is that $T_{JShuffled\to I}(k,l)$ converges to zero at a long sample size. Consequently, any non-zero value of $T_{JShuffled\to I}(k,l)$ is due to small sample effects.

Later on, the work of Dimpfl et. al. [40] improves the bias correction by adding an inferences perspective of the estimated information flows. They proposed to use the Horowitz approach [41], which bootstrap the modelled Markov process. The idea is to simulate process *J* based on the calculated transition probabilities, where the dependencies between *J* and *I* are destroyed, but the dynamics of the series *J* is not changed. Transfer entropy is then estimated using the simulated time series. Then, this procedure is repeated several times to create a null distribution of no information flow, which can be used to test for statistical significance. The proposed equation has the same structure as ETE [40]:

$$ETE_{I\to I}^{boot}(k,l) : T_{J\to I}(k,l) - T_{J\ boot\to I}(k,l) \qquad (3.8)$$

where $T_{boot\to I}$ indicates the average over the estimates derived from the null bootstrap distribution.

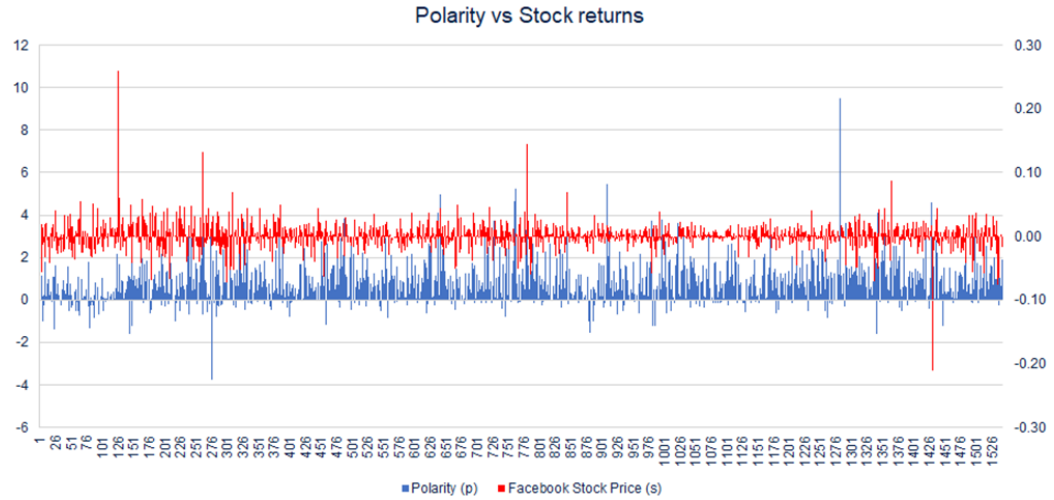
The null hypothesis p-value that measures if there is no information exchange is given by$1-\hat{q}_{TE}$, in which$\hat{q}_{TE}$ denotes the quantile of the simulated distribution that corresponds to the transfer entropy estimations when the dependencies from *I* to *J* are destroyed.

2.3 Results

Evidence of the influence of collective behavior on stock price returns has been presented. This study provides an effective way to measure the relationship between social media and stock returns. Even so, it provides a potential way to measure the impact of social media in stock price performance. Taking the data sets one step further, we performed statistical analysis of the combination of all the vectors by the three methods (Pearson correlation, Granger causality and Effective Shannon Transfer Entropy).

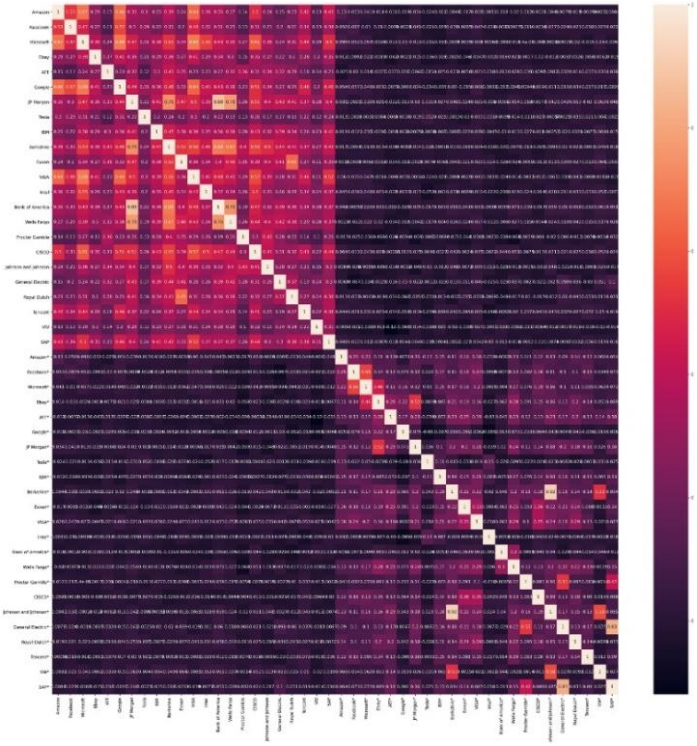
In Figure 2.2 we compare the behavior of the stock price returns and Sentiment Index of Facebook (ticker $FB). It can be observed that the activity in both vectors increases in the same periods of time. Short after its IPO in 2013, the stock movements and the activity in the Sentiment Index increases. The other period that can be observed with high movement is during 2018, where the company suffered largely negative publicity due to the Cambridge Analytica scandal.

Figure 2.2. Time series of daily returns of Facebook vs Facebooks' Sentiment Index. It can visually be appreciated the synchronized upward or downward movements between the Sentiment Index and the company performance.



The information flux between two data sets is usually measured with the correlation factor, which can be described as the statistical relationship between two variables. The most known method is the Pearson correlation coefficient which measures the linear sensitivity between two variables [42]. In financial analysis is not only important to know the level of the relationship between two variables, but the causal dependence between them. In Figure 2.3 we have the linear correlation matrix of the companies daily returns and the Sentiment Indexes(marked with * for each company) constructed for this research. The results show a clear positive correlation between companies' daily returns and the indexes. Counter-intuitively, the matrix shows a little or no correlation between the companies and the indexes.

Figure 2.3. Correlation matrix of the companies and the Sentiment Indexes, it can be appreciated that the companies are considerably correlated and the same for the Sentiment Indexes, yet there is little to no correlation between stocks and Sentiment Indexes.



In order to measure the signal transfer (intensity and direction) between the Sentiment Index and the stock price individually, Shannon Transfer Entropy [43] was applied. The results support the theory of the effect of news media in asymmetric stock returns and the theory that the stock market performance is influenced by collective behavior [44].

To analyze the information flux, we assigned the X variable to the Stock Prices vector and the Y variable to the Sentiment Index vector and the intensity of the signal was calculated:

$$Intensity = ETE(Y \rightarrow X)/[ETE(Y \rightarrow X) + ETE(X \rightarrow Y)]$$

With this expression, if the information signal was greater from ($Y \rightarrow X$), it would take a positive sign, meaning that the information flows from social media to the stock market, which is the main interest of our study. The contrary sign could be expected if the signal from ($X \rightarrow Y$) was greater, meaning that the stock market is affecting the activity of social media. This was performed exclusively to differentiate the direction of the signal in Figure 2.5.

We used the implementation RTransferEntropy to estimate the $ETE_{J\rightarrow I}^{boot}(k,l)$ considering the configuration $J = X$, $I = Y$, $k = 3$, $l = 3$, and a bootstrap simulation of 300 shuffles. The maximum p value considered was ($p \leq 0.10$) The quantiles selected to discretize the time series were c(5, 95), meaning that the 5% and 95% quantiles of the empirical distribution are used.

TE quantifies the information provided by the past of the process X influencing the present of the process Y, that is not already provided by the past of Y. With this implication, in this work we to quantify the information provided by the past of Y on the shifted portion of the system X.

Simulations were performed considering lags = 1,2 and 3 for both variables X and Y, obtaining better results with k = l = 3. With these results, our work supports the theory that the sentiment signal is observable with more clarity in the stock price in a 3-day lag, considering that we are working with daily observations. This would mean that today's polarity will have its maximum effect on the stock price 3 days in the future.

For the construction of the X vector(Eq 4), simple differentiation was considered, in order to work with the stationary expression of the stock price, as an alternate approach since most literature focus on working with the stock performance. The X vector (Eq 4) was tested for unit root using two methods, Dickey-Fuller and Phillips-Perron. In every test, the results concluded that the X vector was stationary. The same process was performed in the Y vector (Eq 1) obtaining the same results. This to ensure that the model was executed with stationary vectors.

The first results are presented in Table 2.2. We can observe that there is a clear effect of the stock price performance given the Sentiment Index in TE simulations (p value = < 0.10) towards the stock price ($Y \rightarrow$ X). An important remark is that these twelve companies are from different industries and capitalize in different stock markets. We have Tech companies, Oil and Gas, Banking, Software and Automobile industries.

The Effective Transfer Entropy was calculated as well and presented in the Tables 2–4. In Table 2.2, the information flux is 67% stronger in the direction of the stock market ($Y \rightarrow X$). The strongest signals where for Tesla (Ticker $TSLA) and Exxon (Ticker $XOM) with an 80% information flux. As in Table 2.3, the information flux is 60% stronger in the direction of the Sentiment Index($X \rightarrow Y$) in average. The strongest signals are AT&T and SAP, both companies with 100% of the signal flux towards the Sentiment Index. Finally, the last six companies tested back with no statistical conclusive results for Shannon TE. Meaning, that the p value is greater than 0.10 in any direction. Results are presented in Table 2.4.

Table 2.2. Information flux from Y -> X TE

Shannon Transfer Entropy Results:							
X -> Y TE				Y->X TE			
Company	**TE**	**ETE**	**p-value**	**TE**	**ETE**	**p-value**	**Intensity**
Amazon	0.0369	0.0161	0	0.0503	0.0211	0	0.5672
Facebook	0.0232	0.0034	0.1967	0.0295	0.0036	0.1	0.5142
JP Morgan	0.0248	0.0054	0.19	0.0454	0.0209	0	0.7946
Tesla	0.0268	0.0047	0.1367	0.0411	0.02	0.0033	0.8097
IBM	0.0209	0.0027	0.1733	0.0284	0.0078	0.06	0.7428
Berkshire	0.0251	0.0036	0.1033	0.0342	0.008	0.0367	0.6896
Exxon	0.0249	0.0047	0.13	0.0437	0.0191	0	0.8025
Visa	0.0458	0.012	0.04	0.0562	0.0186	0	0.6078
Wells Fargo	0.0284	0.0075	0.07	0.0308	0.0087	0.0067	0.537
Royal Dutch	0.0145	0.0017	0.21	0.0231	0.0064	0.06	0.7901
Ten cent	0.0366	0.0135	0.02	0.0479	0.0182	0.0033	0.5741
Volkswagen	0.0266	0.0031	0.19	0.0273	0.007	0.0533	0.693

Table 2.3. Information flux from X ->Y TE

Shannon Transfer Entropy Results:							
X -> Y TE				Y->X TE			
Company	**TE**	**ETE**	**p-value**	**TE**	**ETE**	**p-value**	**Intensity**
At&t	0.0343	0.0095	0.0633	0.021	0	0.3433	-1

Intel	0.0487	0.0239	0	0.0323	0.0071	0.0467	-0.229
Johnson & Johnson	0.0291	0.0062	0.0633	0.0261	0.0031	0.1233	-0.3333
General Electric	0.0331	0.0106	0.0067	0.0317	0.0086	0.0033	-0.4479
SAP	0.0265	0.0058	0.0567	0.0197	0	0.5967	-1

Table 2.4. Nonexistent information flux.

Shannon Transfer Entropy Results:							
X -> Y TE				**Y->X TE**			
Company	**TE**	**ETE**	**p-value**	**TE**	**ETE**	**p-value**	**Intensity**
Microsoft	0.0169	0	0.63	0.0188	0	0.7933	NA
eBay	0.0183	0	0.5933	0.0241	0.0027	0.1467	NA
Google	0.0187	0	0.78	0.0209	0	0.6067	NA
Bank of America	0.0231	0.0045	0.24	0.0277	0.004	0.11	NA
Procter & Gamble	0.0297	0.0031	0.22	0.0212	0	0.45	NA
Cisco	0.0213	0	0.4567	0.0297	0.0037	0.1367	NA

The most widely known method for measuring information flux between time series is Granger Causality [45]. This method has its limitations towards confirming information flow, compared with Transfer Entropy (TE) [30], which considers mutual information and dynamics transports. Schreiber concludes in his paper that TE is capable of detecting the direction and intensity of signal exchange between two systems whilst ignoring static correlations. A virtue of the TE method is that it allows quantifying information transfer without being bounded by linear dynamics [30].

In Figure 2.4 the results of Granger Causality tests are presented. In Figure 2.5 we present the results of the Shannon ETE for the combination of all of our variables, combining stock price returns and Sentiment Indexes. Both results are very similar with a few minor variations. For example, it is visible in Figure 2.4 some variables with high communication such as Ebay* towards VW and BRKA* receiving signals from MSFT*, EBAY*, JnJ* and VW*. Comparing the same variables results in Figure 2.5 it is restated that for ETE, EBAY* sends a signal to VW, while BRKA* sends signals to MSFT*, JnJ* and WV.

Figure 2.4. Granger Causality relationship with a lag of 3 and p-value $p \leq 0.10$

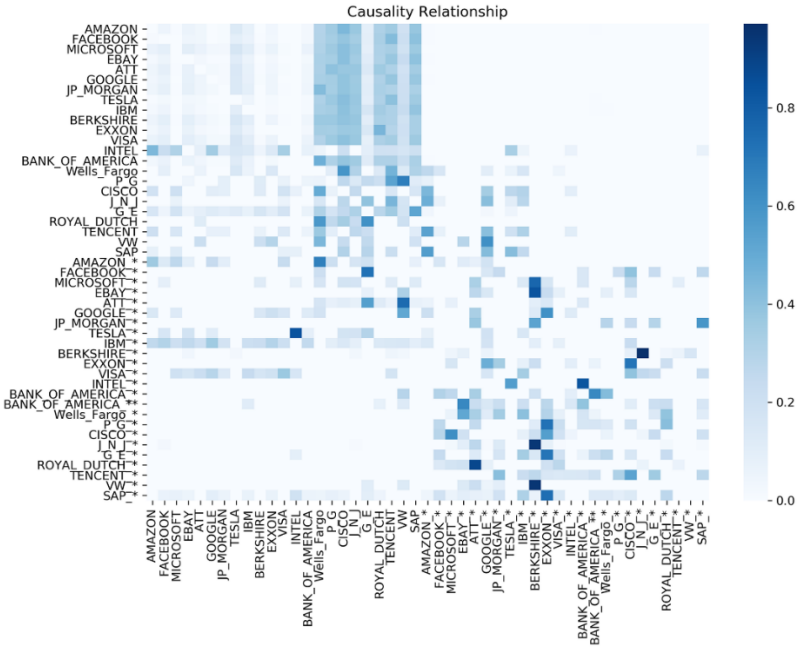
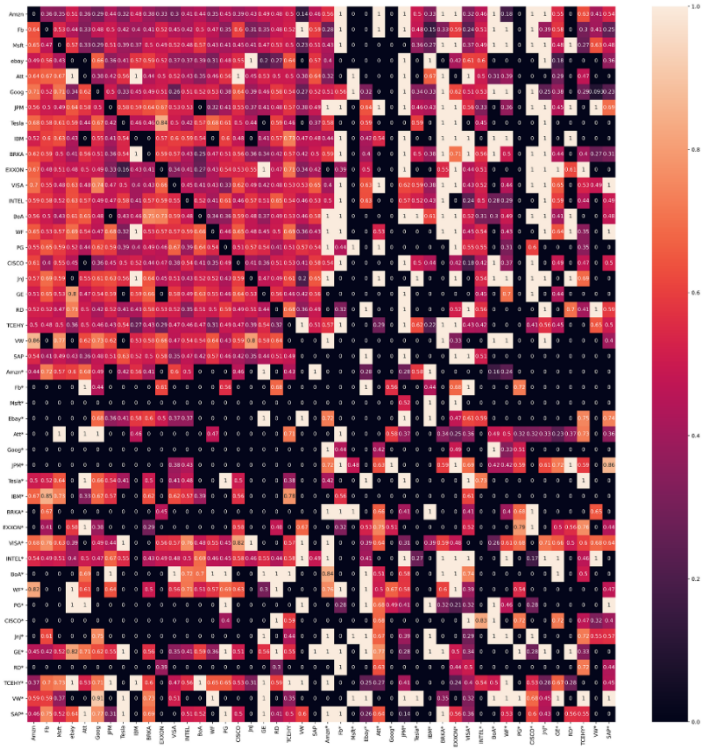


Figure 2.5. Shannon ETE matrix with a lag of 3 and p-value p≤ 0.10. Contrary to the correlation matrix, there is information transfer between stock companies and the Sentiment Index.



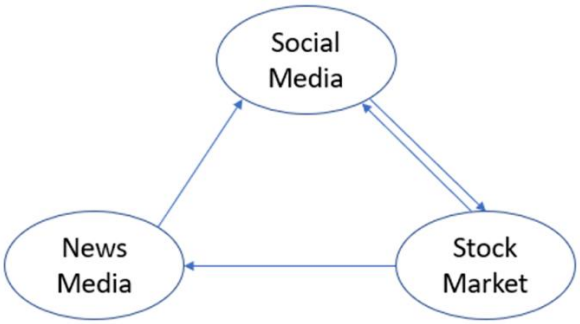
Comparing ETE Figure 2.5 with Correlation matrix Figure 2.3 the main differences are noticeable, the increased signal activity from the stock market to the indexes, and the signal directions. From stock price to stock price, we can observe the communication between our variables very similar to the correlation matrix, yet in the ETE column, the direction of the signal is conspicuous. Being the main difference in the Stock market/Stock indexes quadrants, where the correlation is virtually nonexistent, the ETE signal is very present.

2.4 Discussion

Some studies present evidence of increased drift in stock behaviour after news release of a particular company in a short period of time [46], and that the effect is greater when negative news is released. The information flux measurement between news and financial performance of public traded entities has been proposed [47]. Compared with other computational methods, Transfer Entropy has demonstrated better results in measuring the influence of news in stock returns behavior.

In order to understand the information flux between stock markets and general public opinion, a theoretical framework is presented in Figure 2.6. In this framework, it is stated that the stock market sends signals to both news media and social networks. In addition, the social network also receives information from news media, sending signals to the stock market in consequence. In Table 2.2 it is visible how most companies only receive information from social media, while Amazon, Visa, Wells Fargo and Ten Cent receive and send signals to social media. It has been acknowledged throughout the presented investigation that tweets are receiving the information from News Media since some of them quote the source from news companies and others are just retweeting the information from verified news media accounts.

Figure 2.6. Theoretical information flux in which the results from Table 2 demonstrate information flow from social media towards stock price and in some cases information flow from stock market towards social media.



The presented theoretical framework is similar to the Tetlock statement, in which investors obtain their information from secondary data sources. An expansion to Tetlock's model is proposed by considering the signal feedback from investors and general opinion to stock market performance.

It can be stated from the three analyzed methods, that the stocks are highly dependent on their Sentiment Indexes and in some cases from Sentiment Indexes of other companies, meaning that the ecosystem in which the companies operate, even when some companies are in different markets they are affected and affect each other.

The results of this study compared to the results from Bollen et al. [23] differ mainly in the improved text processing algorithm. The second difference is that our study extends the analysis to 23 different companies, some of them that operate in non-English speaking countries with positive results.

This model obtained positive results considering that our sample covered 6 years with a volume of 200,000 tweets for 23 companies, contrasted to a previous study [48] that used a

sample of 60,000 tweets for a 6 day period of time. In which authors concluded that the sample size prevented their study from acquiring statistically significant results using correlation and regression tests. While both studies conclude that correlations between stock prices and sentiment indexes are nearly non-existent, Transfer Entropy demonstrates statistical significance between Social Media behavior and Stock price reaction.

2.5 Conclusion

In this paper the relationship between stock price performances and social sentiment in public social networks was reviewed, i.e. there is enough evidence to support that what investors and observers of stock markets affect the performance of it independently of the level of participation or location of both the market and the participants issuing the opinion. The database covers a conveniently long period of time (2013–2018) after the 2008 crisis until before the economic slowdown of 2019. The opinions data set was considerably large (over 200,000 tweets covering the same period of time).

By proposing an original simplified Sentiment Index using technological tools as web scraping and machine learning for NLP we evaluate the relationship between variables that apparently are not compatible firsthand. Our simulations were performed for each company and corresponding Sentiment Index. In addition, We have used different statistical methods to measure the performance of our index construction model, in which the results proved to be an improved model over existing methods. It is important to mention that throughout the literature reviewed, it is noticeable that past studies have been consistent in comparing the sentiment analysis versus a single stock. However, they haven't studied the indirect bounce of data among a similar range of stocks. As it is to be proven, there is undoubtedly communication between multiple Sentiment Indexes affecting multiple companies; meaning that the comments regarding a particular company affect indirectly the performance of several others.

The asymmetric theory of information states that negative news has a greater impact on stock prices than positive news. In this study, we present visual evidence to raise the question, expand research aimed to support this theory. Future work can be focused on creating an independent index for both positive and negative news/opinions and measure the marginal effects of each index in the stock price. There is also ground for replicating the model, measuring the extent of the network effects with other exogenous variables in the stock market.

2.6 Further work

In this initial study we have presented a generalized framework to measure the signal from social media sentiment to stock price for daily observations. A natural trend for stock trading strategies is to shift towards high frequency as technology evolves and becomes more available to the average person. The next step following this research, is to adapt the framework for intraday data. A natural issue to solve is to obtain the social media data in real time and the ticks of the stock, which would require high computational capabilities.

Chapter 3: Splitting the signal[2]

3.1 Introduction

There is a large strand of the finance literature interested in demonstrating how the news related to a publicly traded company affect its stock price performance, and how there is a direct correlation between the direction and magnitude of the stock price response and the nature of the published news[1]–[5]. Several studies conclude that the news' impact is related to the media coverage of the company. Furthermore, still more studies propose that if the information is public or private is not relevant, what matter is that the traders have access to it [6].

Some documented cases show how a negative false announcement referred to a given company affected its stock prices heavily during a short period of time, but when the news was revealed to be fake, the stock price recovered some of its value. A first example, on June 14th, 2021, Cristiano Ronaldo stated "Agua, no Coca" in a press conference, this declaration was heavily criticized since that day Coca-Cola stock dropped 1.06% compared to the previous closing price. Analyzing the time frame of the stock behavior we can appreciate that the declaration per se did not affect the performance of the company. Cristiano Ronaldo did the declaration at 9:43 EST, when the stocks had dropped to $55.26 by 9:40 EST, 3 minutes before Ronaldo's declaration. Even more, the stock closed $0.30 above the $55.26 by the end of the trading day. What affected the price more were the headlines "Cristiano Ronaldo removes coke bottles and Coca-Cola stock prices drop" by CNN Spain on June 16th impacting the price on the 18th and 23rd of the same month.

In the other hand we have the exact opposite effect, during the pandemic events in 2020, the work methodology evolved to a virtual presence using communication tools as Skype, Zoom or Google Meets. Being Zoom platform the most common for work and school it became clear that the extended use would soon impact positively in its stock price. But the investors focused in the wrong stock mistaken 'Zoom Technologies' (Ticker $Zoom), a Chinese tech company with 'Zoom'(Ticker $ZM). This derived in a stock price soar of %1800 versus a %132 of the intended company, for the period from February 2020 until April of the same year when the SEC halted the trading of Zoom Technologies.

Ronaldo and Zoom cases illustrate the effect of high volatility after a news released called "drift". This drift is usually present after the news release and its amplitude depends on the nature

[2] Mendoza-Urdiales, R.A.; Núñez-Mora, J.A.; Santillán-Salgado, R.J.; Valencia-Herrera, H. Twitter Sentiment Analysis and Influence on Stock Performance Using Transfer Entropy and EGARCH Methods. *Entropy* **2022**, *24*, 874. https://doi.org/10.3390/e24070874

of the news. Evidence suggests that companies with negative news releases have longer drifts that companies with positive news [7]

Numerous publications aim to build models to predict stock prices, considering the traditional models are not fully successful in doing so. In contemporary markets, stockholders' opinions are considered faithful indicators of the future values of their investment holding. With the common use of social networks, the opinion of the stockholders has been more present than ever before. Social networks flow of opinions, in combination with the traditional prediction models, have improved the rate of success of predictions methodology significantly. There are several published studies that report new models to predict stock prices using mostly social opinions [8]. While some studies achieve some degree of prediction capability [5][6], some others conclude that social sentiment is not useful for stock price prediction[11][12]. Sentiment analysis of social media has also been used to study the effect of news flows on the price of crypto currencies [13]. It may be said that considering that this category of assets has not yet gained the trust of investors, their price is more susceptible to volatility and the correlation between news releases and price behavior is more accentuated. However, all these studies have focused on predicting the movements of the stock market or individual stock prices but have not been focused on determining the magnitude of the stock price effect of bad news compared to good news for specific companies. The present study aims to fill that gap in the literature. Using Artificial Intelligence to build an index that quantifies the influence of positive and negative news on companies is an innovation on the study of the differentiated impact they have on stock prices, so our ex-ante expectations, consistent with the existing literature on the subject, are that the effects of negative news should be larger than the repercussion of positive news. To that end, an Artificial Intelligence method is developed to build an index capable of measuring the impact of positive and negative news on the stock price of a company.

Equity returns are asymmetric when negative returns have larger volatility periods than positive returns. This phenomenon was originally reported by [14] . According to that study, declines in equity values are not matched by declines in the value of debt, so negative returns influence the leverage of the firm's capital structure. However, they conclude that the financial leverage effect is not sufficient to explain the asymmetry in returns.

The response of stock prices to the revelation of economic information has been studied by different authors. [15], for example, analyzed the reaction of the stock price of companies to news for the period 1953-1978, and found that stock prices react slowly and weakly to news about inflation. [16] studied the period between September 1977 and October 1982 to measure expectations based on survey results, used them to justify daily stock price movements, and concluded that stock prices are sensitive to monetary policy announcements, but not to news regarding price indices, unemployment, and industrial production. [17] measured news from 1871 to 1986 using Vector Autoregressions and concluded that at least one third of the monthly returns' variance can be explained by the news.[18] concluded that the stock market not only responds to macroeconomic news, but that the nature of its reaction depends on the current state of the economy. People put more attention to losses of utility than to gains of equal magnitude; this feature is called "loss aversion" and was originally presented in the Prospect Theory literature in [19]. Such behavior is latent at the level of macroeconomic aggregates as consumption usually

falls by a larger percentage during economic recessions, compared to consumption increments during economic growth. The evidence on the asymmetric response to positive and negative information has also been extensively studied in the capital markets. [20]analyzed the news impact and express the asymmetry of good *vs*. bad news in investors' sentiment in online forums and concluded that investors respond in a more conservative manner to positive news during harsh economic market conditions than during economic growth periods.

The present study includes data corresponding to the period that followed the 2008 crisis, and through the whole period of observation there was a positive performance in global GDP. Additionally, most of the companies included are based in the U.S. So, taking the SP500 as a market indicator for state of the U.S. economy, it is possible to conclude that our assumption of positive economic growth during the period analyzed is confirmed.

There have been major advances in recent years developing methodologies to improve accuracy in prediction of events for financial analysis (crises, market crashes, out of the ordinary portfolio returns in one day, relation between environmental changes and a company behavior) but most of the research areas that have been studied separately, when in fact, there may be a logical approach to cross-relate them. There are studies that strongly suggest a direct relationship between social media behavior and market reactions [10], [21]. It is widely known that portfolio managers and financial analysts react regarding on behalf of their portfolios when a thread of news is released, there is a research stream that analyzes the connection between a company Corporate Social Responsibility policy (CSR)and the capability of the company to adapt to internal or environmental changes. This theory supports the existence of a negative correlation between the confidence from investors and the possibility for the firm to be at default risk[22]; this can be interpreted as how investors react positively to a company's CSR policy, reflecting on a company's financial health. If it were to be seen as a risk management tool, and a CSR progress monitor of a company was carried out, certainly, the risk of news that could affect this company stock price could be measured and its impact could be reduced for the best possible outcome on the company's behalf. Different research streams analyzers acknowledge that there are variables which are being underseen that may affect the outcome of the results on experiments since most studies are done in controlled environments under laboratory circumstances. If we modify the experiment environment with variating stress conditions, the behavior of the studied models will most certainly be unknown.

The main advantage of "Big Data" is to analyze large pools of information without the need for classifying facts and risking to unintentionally discard data that might be relevant to further analysis. An additional advantage is that current technology can perform an accurate real-time analysis from these large pools of information, yet most of the existing models are designed to classify data under the researcher criteria.

Another early research ventured to predict stock market indicators such as Dow Jones, NASDAQ, and S&P 500 by analyzing Twitter posts, collecting twitter feeds for six months, and taking a randomized subsample of nearly one hundred of the full volume of all tweets [23]. The researchers measured collective general feelings considering fears and hopes on each day and analyzed the correlation between these indices and the stock market indicators. The research team found a percentage of emotionally negatively twitter posts correlated with Dow Jones,

NASDAQ, and S&P 500 outcomes, and found a significant positive correlation to Chicago Board Options Exchange Market Volatility Index (VIX) indicators. Determining that analyzing twitter for emotional responses of any kind provides a hint of how the stock market will be behaving the following day.

There is another publication in which Twitter was taken as the main source of news due to a clear trend to adopt this platform to disperse news amongst financial traders. In this paper, the researchers applied sentiment analysis algorithms, identified news trends in the web, and compared these trends against financial market movements finding a positive correlation [24].

A particular approach that is not found in many research streams was presented by [25], a large amount of information on Twitter on a short time was analyzed. 60,944 tweets were analyzed aiming to find a correlation between political action in social media and stock market behavior on a 6-day window of time (from May 4th, 2016, to May 9th, 2016) where political campaigns in the UK were being held. [26] analyzed Microsoft news on Twitter not only regarding the company stock, but also the opinions on the products and services of the company. In their paper, the researchers proved that a strong statistical inference exists between rising/falling stock prices of a company accordingly to the public opinions and emotions expressed about that company on Twitter. The main contribution of their work is the development of a sentiment analyzer that can judge the type of sentiment presented in the tweet. The tweets are classified into three categories: positive, negative, and neutral. Initially, they intended to prove that positive emotions or sentiments about a company would reflect in its stock price. However, neutral, and negative tweets also affected the stock price. [27] Applies Random Matrix Theory and to analyze 64,939 news from the perspective of information theory. This study finds a correlation between the flow of the New York Times news and 40 world financial indices for a 10-month period between 2015-2016. The study describes a dynamic movement between the flow of information and stock price behavior. The model was also tested with and without white noise. A delay between the time the news is published, and the reaction of the stock price is reported.

3.2 Materials and Methods

In Figure 3.1 we compare two time series, the stock price of Tesla paired with the sentiment obtained from Twitter. It can be visually appreciated the synchronized movements from both time series.

Figure 3.1. Tesla daily stock price versus twitter sentiment It can be appreciated that the

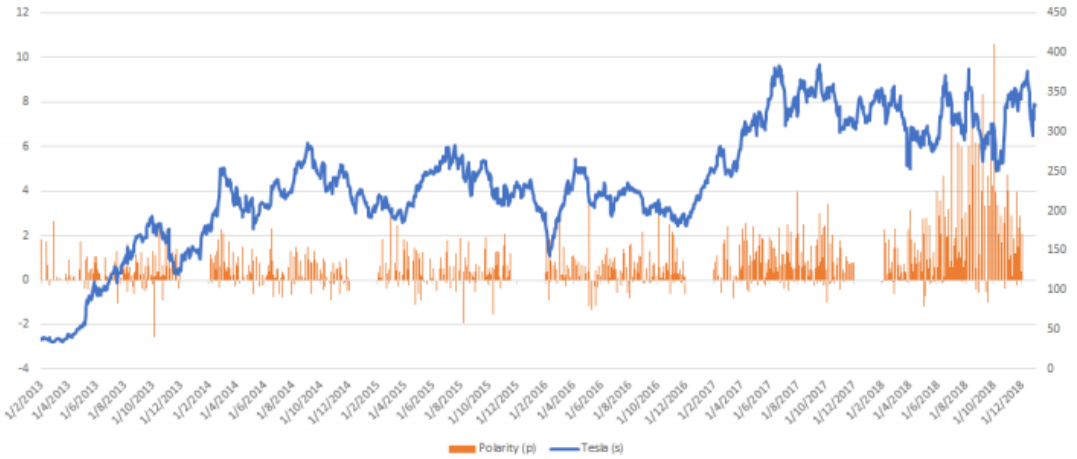
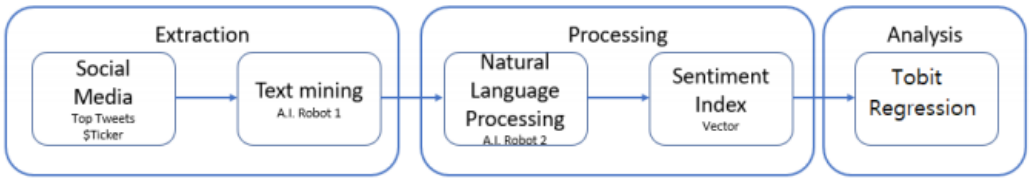


Figure 3.2 we present the framework of our method for the analysis



3.2.1 Text mining

Text mining is a data mining method within the web scraping category that retrieves information from selected web pages to create large pools of data that may be analyzed to discover patterns [14]. This was the first step in the analysis.

Twitter was scanned with a JSON routine written in Java. The advantage of this robot (Robot 1) is that the information is extracted and structured in two columns, date, and tweet. A second text mining technique considered the following variables:

1. Date: The period that was selected covered from January 1st, 2009, until December 1st, 2018. Which covered most of a full global economic cycle, i.e., from the aftermath of the 2008 financial crisis, until the last months before the economic slowdown of 2019.
2. Language: The language selected for analyzing the information is English.
3. Key words: The only word that needed to be mentioned in each tweet was the company ticker (abbreviation used for trading preceded with the $ symbol).
4. Top tweets: The page allows the search engine to classify the results in top tweets a sample of 1% of the most commented and shared tweets.

The criteria were applied for the 23 companies considered in the study and each company was mined individually, meaning that there could exist tweets that mention 2 or more companies, in that case, if Robot 1 detected the same opinion with 2 different tickers, this opinion was used in our study in both Tobit regressions. In the final part, each unprocessed database was ordered

chronologically. This allowed to compare data sizes and mention frequency since some companies were founded and traded well after 2009 (Facebook IPO was in 2012).

3.2.2 Sentiment Analysis

Python was the language in which a Natural Language Processing algorithm was coded, and the library used for said task was *TextBlob*. This library calculates the sentiment by breaking each text analyzed individually into the words that compose the text. Single letter words are ignored and for the rest of the text, a numeric value is given for each word that is already assigned inside the library, a value for polarity and subjectivity. When composed expressions are used (i.e., 'very1 great2') the library recognizes the emphasizing word 'very' that precedes 'great', for which polarity is ignored and multiplies the intensity for the following words polarity.

In addition to the existing algorithm, helping the software in cleaning each tweet phrase, we improved the technique changing abbreviations to the full extent of the words (i.e., 'ive', to 'I have', 'im' to 'I am', etc.), this step was very needed since the abbreviation of words is very common in twitter given the limited space for each tweet (140 characters). Given that for words that the library does not detect or identifies, the resulting assigned value is zero. By cleaning each tweet, we reduced considerably the margin error. Example:

We will take a real tweet from May 6, 2018:

$Tesla starts brutal review of contractors, firing everyone that is not 207 vouched for by an employee via @FredericLambert

We can understand that the user is stating that the company Tesla does not have a good relationship with its manufacturing contractors due to its firing policies. The first step is to clean the tweet from characters other than letters and from abbreviations. The algorithm returned the clean sentence:

```
tesla starts brutal review of contractors firing everyone that is not vouched for by an employee via fredericlambert
```

We can observe that all the characters that were not letters disappeared.

The second step the algorithm breaks the tweet in sentences and words. Since Fredic Lambert is a name, it was not considered by our algorithm in the dictionary and does not affect the sentiment of the sentence.

```
[Sentence ("tesla starts brutal review of contractors firing everyone that is not vouched for by an employee via fredericlambert")]
```

```
WordList(['tesla', 'starts', 'brutal', 'review', 'of', 'contractors', 'firing', 'everyone', 'that', 'is', 'not', 'vouched', 'for', 'by', 'an', 'employee','via', 'fredriclambert'])
```

Finally, it analyzes the polarity and subjectivity for the sentence adding the individual score for each word. The individual score is also prerecorded in the library.

```
Sentiment (polarity=-0.875, subjectivity=1.0)
```

The result for this example was a Polarity (P) =-0.875 meaning that it haves an 87.5% of negativity according to our algorithm.

The Processing was applied for each tweet in our unprocessed data with the help of our automated robot (Robot 2).

3.2.3 Index Construction

Using the positive or negative values of polarity for the analyzed tweets, they were added for each day of observation. In Eq 1 the number of negative tweets was added for the day t and in Eq 4 the N vector was constructed for all the structured negative tweets. The same structure was followed for the positive tweets in Eq 2 and the positive vector P in Eq 3.

$$\boldsymbol{N_t = \sum_{n=1}^{i} y_n} \qquad (\mathbf{3.1})$$

$$\boldsymbol{P_t = \sum_{n=1}^{i} x_n} \qquad (\mathbf{3.2})$$

$$\boldsymbol{P_t = [P_t, P_{t+1}, P_{t+2}, \dots]} \qquad (\mathbf{3.3})$$

$$\boldsymbol{N_t = [N_t, N_{t+1}, N_{t+2}, \dots]} \qquad (\mathbf{3.4})$$

3.2.4 Tobit model

Here, we propose as model a limited dependent variables regression [30] to analyze the relationship between companies' stock prices and their corresponding social sentiment index. The model is more suitable than ordinary least squares because it considers that the underlying stock value is censored when the stock price becomes zero. A plausible interpretation is that there is a fundamental value of equity, which can become negative when the stock price becomes zero or close to zero. For example, the fundamental value of equity can be closer to the book value of equity than to the stock price [31]. [32]reports that firms that start to report negative book values of equity are close to financial distress. Because of the limited liability of stockholders, the market stock price cannot be negative. When the stock price is close to zero, it is frequently because the firm is closed to bankruptcy. [33]find that firms that grabbed attention in social media before bankruptcy fillings had significant negative abnormal returns, even before becoming no compliant. Efforts have been made to consider the bankruptcy risk in modeling stock prices and returns [34], [35]. Some of these models consider a lower bound in stock prices, as in [36], [37]. [36] Rubinstein´s (1983) model separates the firm´s assets in risky and riskless ones. The paper interprets the lower bound as the firm´s riskless assets, which can come from past investments.

The risky ones can be those relatively risky, such as the ones in investment opportunities. The proportion of riskless assets can be related to the book to market ratio. As a result, the value of a call option on the risky assets is a displaced [38] option pricing formula. Volatility is variable and depends on the proportion of risky assets. [37] proposes a model in which the bound can be different from zero. The model here can be represented as having the following expression:

$$\boldsymbol{S_t = P_t + N_t} \qquad (\mathbf{3.5})$$

Where: S is the vector of the dependent variable; in this case, it represents the closing daily stock prices, $\boldsymbol{S_t}$ is the closing price of the stock in the day t. P is the vector of the first independent variable, being $\boldsymbol{P_t}$ the sum of a day's positive comments. $\boldsymbol{x_i}$ represents each of the positive comments on a given stock in day. N is a vector of the second independent variable, where $\boldsymbol{N_t}$ is the sum of a day's negative comments on a ticker, and $\boldsymbol{y_i}$ represents each one of the of negative comments in day

3.3 Results

The companies included in the sample were among the 20 largest market-capitalization during the last 20 years and correspond to different economic sectors such as: Banking (Visa, JP Morgan, Royal Dutch), technology (Microsoft, Google, IBM, Intel); investment funds (Berkshire), oil (Exxon Mobil); and others (e.g., Procter and Gamble, General Electric, Tesla, and AT&T). The use of our computerized algorithm yielded both a global and a weighted index of the positive and negative impact of the news on the price of the sample stocks. Since different companies have different presence on the internet, the number of mentions each received in Twitter varied. Yet, in 69% of the cases the negative index was greater than the positive index supporting the economic theory that negative news affects the price of a stock greater than positive news.

Tobin's regression estimation results are presented in Table 1. The companies which data was analyzed are listed in the first column; the McFadden pseudo-R^2 for each company's regression is reported in the second column. This measure is comparable to a linear regression's R^2, but for a Tobit regression a range from 0.1 to 0.4 can be considered an excellent fit. In the third column, the p-value of the regression is expressed and the Log likelihood in the fourth column.

The marginal effects and their corresponding p-values are presented in the columns 5 to 8. Considering that the marginal effects for the Negative index (Column 5) and the Positive Index (Column 7) are the main contributions of our results, we grouped the 16 companies that the Negative index is greater in magnitude with a negative sign, in concordance with our main hypothesis that the negative news negatively affect the performance of stock prices and have a greater impact than the positive influence of positive news in the stock prices.

In Table 1 there are 3 other groups of companies that presented different results than expected. The second group present 3 companies (Microsoft, Volkswagen, and SAP) and returned a positive sign in both indexes, N and P indicating that both negative and positive news affect positively the stock price performance. Our main hypothesis regarding these results is that these companies have a better management of news releases.

In the third group another 3 companies (Bank of America, Procter Gamble and Royal Dutch) present the signs inverted, a positive sign for the N index and a negative sign in the P index. Our hypothesis for these results is that the performance of the companies in the period analyzed presented a negative slope, meaning that the companies performed negatively and thus the inverted signs.

The final group in Table 1 only presents one company, Johnson and Johnson, with the correct sign in the P and N indexes but a p-value above 0.05, which discards the results.

Table 3.1. Description of the Database and Tobit's Regression Estimation Results

	Company	Tobit			Margins			
		Pseudo R2	Prob > chi2	Log likelihood	Neg	(p-value)	Pos	(p-value)
N index is greater than P index in magnitude and with expected signs	Amazon	0.2274	0	-12517.421	-0.0000296	0	0.0000104	0
	Facebook	0.2617	0	-10160.805	-0.0001325	0	0.0000328	0
	eBay	0.2849	0	-9583.8491	-0.0007359	0	0.0001541	0
	AT&T	0.1205	0	-9540.0572	-0.0003445	0	0.0000782	0
	Google	0.1819	0	-21243.784	-0.0000551	0	9.92E-06	0
	JP Morgan	0.2915	0	-11907.441	-0.0012672	0	0.0002868	0
	Tesla	0.1054	0	-11127.786	-0.000018	0	5.42E-06	0
	IBM	0.1939	0	-6030.848	-0.0024723	0	0.0003284	0
	Intel	0.281	0	-9344.8184	-0.000457	0	0.0001201	0
	Berkshire	0.2572	0	-16116.298	-0.0003949	0	0.0000898	0
	Exxon	0.1107	0	-6622.4951	-0.0010151	0	0.0001903	0
	Visa	0.4227	0	-10587.51	-0.0023252	0	0.0005323	0
	Wells Fargo	0.2387	0	-10952.302	-0.0003765	0	0.0000935	0
	CISCO	0.2773	0	-8308.661	-0.0016725	0	0.0003607	0
	General Electric	0.1342	0	-7904.6195	-0.0017143	0	0.0002533	0
	TenCent	0.2591	0	-8939.5771	-0.0012741	0	0.000642	0
N (+) and P (+)	Microsoft	0.2302	0	-12911.447	0.0000269	0	2.43E-06	0
	VW	0.127	0	-4722.659	0.001491	0	0.0004762	0

	SAP	0.2119	0	-11430.718	1.24E-07	0	6.29E-06	0
N (+) and P(-)	Bank of America	0.1278	0	-10512.037	0.0039597	0	-0.0002774	0
	P&G	0.1296	0	-9226.4977	0.0009822	0	-0.0000711	0
	Royal Dutch	0.0044	0	-13139.409	0.0008852	0	-0.0001158	0
NA	Johnson & Johnson	0.2265	0	-11929.585	-0.0168759	**0.58**	0.0593656	0

3.4 Discussion

According to the Efficient Markets Hypothesis [39], it is impossible to predict stock prices as they respond to the arrival of new information, and the news cannot be anticipated. What may be said is that on the arrival of news, the impact they have on stock prices depends on the character of the information they contain. The diversity of factors that may affect a stock price is difficult to conceptualize. But there are certain environmental factors that can be measured and recorded to explore their influence on stock prices. The language contained in the comments and references on given companies in different social media platforms is likely to impact the price of their stock. This paper reports the outcome of an experiment in which a Natural Language Processing algorithm is employed to classify tweets referred to companies as positive (favourable) or negative (unfavourable). These tweets reflect the opinion (qualitative and subjective) of people interested in the companies may be because they are investors or possibly because they are market analysts, but it is easy to visualize they have some interest in sharing their views. The daily number of tweets of each kind (positive and negative) are converted into indices (a positive index and negative index) and are used as explanatory variables of a Tobit regression, with the stock price of companies as the dependent variable. The coefficients are estimated by the regressions confirm the extensively documented fact that the negative news have a larger impact on stock prices than positive news. However, the original contribution of the report consists in the documentation of whether the frequently reported regularity that negative news have a larger impact on stock prices than positive news can be confirmed with the utilization of social media information flows in the form of tweets. The output of the regression allows the comparison between the coefficients of the positive and negative indexes and the clearly prevalent larger absolute value of those that correspond to the latter type, indicates that the experiment's results were highly consistent with what was to be expected.

3.5 Further work

In this chapter we have proposed an initial approach to separate the sentiment index, measure independently the impact of positive and negative shocks. An intuitive development over the proposed methodology would be to focus on the sentiment measurement, focus on improving the library used for said purpose. By increasing the sentiment accuracy, the independent measurement of positive and negative impact of news in the stock price performance would improve naturally.

A different improvement could be considering the sampling size to the full universe of comments of social media instead of the top comments. Doing this, would allow to create an intermediary rating for the comments based on retweets, likes, shares, comments and importance of users, by reach and level of importance (such as analysts or news media).

Chapter 4: A new approach to creating sentiment index

4.1 Introduction

There is a large strand of the finance literature concerned with demonstrating how the news related to a publicly traded company affects its stock price performance and how there is a direct correlation between the direction and magnitude of the stock price response and the nature of the published news [1–5]. The authors of several studies have concluded that the news' impact is related to the media coverage of companies. Furthermore, the authors of other studies have proposed that whether information is public or private is not relevant—what matter is that traders have access to it [6]..

Recently, there have been major advances in developing methodologies to improve accuracy in the prediction of events for financial analysis (crises, market crashes, out of the ordinary portfolio returns in one day, relation between environmental changes and company behavior), but most of research areas have been studied separately even though there may be a logical approach to cross-relate them. Some studies have strongly suggested a direct relationship between social media behavior and market reactions [10,19]. It is widely known that portfolio managers and financial analysts react regarding on behalf of their portfolios when a thread of news is released, and there is a research stream focused on the connection between a company's corporate social responsibility policy (CSR) and the capability of a company to adapt to internal or environmental changes. This theory supports the existence of a negative correlation between confidence from investors and the possibility for a firm to be at default risk [20], an idea that can be interpreted as how investors positively react to a company's CSR policy, reflecting on a company's financial health. If seen as a risk management tool (and the CSR progress monitoring of a company was also carried out), the risk of news that could affect a company's stock price could be measured and its impact could be reduced for the best possible outcome on a company's behalf. Analyzers in different research streams have acknowledged that there are underseen

variables that may affect the outcome of results in experiments since most studies are conducted in controlled environments under laboratory circumstances. If we modify the experiment environment with variating stress conditions, the behavior of the studied models will most certainly be unknown.

The main advantage of using “Big Data” is the ability to analyze large pools of information without the need to classify facts and the risk of unintentionally discarding data that might be relevant to further analysis. An additional advantage is that current technology can be used to perform accurate, real-time analysis from these large pools of information, though most of the existing models were designed to classify data under the researcher criteria.

Another early research team ventured to predict stock market indicators such as Dow Jones, NASDAQ, and S&P 500 by analyzing Twitter posts through the collection of Twitter feeds for six months and analyzing a randomized subsample of nearly one hundred of the full volume of all tweets [21]. The researchers measured collective general feelings, considering fears and hopes on each day, and analyzed the correlation between these indices and stock market indicators. The research team found that a percentage of emotionally negatively Twitter posts correlated with Dow Jones, NASDAQ, and S&P 500 outcomes, as well as a significant positive correlation to Chicago Board Options Exchange Market Volatility Index (VIX) indicators. They determined that analyzing Twitter for emotional responses of any kind provides a hint of how the stock market will behave the following day.

The authors of another study took Twitter as the main source of news due to a clear trend to adopt this platform to disperse news amongst financial traders. In this paper, the researchers applied sentiment analysis algorithms, identified news trends on the web, and compared these trends against financial market movements; they found a positive correlation [22].

A particular approach not used in many research streams was presented by the authors of [23], in which a large amount of information on Twitter over a short time was analyzed. The researchers analyzed 60,944 tweets with the aim of finding a correlation between political action in social media and stock market behavior over a 6-day time window (from 4 May 2016 to 9 May 2016) when political campaigns in the UK were being held. The authors of [24] analyzed Microsoft news on Twitter regarding not only company stock but also opinions on the products and services of the company. In their paper, the researchers proved that a strong statistical inference exists between the rising/falling stock prices of a company accordingly to the public opinions and emotions expressed about that company on Twitter. The main contribution of their work was the development of a sentiment analyzer that can judge the type of sentiment presented in a tweet. The tweets are classified into three categories: positive, negative, and neutral. Initially, the researchers intended to prove that positive emotions or sentiments about a company would reflect in its stock price. However, neutral and negative tweets were also found to affect the stock price. The authors of [25] applied random matrix theory to analyze 64,939 news pieces from the perspective of information theory. This study revealed a correlation between the flow of the New York Times news and 40 world financial indices for a 10-month period between 2015 and 2016. The authors of the study described a dynamic movement between the flow of information and stock price behavior. The model was also tested with and without white noise. A delay between the time the news is published, and the reaction of the stock price was reported.

In [26], quantitative and qualitative factors influencing the dynamics of stock markets were integrated in convolutional neural networks and bidirectional short-term memory to obtain better predictions of stock market movements. Investigating the same problem of stock markets' dynamics prediction using multiple sentiment analysis, the authors of [27] applied seven machine learning classification algorithms. Support vector machines, linear regression and convolutional neural networks were applied to review the shock of Brexit on the European Union's stock markets in [28] using sentiment analysis. The results indicated that deep learning was the best of the studied techniques for the prediction of stock markets. In Korea, the authors of [29] analyzed the common and preferred stock prices, and the difference between them were explained by both corporate factors (as expected) and the sentiment of investors. Similar studies using Twitter information, such as [30], found that sentiment has a positive impact on the effective spread of liquidity when considering the S&P 500 index, among other measures of liquidity. In [31], the impact of sentiment on the volatility of S&P 500 Environmental & Socially responsible index was demonstrated. In [32], sentiment was classified as neutral, positive, or negative, and this polarity was shown to have an impact on the Indian banking index the Bank Nifty.

In the biotechnological sector, social media (volume of Twitter) has been shown to have an impact on the revenues of the companies. The authors of [33] showed that the cumulative average of abnormal returns following initial public offerings are positively affected by sentiment; therefore, the success of large firms (contrary to small firms) is a consequence of the attention received by investors.

This paper was aimed to extend the well-established field of study dedicated to understanding the relationship between social media activity and stock market performance. We began with the theory that general sentiment reflects the economic environment that is generated by news media or the direct observation of the stock market, as well as how this sentiment feeds back into the stock market. The researchers of the aforementioned studies attempted to solve the questions: "Does the stock market affect general sentiment or is it the other way around?" and, furthermore, "Does positive news affect greater and longer than negative news?"

The main contribution of this paper is a standardized framework to measure the sentiment of social media comments and to quantify the impact of populations' optimistic sense on positive performance in the market and the time that it takes for a positive signal to travel from the general population to stock performance. The same framework was applied to quantify the impact of pessimistic feelings of the general population on the negative performance of the stock market and the time that it takes the market to receive a negative signal.

We emphasize a major contribution to the sentiment research field: the methodology in classifying the sentiment of the tweets and the indexes construction introduced in this paper. We were able to demonstrate how the negative index for a company affects additional companies more frequently than the positive index with two different approaches. By applying transfer entropy, we demonstrated that the negative index from a particular company directly or indirectly affects not only that company but also other companies.

We used the EGARCH model, with each studied company's performance as the dependent variable and the corresponding sentiment indexes as the independent variable. This method can be used to measure the direct impact of sentiment indexes, expecting a greater coefficient with a

negative sign for the negative index and a smaller coefficient with a positive sign for the positive index.

The paper is structured as follows. In the first part, we present a framework that shows how the data were extracted from social media, processed, and catalogued to construct our sentiment index. In the second part, we describe how the transfer entropy and EGARCH approaches were applied to the resulting vectors of stock price performance paired with the negative and positive indexes. In the third part, we present the results demonstrating that effective transfer entropy was used to confirm that the negative index affects stock price performance more than the positive index. Finally, to support the asymmetric economic theory that negative news has a greater effect than positive news, EGARCH was applied to stock price performance with the corresponding negative and positive indexes.

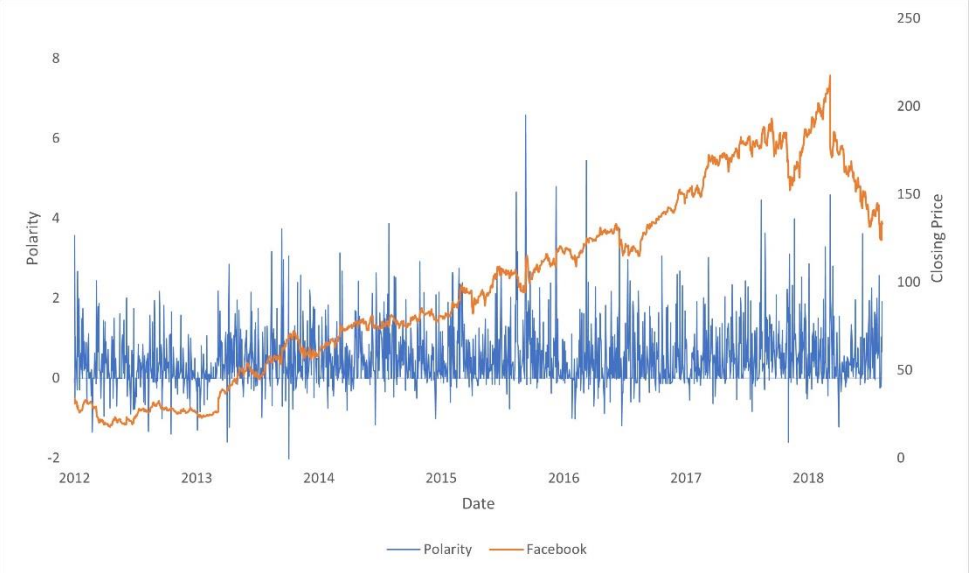
4.2 Materials and Methods

In Figure 4.1, we compare two time series: the stock price of Tesla and the sentiment obtained from Twitter. The synchronized movements from both time series can be visually appreciated. We can infer two hypotheses from this example:

a) There is a relationship between stock price movements and the polarity of the top comments mentioning the ticker of a company.
b) The positive movements in polarity are larger and have more “density” than negative movements.

The previous hypotheses, when initially tested with classic statistical methodologies such as correlation analyses or regressions (OLS, Panel, or Pooled OLS), provided little evidence to support further analysis. Correlations were virtually non-existent between stock performance and sentiment indexes, and R squared in regressions were nearly zero, even when there was statistical significance between the indexes as independent variables and stock performance as the dependent variable.

Figure 4.1. Tesla daily stock price versus Twitter sentiment. It can be appreciated that the sentiment moves accordingly to the stock price.

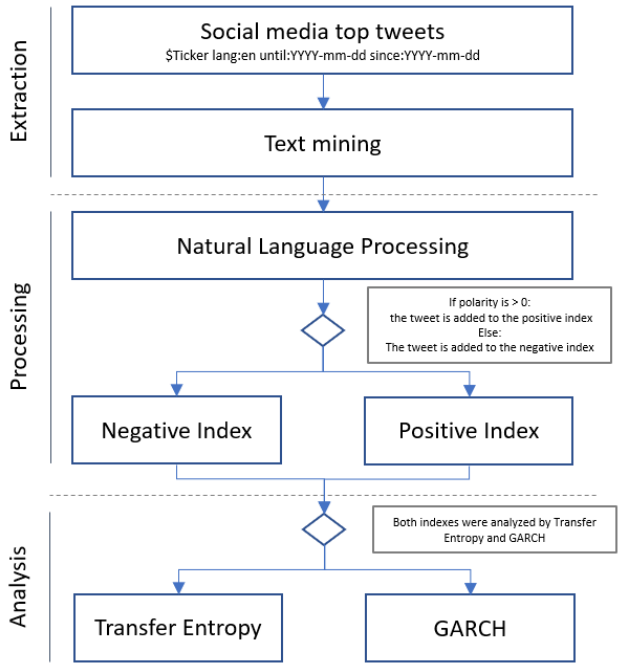


In Figure 4.2, we present our 3-step framework created to test our initial hypotheses that stock prices are affected by Twitter comments' polarity in an asymmetric magnitude, which we present as follows:

1) Extraction: A JSON artificially intelligent (AI) robot that looked for the top tweets that mentioned the tickers of 24 companies (i.e., for Tesla, the ticker would be $TSLA in the English language for the period of 2009–2019) was created.
2) Processing: The tweets were processed using natural language processing to calculate the weighted and normalized polarity. The grading polarity ranged from [−1, +1], so 0 refers to a completely neutral comment and +1 refers to a 100% positive text.

   Sentiment Index: With the polarity already calculated, the tweets were classified as positive or negative and assigned to the corresponding index. If a tweet was graded as 0 (completely neutral), it was discarded.
3) Analysis: The index vectors were paired with the companies' daily closing performance and standardized. For each vector, including company performance, we subtracted the average from the daily observation and divided that value by its standard deviation. Under this treatment, we worked with normal distributions for the final data frame.

Figure 4.2. Framework created to extract, process, and analyze the data from social networks and their impact in stock performance.



The framework was applied to 24 of the largest market-capitalized publicly traded companies that operate in different stock markets and countries, meaning that this method has the flexibility to be applied in different environments.

In Table 1, we present the initial list of target companies and the tickers used as keywords in the social network. Additionally, we introduce the tag used to identify each company in the rest of the figures. The negative and positive indexes were identified by adding the _neg and _pos prefixes, respectively, to each company tag.

Table 4.1. List of the 24 publicly traded companies considered in this study. We include the ticker used in the search word, the country of origin of the company, and the tag (*) used to identify the company in the rest of the figures.

N	Company	Ticker	Country	Tag *
1	Amazon	$AMZN	USA	amazon
2	Facebook	$FB	USA	face
3	Microsoft	$MSFT	USA	microsoft
4	eBay	$EBAY	USA	ebay
5	AT&T	$ATT	USA	att
6	Google	$GOOG	USA	google
7	JP Morgan	$JPM	USA	jpm
8	Tesla	$TSLA	USA	tesla
9	IBM	$IBM	USA	ibm
10	Intel	$INTEL	USA	intel
11	Berkshire Hathaway	$BRKA	USA	brka
12	Exxon	$XOM	USA	exxon
13	Visa	$V	USA	visa
14	Bank of America	$BOA	USA	boa
15	Wells Fargo	$WFF	USA	wf
16	Procter & Gamble	$PG	USA	pg
17	Cisco	$CSCO	USA	csco
18	Johnson & Johnson	$JNJ	USA	jnj
19	General Electric	$GE	USA	ge
20	Royal Dutch	$RDSA	Netherlands	rdsa
21	Ten Cent	$TCEHYN	China	tencent
22	Volkswagen	$VW	Germany	vw
23	SAP	$SAP	Germany	sap
24	Twitter	$TW	USA	tw

4.2.1 Text mining

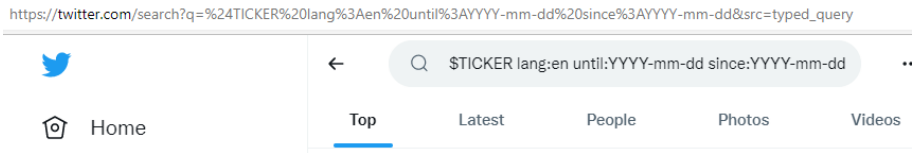
Text mining is a data-mining method within the web-scraping category that is used to retrieve information from selected web pages to create large pools of data that may be analyzed to discover patterns [12]. This was the first step in the analysis.

Twitter was scanned with a JSON routine written in Java. The advantage of this robot (Robot 1) is that the information was extracted and structured in two columns: date and tweet. A second text-mining technique considered the following variables:

Date: The selected period comprised from 1 January 2009 to 1 December 2018, which covered most of a full global economic cycle, i.e., from the aftermath of the 2008 financial crisis to the last months before the economic slowdown of 2019.
Language: The language selected for analyzing the information was English.
Key words: The only word that needed to be mentioned in each tweet was the company ticker (abbreviation used for trading preceded with the $ symbol).
Top tweets: The search engine was able to classify the results of top tweets from a sample of 1% of the most commented and shared tweets.

To automate the search criteria for the companies, a list was created with the tickers, language, and periods of interest. This list was then structured as a search criterion and iterated by the JSON automated robot to extract the results in a structured data frame. In Figure 4.3, we present how the structure was the same for any search term and allowed for such automation.

Figure 4.3. Text structure used as input for the search of samples used in the JSON data-mining robot. (https://twitter.com/search-advanced/ accessed on 15 Dec 2019)



The criteria were applied for the 24 companies considered in the study, and each company was mined individually, meaning that tweets that mentioned 2 or more companies could co-exist; in that case, if Robot 1 detected the same opinion with 2 different tickers, this opinion was used in our study in both the EGARCH model and the transfer entropy measurement. In the final part, each unprocessed database was chronologically ordered. This allowed us to compare data sizes and mention frequency since some companies were founded and traded well after 2009 (Facebook IPO was in 2012).

4.2.2 Sentiment Analysis

Python was the language in which the natural language processing algorithm was coded, and the library used for said task was TextBlob (https://textblob.readthedocs.io/en/dev/ accessed on 21 March 2020). This library calculated sentiment by breaking each individually analyzed text into the words that composed it. Single letter words were ignored, and for the rest of the text, a numeric value for polarity and subjectivity was given to each word that was already assigned inside the library. When composed expressions were used (e.g., 'very1 great2'), the library recognized the emphasizing word 'very' that preceded 'great', for which polarity was ignored, and multiplied the intensity for the following words' polarity.
In addition to the existing algorithm, to help the software clean each tweet phrase, we improved the technique by allowing it to replace abbreviations with the full extent of the words (e.g., 'ive' to 'I have' and 'im' to 'I am'); this step was needed since the abbreviation of words is very common in Twitter due to the limited space for each tweet (140 characters). For words that the library did

not detect or identify, the resulting assigned value was zero. By cleaning each tweet, we considerably reduced the margin error.

For example, we will use a real tweet from 6 May 2018:

'$Tesla starts brutal review of contractors, firing everyone that is not vouched for by an employee via @FredericLambert'

We can understand that the user is stating 'that the company Tesla does not have a good relationship with its manufacturing contractors due to its firing policies. The first step was to clean the tweet from characters other than letters and from abbreviations. The algorithm returned the clean sentence:

'tesla starts brutal review of contractors firing everyone that is not vouched for by an employee via fredericlambert'

We can observe that all the characters that were not letters disappeared.

In the second step, the algorithm broke the tweet in sentences and words. Since Frederic Lambert is a name, it was not considered by our algorithm in the dictionary and did not affect the sentiment of the sentence.

'[Sentence ("tesla starts brutal review of contractors firing everyone that is not vouched for by an employee via fredericlambert")]'

'WordList (['tesla', 'starts', 'brutal', 'review', 'of', 'contractors', 'firing', 'everyone', 'that', 'is', 'not', 'vouched', 'for', 'by', 'an', 'employee', 'via', 'fredriclambert'])'

Finally, the algorithm analyzed the polarity and subjectivity for the sentence, adding the individual scores for each word. The individual score was also prerecorded in the library.

Sentiment (polarity = −0.875, subjectivity = 1.0)

The result for this example was a polarity (P) = −0.875, meaning that it had an 87.5% score of negativity according to the algorithm.

Each tweet in our unprocessed data was processed with the help of our automated robot (Robot 2).

4.2.3 Index Construction

After the polarity was calculated for each tweet, we categorized the tweets into positive and negative categories. Tweets with polarity (0, 1] were tagged as positive sentiments and tweets with polarity [−1, 0) were tagged as negative sentiments. For each company, the number of tweets mentioning the company ticker was calculated daily, with y representing those of the negative index and x representing those of the positive index.

Regarding each company stock price, we created the performance vector by calculating the daily performance, having I representing the daily closing price and R the daily performance. Finally, each vector was standardized; in this manner, we ensured the measurement of the effect of sentiment on the performance of the companies.

$$N_t = \sum_{n=1}^{i} y_n \qquad (4.1)$$

$$Z_{Neg_t} = \frac{N_t - \mu_N}{\sigma_N} \tag{4.2}$$

$$Neg = \left[Z_{Neg_t}, Z_{Neg_{t+1}}, Z_{Neg_{t+2}}, \dots\right] \tag{4.3}$$

$$P_t = \sum_{n=1}^{i} x_n \tag{4.4}$$

$$Z_{Pos_t} = \frac{P_t - \mu_P}{\sigma_P} \tag{4.5}$$

$$Pos = \left[Z_{Pos_t}, Z_{Pos_{t+1}}, Z_{Pos_{t+2}}, \dots\right] \tag{4.6}$$

$$R_t = ln(I_t) - ln(I_{t-1}) \tag{4.7}$$

$$Z_{R_t} = \frac{R_t - \mu_S}{\sigma_S} \tag{4.8}$$

$$S = \left[Z_{R_t}, Z_{R_{t+1}}, Z_{R_{t+2}}, \dots\right] \tag{4.9}$$

4.2.4 Transfer Entropy

The TE from *X* to *Y* can be defined as the average information included in *X* excluding the information reflected by the past state of *Y* for the next state of *Y*. In consequence, TE is defined as follows:

$$T_{X\to Y}(k,l) = \sum_{k,l} p\left(y_{t+1}, y_t^{(k)}, x_t^{(l)}\right) log \frac{p\left(y_{t+1} | y_t^{(k)}, x_t^{(l)}\right)}{p(y_{t+1} | y_y^{(k)})} \tag{4.10}$$

In which y_{t+1} of Y is affected by the k previous states of Y, i.e., the lagged values affecting the current value of Y. In addition, Y is affected by l previous states of X, i.e., the lagged values affecting the current value of X.

X and Y represent two random discrete variables with marginal distributions p(y) and p(x), respectively, with joint probability distributions p (y, x) and dynamics corresponding to Markov processes with order k for system X and l for system Y. One Markov property, which considers the probability of observing Y at time t + 1 in state i conditional on the previous observation of k, is as follows:

$$p(y_{t+1} | y_t, \dots, y_{t-k+1}) = p(y_{t+1} | y_t, \dots, y_{t-k}), y_i \in Y \tag{4.11}$$

The transfer entropy calculation in Equation 11.0 can be applied to discrete data. Since the methodology in this study was applied to financial continuous data, the data were discretized by partitioning them into quantiles. A time series y(t) was partitioned to obtain the symbolically encoded sequence S(t). This sequence replaced the value in the observed time series by discrete states {1, 2, …, n − 1, n}. Denoting the pre-selected number of bins by q1, q2, q3,…, qn, where q1 < q2 < q3, … ,< qn, each value in the original time series was replaced by an integer. Equation 11.0 could be considered biased, mainly by finite sample effects in this case. In addition, higher

signal transfer from time series with higher entropy was expected. To reduce bias, effective transfer entropy was proposed in [34]:

$$ETE_{J\to I}^{boot}(k,l): T_{J\to I}(k,l) - T_{boot\to I}(k,l) \quad (4.12)$$

where $T_{boot\to I}$ indicates the average over the estimates derived from the null bootstrap distribution. The null hypothesis p-value that measured if there was no information exchange is given by $1 - \hat{q}_{T_E}$, in which $\hat{q}_{T_E}$ denotes the quantile of the simulated distribution that corresponds to the transfer entropy estimations when the dependencies from I to J are destroyed. When shuffling the observations of the variables, single observations were randomly arranged into groups that could not occur in the present sample. In consequence, we expected to derive an improved estimation within the J variable corresponding more closely to those observed in the actual sample.

4.2.5 EGARCH

The exponential GARCH model was proposed by Nelson (1991), who presented the conditional equation of variance:

$$ln(\sigma^2) = \omega + \beta ln(\sigma_{t-1}^2) + \gamma \frac{u_{t-1}}{\sqrt{\sigma_{t-1}^2}} + \alpha \left[\frac{|u_{t-1}|}{\sqrt{\sigma_{t-1}^2}} - \sqrt{\frac{2}{\pi}} \right] \quad (4.13)$$

Since the log (σ_t^2) is modeled, when having negative parameters, the value σ_t^2 will be positive. This helped us avoid imposing restrictions to the model parameters. Additionally, asymmetries are permissible under EGARCH use since the relationship between volatility and returns is negative, γ will consequently be negative, meaning that the negative shocks at time *t* have a stronger impact in the variance at time *t + 1* than positive shocks. This asymmetry is known as the leverage effect because the increase in volatility is derived from the increased leverage induced by a negative shock.

4.3 Results

The companies included in the sample were among the 20 largest market-capitalized companies of the last 20 years and correspond to different economic sectors such as banking (Visa, JP Morgan, and Royal Dutch), technology (Microsoft, Google, IBM, and Intel), investment funds (Berkshire), oil (Exxon Mobil), and others (e.g., Procter and Gamble, General Electric, Tesla, and AT&T). The use of our computerized algorithm yielded both global and weighted indexes of the positive and negative impacts of the news on the price of the sample stocks. Since different companies have different levels of presence on the internet, the number of mentions each received in Twitter varied.

We used different methods to measure the influence of social sentiment on the effect of stock performance. Using effective transfer entropy with lags = 1 and 2 and *p*-value ≤ 0.10, we found evidence that there is a relation between sentiment vectors and stock prices. More importantly,

we were able to split the sentiment signal into negative and positive signals and to demonstrate that negativity in social sentiment has more frequent effects of greater impact than positivity. In Table 2 and Table 3, we present the frequency that each category of vectors affected the other groups.

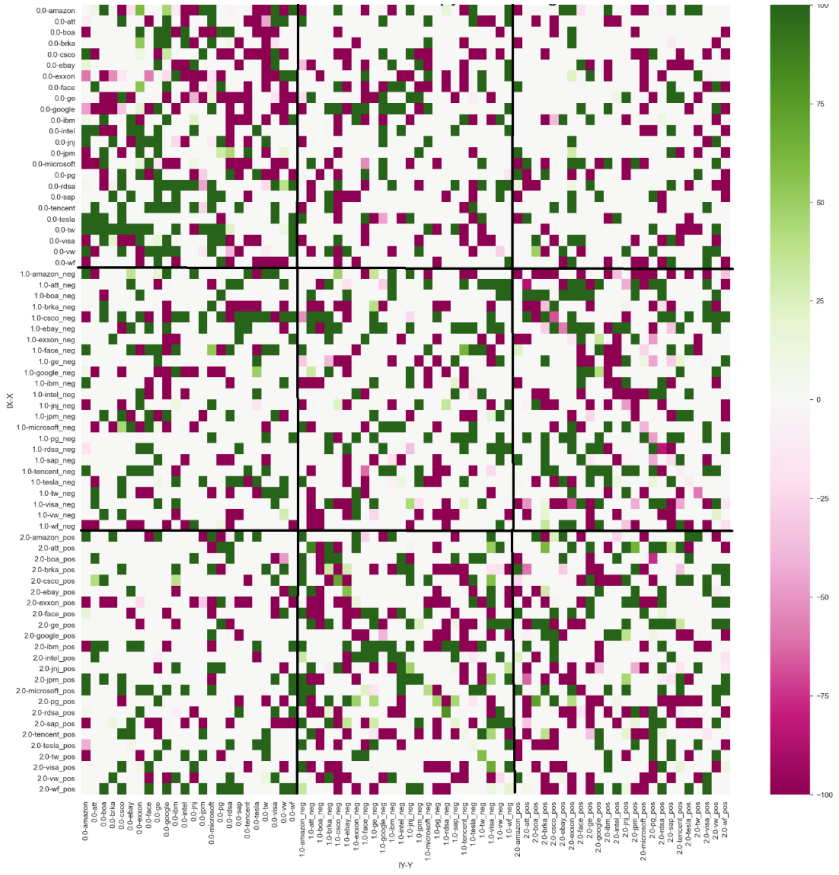
To calculate the intensity of the information transfer between vectors, we calculated the intensity signal with the ETE:

$$Intensity = \frac{ETE(Y \to X)}{ETE(X \to Y) + ETE(Y \to X)} \tag{4.14}$$

With this expression, we included the sign of the signal and were able to identify its direction.

The results of $ETE_{J\to I}^{boot}(k,l)$ with lag k =1 and lag l= 1 and bootstrap simulations of 500 shuffles with p-value $\leq$ 0.10 (Figure 4.4) showed that the negative index sent information to the stock companies for a total of 104 times while receiving information from the stock companies 75 times. The positive index sent information to the stock companies on 92 occasions and received signals 65 times. These results prove that splitting the signals into positive and negative categories is possible and that negative news has more influence on stock performance than positive news. A summary of these results can be found in Table 2.

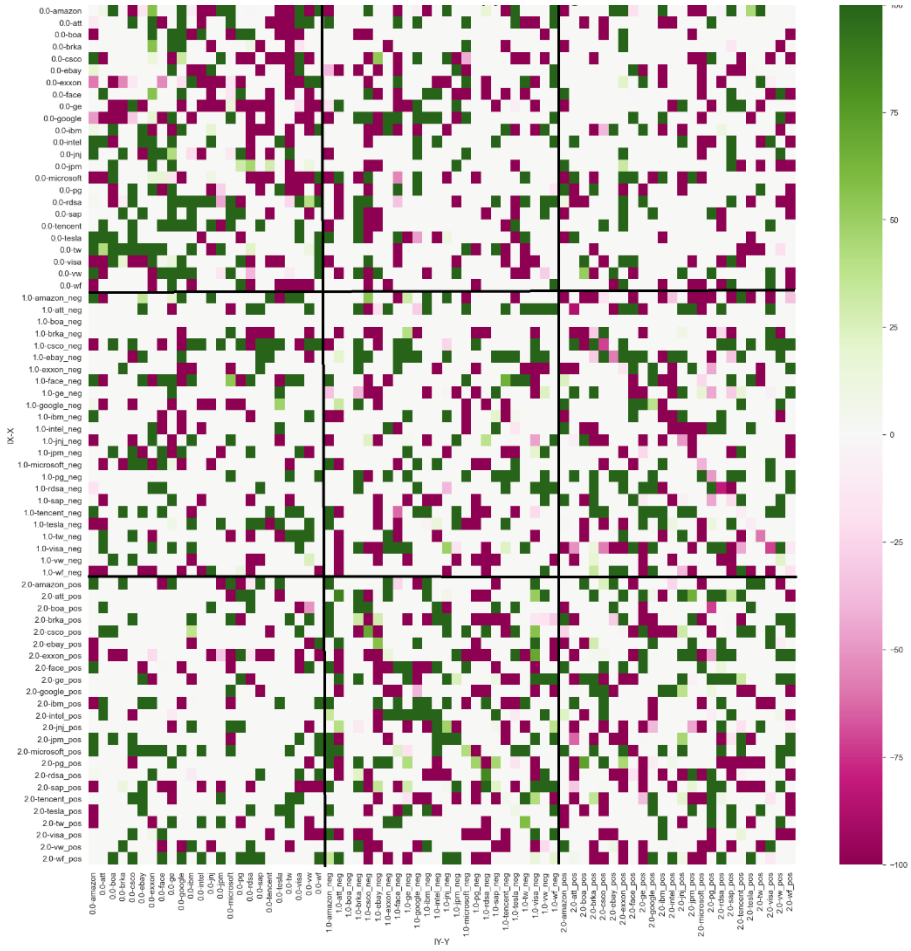
Figure 4.4. Intensity of effective transfer entropy with no lags.



The results of $ETE_{J\to I}^{boot}(k,l)$ with lag k = 1 and lag l = 2 and bootstrap simulations of 500 shuffles with p-value $\leq$ 0.10 (Figure 4.5) demonstrated a similar structure, meaning that the

negative index sent information to the stock companies 97 times and received signals on 75 occasions. The positive index sent signals to the stock performance group for 89 events and received information from stock companies 67 times. A summary of these results can be found in Table 3.

Figure 4.5. Intensity of effective transfer entropy with lag $Y = 2$.



Remarkably, transfer entropy results showed that the negative sentiment index influenced more frequently stock companies than the positive index. Even more, the negative index influenced with greater magnitude the stock performance than the positive index. This is the first major finding of this study.

To present a proposal for measuring the direction and percentage of stock performance that is affected by movement in negative and positive indexes, we fitted an EGARCH model for each company using their corresponding sentiment indexes in addition to the variation of tweets and performance of ACWI (ACWI is an index that represents the performance of the world global stock market, named All Country World Index created by MSCI. For ACWI, we calculated its closing daily performance, and for both the number of tweets and ACWI, we standardized the vectors in order to provide the same treatment as the other variables):

$$S_t = \propto_t + \beta_1 Pos_t + \beta_2 Neg_t + \beta_3 Tweets + \beta_4 ACWI \qquad (4.15)$$

For the variation of tweets (tweet variable), we calculated the daily variation and applied standardization to the transformed time series; the tweet variable can be considered as the

addition of positive and negative indexes. The results for this variable were non-conclusive since only 9 of the 20 companies were presented in the results. Even when a causal relationship was found in 45% of the cases, it was not until the signal was split in positive and negative that results could be considered conclusive.

Our second was that in 83% of the cases considered by the EGARCH model, the negative index coefficient was greater than the positive index coefficient and the signs of the coefficients were negative for the negative index and positive for the positive index. This is a finding of asymmetry, i.e., the effect of the negativity was stronger and lasted longer than that of positivity. We present the results of EGARCH for each company that fulfilled the expected results in Table 4.

For each independent variable, we included the coefficient value with the corresponding *p*-value and T statistic. As expected, the positive index of tweets (classified as positive) had a positive impact on the returns of the companies. Similarly, the negative index of tweets (classified as negative) had a negative impact on the returns of the companies. Additionally, both models showed that the negative impact was greater in absolute value than the positive impact (impacts were given by the absolute value of the coefficients), results in accordance with the premise that the indexes can be used to demonstrate the impact of the population's sentiment in the stock performance.

We observed an asymmetric effect on the variance via the gamma coefficient in the variance equation, which represents the leverage effect—a well-known financial phenomenon that involve returns. Negative shocks were found to provoke increases in volatility more than positive shocks. A simple GARCH cannot explain this fact in typical finance time series.

The results for most of the studied companies were expected, i.e., the negative index presented a negative sign and was greater than the positive index in absolute terms and the positive index presented a positive sign and was smaller than the negative index in absolute terms. The only companies that did not fulfilled the expected results were Berkshire Hataway, Cisco, Royal Dutch, and Volkswagen; we attribute this to the smaller sample of tweets obtained for these companies, which were the smallest of those analyzed.

Table 4.2. Summary of signal events with lag X = lag Y = 1.

		Y					
		Stocks		**Negative Index**		**Positive Index**	
		X->Y	**Y->X**	**X->Y**	**Y->X**	**X->Y**	**Y->X**
X	Stocks	151	149	75	104	65	92
	Negative Index	104	75	101	96	120	129
	Positive Index	92	65	129	120	128	127

Table 4.3. Summary of signal events with lag X = 1, lag Y = 2.

		Y					
		Stocks		**Negative Indexes**		**Positive Indexes**	
		X->Y	**Y->X**	**X->Y**	**Y->X**	**X->Y**	**Y->X**
X	Stocks	145	151	75	97	67	89
	Negative Index	97	75	94	95	111	121
	Positive Index	89	67	121	111	123	128

Table 4.4. EGARCH results

Company	R2	Constant			Positive Index			Negative Index			Number of Tweets			ACWI		
		Coefficient	T stats	P-value	Coefficient	T stats	P-value	Coefficient	T stats	P-value	Coefficient	T stats	P-value	Coefficient	T stats	P-value
Amazon	28%	-0.01	-0.55	0.59	0.10	2.68	0.01	-0.10	-2.24	0.02	0.06	2.05	0.04	0.49	22.46	0.00
At&t	28%	0.01	10.78	0.00	0.01	2.15	0.03	-0.02	-3.46	0.00	0.01	2.68	0.01	0.52	30.99	0.00
Bank of America	46%	-0.01	-3.09	0.00	0.05	7.31	0.00	-0.08	-24.64	0.00	0.03	6.44	0.00	0.60	26.31	0.00
eBay	31%	0.02	28725.15	0.00	0.02	2.82	0.00	-0.02	-679.17	0.00	0.01	9.24	0.00	0.56	206.23	0.00
Exxon	41%	0.00	0.57	0.57	0.02	4.26	0.00	-0.04	-2.65	0.01	0.00	0.97	0.33	0.58	30.01	0.00
Facebook	12%	-0.07	-18.52	0.00	0.06	8.19	0.00	-0.17	-38.35	0.00	0.00	0.65	0.52	0.33	38.32	0.00
General Electric	37%	0.02	3.17	0.00	0.03	18.08	0.00	-0.06	-4.44	0.00	0.01	4.20	0.00	0.59	37.43	0.00
Google	35%	-0.01	-4.31	0.00	0.03	3.44	0.00	-0.04	-5.44	0.00	0.02	4.43	0.00	0.56	73.29	0.00
IBM	39%	0.02	1.19	0.23	0.03	2.39	0.02	-0.13	-53.64	0.00	0.00	0.20	0.84	0.59	139.32	0.00
Intel	38%	-0.02	-44.20	0.00	-0.01	-10.64	0.00	0.02	7.96	0.00	-0.01	-9.35	0.00	0.61	23.08	0.00
Johnson & Johnson	36%	-0.01	-2.08	0.04	0.02	2.72	0.01	-0.07	-8.62	0.00	0.00	-0.15	0.88	0.56	291.59	0.00
JP Morgan	56%	-0.01	-0.44	0.66	0.03	1.96	0.05	-0.08	-3.11	0.00	0.00	0.18	0.86	0.72	44.29	0.00
Microsoft	41%	0.00	0.34	0.73	0.04	4.74	0.00	-0.06	-4.43	0.00	0.02	1.40	0.16	0.61	32.47	0.00
Procter and Gamble	9%	-0.12	-3180.72	0.00	0.04	3587.49	0.00	-0.07	-5544.68	0.00	0.02	20240.00	0.00	0.34	2814.54	0.00
SAP	51%	0.00	-0.64	0.52	0.02	2.82	0.00	-0.04	-2.59	0.01	0.01	0.74	0.46	0.70	39.34	0.00
Tencent	30%	-0.02	-3.56	0.00	0.02	13.48	0.00	-0.02	-3.78	0.00	0.01	2.62	0.01	0.54	43.81	0.00
Tesla	11%	-0.04	-1.93	0.05	0.10	3.03	0.00	-0.15	-3.29	0.00	0.03	1.39	0.16	0.32	13.38	0.00
Twitter	14%	0.02	1.13	0.26	0.18	24.74	0.00	-0.32	-11.49	0.00	0.01	0.24	0.81	0.26	5.66	0.00
Visa	39%	0.00	0.02	0.98	0.02	2.69	0.01	-0.03	-3.30	0.00	0.01	1.86	0.06	0.63	42.93	0.00
Wells Fargo	54%	0.00	0.07	0.94	0.02	9.45	0.00	-0.03	-2.49	0.01	0.01	0.79	0.43	0.70	40.37	0.00

4.4 Discussion

The authors of [36] found a strong causal flow of information between the prices and sentiments of different studied companies, as well as flow in directions from prices to sentiment and sentiment to prices in the top 50 S&P companies for the period of 2018–2020; in the present

study, we found a separation of negative and positive sentiments for a selection of 24 companies for a 10-year period that included a full economic cycle, with periods of high stress and high positive growth.

Transfer entropy was applied for the in-depth analysis (it can be seen as liquidity) of the market with intraday frequency for the Warsaw Stock Exchange in [37] and in portfolio selection (multiperiod and fuzzy returns) in [38]; in the same line of portfolio selection using entropy, the authors of [39] applied an approach considering mean, variance, and skewness. The research directions embodied by mentioned studies are proposed for future opportunities to improve the speed and efficiency of existing approaches used to measure signals in real time.

Another application of transfer entropy currently is in asset selection for robust portfolio construction. For instance, the authors of [40] applied a numerical method to maximize entropy, analyzing the flow of information in Chinese and American stock markets, and the authors of [41] found that bank sector in China and the technology sector in the USA are the most prominent in information flow, which is highly influenced by the heavy presence of technological companies on social media. Moreover, the bank and energy sectors of China and the USA, respectively, are the largest in terms of the net flow of information. The transfer entropy methodology is now being used in other areas, e.g., sentiment related to different assets such as gold, cryptocurrencies, and bonds were studied in [42] during interesting moments associated with negative sentiment (tweets of Elon Musk and Dogecoin); the effect of Elon Musk tweets was also measured in this study (encompassed in tweets that mention Tesla's ticker ($TSLA). And a last example would be [43] in which the predictive power of a wide range of determinants on bitcoins' price direction under the continuous transfer entropy approach as a feature selection criterion was tested.

4.5 Conclusion

In our previous study [44], we presented an early version of a general sentiment index that included the aggregate daily sentiment for the comments mentioning a given stock. In this study, we successfully split the signal into positive and negative categories for the same sample of companies.

By including the tweet variable in the EGARCH modeling, we observed that finding a direct causal relationship is difficult due the low success rate of measuring a statistically significant signal from the tweet variable towards stock performance. It is not until we applied the natural language processing treatment to the data and split them into positive and negative categories that we found that the negative sentiment observed in the general population was translated to the stock market with a greater negative effect than the positive sentiment's positive effect.

The daily numbers of tweets of each kind (positive and negative) were converted into indices (a positive and negative index) and used as explanatory variables of transfer entropy and EGARCH models, with the stock performance of companies as the dependent variable. The coefficients estimated by the regressions confirmed the extensively documented fact that negative news has a larger impact on stock prices than positive news. However, the original contribution of this report is the documentation that the frequently reported regularity that negative news has a larger impact on stock prices than positive news could be confirmed with the utilization of social media information flows in the form of tweets. The output of the GARCH model allowed for comparisons

between the coefficients of positive and negative indexes, and the clearly larger absolute values of those that correspond to the negative index indicated that the experiment's results were highly consistent with what was expected.

4.6 Following steps

In this last chapter we presented a method for measuring the impact of positive and negative comments in the stock performance separately. Also, the volatility factor was included indirectly in the EGARCH modelling but was largely ignored due the length of the study. An intuitive expansion of this study would focus on the volatility impact of the sentiment indexes in the stock performance.

Furthermore, the EGARCH models could be adapted to high frequency performance with the increase of sampling to the same level. With this improvement, it could be possible to measure in real time the impact of social sentiment in the stock performance.

References

Chapter 2

1. DeGennaro R., Shrieves R. Public information releases, private information arrival and volatility in the foreign exchange market. Journal of Empirical Finance. 1997; 4:295–315. https://doi.org/10.1016/ S0927-5398(97)00012-1
2. Fabrizio Lillo, Salvatore Miccichè and Tumminello Michele and Piilo Jyrki and Mantegna Rosario N. How news affect the trading behavior of different categories of investors in a financial market. Quantitative Finance.2015; 15:213–229. https://doi.org/10.1080/14697688.2014.931593
3. Vega C. Stock price reaction to public and private information. Journal of Financial Economics. 2004; 82:103–133. https://doi.org/10.1016/j.jfineco.2005.07.011
4. Fama Eugene. Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance.1970; 25(2):383–417. https://doi.org/10.1111/j.1540-6261.1970.tb00518.x
5. Lim KP, Brooks R. The evolution of stock market efficiency over time: A survey of the empirical literature. Journal of Economic Surveys.2011; 25(10):69–108.
6. Shi Huai-Long, Jiang Zhi-Qiang, Zhou Wei-Xing. Time-varying return predictability in the Chinese stock market. Reports in Advances of Physical Sciences.2016; 1(1):1–11.
7. Arendas Peter, Bozena Chovancova. The Adaptive Markets Hypothesis and the BRIC Share Markets. Ekonomicky´časopis.2015; 63(10):1003–1018. ISSN 0013-3035
8. Adaramola Anthony Olugbenga, Obisesan Oluwaseun Grace. Adaptive Market Hypothesis: Evidence from Nigerian Stock Exchange The Journal of Developing Areas.2021;(55) 2:153–165
9. Dzung Phan Tran Trung, Hung Pham Quang. Adaptive Market Hypothesis: Evidence from the Vietnamese Stock Market. Journal of Financial Risk and Financial Management.2019; 12(81):193–206.
10. Tahmina Akhter, Othman Yong. Can Adaptive Market Hypothesis Explain the Existence of Seasonal Anomalies? Evidence from Dhaka Stock Exchange, Bangladesh. Contemporary Economics.2021; 15(2):198–223. https://doi.org/10.5709/ce.1897-9254.444
11. Mahdi Moradi, Mehdi Jabbari Nooghabi, Mohammad Mahdi Rounaghi. Investigation of fractal market hypothesis and forecasting time series stock returns for Tehran Stock Exchange and London Stock Exchange. International Journal of Finance and Economics.2021; 26:662–678. https://doi.org/10.1002/ ijfe.1809
12. Ladislav Kristoufek. Fractal Markets Hypothesis and the Global Financial Crisis: Wavelet Power Evidence. Scientific Reports.2013; 3:2857. https://doi.org/10.1038/srep02857
13. Dar Arif Billah, Bhanja Niyati, Tiwari Aviral Kumar. Do global financial crises validate assertions of fractal market hypothesis?. International Economics and Economic Policy.2017; 14:153–165. https://doi.org/ 10.1007/s10368-015-0332-0
14. Hisham Farag, Robert Cressy. Stock market regulation and news dissemination: evidence from an emerging market. The European Journal of Finance.2012; 18:351–368. https://doi.org/10.1080/

1351847X.2011.579740

15. Thompson James H. A Global Comparison of Insider Trading Regulations. International Journal of Accounting and Financial Reporting.2013; 3(1):2162–3082. https://doi.org/10.5296/ijafr.v3i1.3269
16. Tanja Boskovic, Caroline Cerruti, Michel Noel. Comparing European and U.S. Securities Regulations. World Bank.2010; 184.
17. Jones Emily, Knaack Peter. Global Financial Regulation: Shortcomings and Reform Options. Global Policy.2019; 10(2):193–206. https://doi.org/10.1111/1758-5899.12656
18. Carlos Carvalho, Nicholas Klagge, Emanuel Moench. The persistent effects of a false news shock. Journal of Empirical Finance. 2011; 18:597–615.
19. Chan Wesley S. Stock price reaction to news and no-news: drift and reversal after headlines. Journal of Financial Economics. 2001; 18:203–260.
20. Arditi E, Yechiam E, Zahavi G. Association between Stock Market Gains and Losses and Google Searches. PLoS ONE. 2015; 10(10):e0141354. https://doi.org/10.1371/journal.pone.0141354
21. Ali Derakhshan, Hamid Beigyh. Sentiment analysis on stock social media for stock price movement prediction. Engineering Applications of Artificial Intelligence.2019; 85:569–578. https://doi.org/10.1016/j. engappai.2019.07.002
22. Nguyen T.H., Shirai K., Velcin J. Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications in Intelligence.2015; 42:9603–9611. https://doi.org/10.1016/j.eswa. 2015.07.052
23. Johan Bollen, Huina Mao, Xiaojun Zeng. Twitter mood predicts the stock market. Journal of Computational Science.2011; 2:1–8. https://doi.org/10.1016/j.jocs.2010.12.007
24. Johan Bollen, Huina Mao, Xiaojun Zeng. Is all that talk just noise? the information content of internet stock message boards. Journal of Finance.2004; 59:1259–1294. https://doi.org/10.1111/j.1540-6261. 2004.00662.x
25. Tumarkin R., Whitelaw R.F. News or noise? internet postings and stock prices. Financial Analysts Journal. 2001; 57:41–51. https://doi.org/10.2469/faj.v57.n3.2449
26. Steinert L., Herff C. Predicting altcoin returns using social media. PLoS ONE.2018; 13:1–12. https://doi. org/10.1371/journal.pone.0208119
27. Cutler David M., Poterba James M., Summers Lawrence H. What moves stock prices?. Journal of Portfolio Management.1989; 15:4–12. https://doi.org/10.3905/jpm.1989.409212
28. Tetlock P. C., Saar-Tsechansky M., Sofus Macskassy. More than words: Quantifying language to measure firms' fundamentals. Journal of Finance.2008; LXIII:1437–1467.
29. Yinghao Ren,Fangqing Liao, Yongjing Gong. Impact of News on the Trend of Stock Price Change: an Analysis based on the Deep Bidirectiona LSTM Model. Procedia Computer Science.2020; 174:128–140. https://doi.org/10.1016/j.procs.2020.06.068
30. Thomas Schreiber. Measuring Information Transfer. Physical review letters.2000; 85:461–464. https:// doi.org/10.1103/PhysRevLett.85.461
31. Yao Can-Zhong and Li Hong-Yu. Effective Transfer Entropy Approach to Information Flow Among EPU, Investor Sentiment and Stock Market. Frontiers in Physics.2020; 8:206. https://doi.org/10.3389/

fphy.2020.00206

32. Dewi Luisiana Citra, Chandra Meiliana Alvin. Social Media Web Scraping using Social Media Developers API and Regex. Procedia Computer Science.2019; 157:444–449. https://doi.org/10.1016/j.procs. 2019.08.237
33. Aggarwal, Anupama. TextBlob. TextBlob.2013;. https://textblob.readthedocs.io/en/dev/quickstart.html
34. Barnett Lionel and Barrett Adam B and Seth Anil K. Granger causality and transfer entropy are equivalent for Gaussian variables. APS.2009; 103(23):238–701.
35. Katerina Schindlerova. Equivalence of Granger Causality and Transfer Entropy: A Generalization. Applied Mathematical Sciences.2011; 5(73):3637–3648.
36. Barnett Lionel and Bossomaier Terry. Transfer entropy as a log-likelihood ratio. APS.2012; 109(13):138105.
37. Bossomaier Terry and Barnett Lionel and Harre´ Michael and Lizier Joseph T. An introduction to transfer entropy. Springer.2016; 71(4):65–95.
38. García-Medina Andrés and González Farías Graciela. Transfer entropy as a variable selection methodology of cryptocurrencies in the framework of a high dimensional predictive model. PloS one.2020; 15(1):e0227269. https://doi.org/10.1371/journal.pone.0227269
39. Robert Marschinski and Holger Kantz. The European Physical Journal B-Condensed Matter and Complex Systems. Springer.2002; 30(2):275–281.
40. Dimpfl Thomas and Peter Franziska Julia. Using transfer entropy to measure information flows between financial markets. Studies in Nonlinear Dynamics & Econometrics.2013; 17(1):85–102.
41. Horowitz Joel L. Bootstrap methods for Markov processes. Econometrica.2003; 71(4):1049–1082.
42. Agustin Garcia Asuero, Gustavo Gonzalez, Ana Sagayo. The Correlation Coefficient: An Overview. Critical Reviews in Analytical Chemistry.2006; 36:41–59. https://doi.org/10.1080/10408340500526766
43. Prokopenko M., Lizier J. Transfer Entropy and Transient Limits of Computation. Scientific Reports.2014; 4:5394. https://doi.org/10.1038/srep05394
44. Piskorec M., Antulov-Fantulin N., et al. Cohesiveness in Financial News and its Relation to Market Volatility. Scientific Reports.2014;:5038.
45. Granger C. W. J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica.1969; 37:424–438. https://doi.org/10.2307/1912791
46. Chan Wesley S. Stock price reaction to news and no-news: drift and reversal after headlines. Journal of Financial Economics.2003; 70:223–260.
47. García-Medina Andrés and Sandoval Leonidas Junior and Efraı´nañuelos Urrutia B and Martinez Arguello A.M. Correlations and flow of information between the New York Times and stock markets. Physica A: Statistical Mechanics and its Applications.2018; 502:403–415. https://doi.org/10.1016/j.physa.2018.02.154
48. Nisar Tahir M., Yeung Man. Twitter as a tool for forecasting stock market movements: A short-window event study. The Journal of Finance and Data Science 4.2018:101–119. https://doi.org/10.1016/j.jfds.

2017.11.00

Chapter 3

1. R. P. DeGennaro and R. E. Shrieves, “Public information releases, private information arrival and volatility in the foreign exchange market 4, 295-315,” J. Empir. Finance, vol. 4, no. 4, pp. 295–315, 1997, doi: https://doi.org/10.1016/S0927-5398(97)00012-1.
2. A. Atkins, M. Niranjan, and E. Gerding, “Financial news predicts stock market volatility better than close price,” J. Finance Data Sci., vol. 4, pp. 120–137, 2018.
3. F. Audrino, F. Sigrist, and D. Ballinaria, “The impact of sentiment and attention measures on stock market volatility,” Int. J. Forecast., vol. 36, pp. 334–357, 2020.
4. Y. Ren, F. Liao, and Y. Gong, “Impact of News on the Trend of Stock Price Change: an Analysis based on the Deep Bidirectiona LSTM,” Procedia Comput. Sci., vol. 174, pp. 128–140, 2020.
5. X. Li, L. Chen, J. Wang, and X. Deng, “News impact on stock price return via sentiment analysis,” Knowl.-Based Syst., vol. 69, pp. 14–23, 2014.
6. C. Vega, “Stock price reaction to public and private information,” J. Financ. Econ., vol. 82, no. 1, pp. 103–133, 2006, doi: https://doi.org/10.1016/j.jfineco.2005.07.011.
7. W. S. Chan, “Stock price reaction to news and no-news: drift and reversal after headlines,” J. Financ. Econ., vol. 70, no. 2, pp. 223–260, 2003, doi: https://doi.org/10.1016/S0304-405X(03)00146-6.
8. A. Derakhshan and H. Beigy, “Sentiment analysis on stock social media for stock price movement prediction,” Eng. Appl. Artif. Intell., vol. 85, pp. 569–578, 2019, doi: https://doi.org/10.1016/j.engappai.2019.07.002.
9. T. H. Nguyen, K. Shirai, and J. Velcin, “Sentiment analysis on social media for stock movement prediction,” Expert Syst. Appl., vol. 42, no. 24, pp. 9603–9611, 2015, doi: https://doi.org/10.1016/j.eswa.2015.07.052.
10. J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” J. Comput. Sci., vol. 2, no. 1, pp. 1–8, 2011, doi: https://doi.org/10.1016/j.jocs.2010.12.007.
11. W. Antweiler and M. Z. Frank, “Is all that talk just noise? the information content of internet stock message boards,” J. Finance, vol. 59, no. 3, pp. 1259–1294, 2004, doi: https://doi.org/10.1111/j.1540-6261.2004.00662.x.
12. R. Tumarkin and R. F. Whitelaw, “News or Noise? Internet Postings and Stock Prices,” Financ. Anal. J., vol. 57, no. 3, pp. 41–51, 2001, doi: 10.2469/faj.v57.n3.2449.
13. L. Steinert and C. Herff, “Predicting altcoin returns using social media,” PLoS ONE, vol. 13, no. 12, 2018, doi: 10.1371/journal.pone.0208119.
14. B. Fischer, “Studies of stock price volatility changes,” Meetings of the Business and Economics, vol. Statistics Section, no. American Statistical Association, pp. 177–181, 1976.
15. G. W. Schwert, “The Adjustment of Stock Prices to Information About Inflation,” J. Finance, vol. 36, no. 1, pp. 15–29, 1981, doi: 10.1111/j.1540-6261.1981.tb03531.x.

16. D. K. Pearce and V. V. Roley, “Stock Prices and Economic News,” J. Bus., vol. 58, no. 1, pp. 49–67, 1985, doi: 10.1086/296282.
17. D. M. Cutler, J. M. Poterba, and L. H. Summers, “What moves stock prices?,” J. Portf. Manag., vol. 15, no. 3, pp. 4–12, 1989, doi: https://doi.org/10.3905/jpm.1989.409212.
18. G. McQueen and V. Roley, “Stock Prices, News and Business Conditions,” Rev. Financ. Stud., vol. 6, no. 3, pp. 683–708, 93, doi: 10.1093/rfs/6.3.683.
19. D. Kahneman and A. Tversky, “Prospect Theory: An Analysis of Decision Under Risk,” Econometrica, vol. 47, no. 2, pp. 263–291, 1979, doi: 10.2307/1914185.
20. A. Al Nasseri, F. Menla Ali, and A. Tucker, “Good news and bad news: Do online investor sentiments reaction to return news asymmetric?,” MIDAS, vol. 1774, pp. 55–66, 2016.
21. Y. Ruan, A. Durresi, and L. Alfantoukh, “Using Twitter trust network for stock market analysis,” Knowl.-Based Syst., vol. 145, no. 1, pp. 207–218, 2018, doi: https://doi.org/10.1016/j.knosys.2018.01.016.
22. W. Sun and K. Cui, “Linking corporate social responsibility to firm default risk,” Eur. Manag. J., vol. 32, no. 2, pp. 275–287, 2014, doi: https://doi.org/10.1016/j.emj.2013.04.003.+
23. X. Zhang, H. Fuehres, and P. A. Gloor, “Predicting Stock Market Indicators Through Twitter ‘I hope it is not as bad as I fear,’” Procedia - Soc. Behav. Sci., vol. 26, pp. 55–62, 2011, doi: https://doi.org/10.1016/j.sbspro.2011.10.562.
24. S. Y. Yang, S. Y. Kevin Mo, and A. Liu, “Twitter financial community sentiment and its predictive relationship to stock market movement,” Quant. Finance, vol. 15, no. 10, pp. 1–27, 2015, doi: 10.1080/14697688.2015.1071078.
25. T. M. Nisar and M. Yeung, “Twitter as a tool for forecasting stock market movements: A short-window event study,” J. Finance Data Sci., vol. 4, no. 2, pp. 101–119, 2018, doi: https://doi.org/10.1016/j.jfds.2017.11.002.
26. V. S. Pagol, K. N. Reddy Challa, and G. Panda, “Sentiment Analysis of Twitter Data for Predicting Stock Market Movements,” Int. Conf. Signal Process. Commun. Power Embed. Syst., vol. SCOPES, 2016, [Online]. Available: https://arxiv.org/pdf/1610.09225.pdf
27. A. García-Medina, L. Sandoval Junior, E. Urrutia Bañuelos, and A. M. Martínez-Argüello, “Correlations and flow of information between the New York Times and stock markets,” Phys. Stat. Mech. Its Appl., vol. 502, no. 15, pp. 403–415, 2018, doi: https://doi.org/10.1016/j.physa.2018.02.154.
28. F. Johnson and S. Kumar Gupta, “Web Content Mining Techniques: A Survey,” Int. J. Comput. Appl., vol. 47, no. 11, pp. 0975–888, 2012.
29. G. G. Chowdhury, “Natural language processing,” Annu. Rev. Inf. Sci. Technol., vol. Chapter 2, pp. 51–88, 2005, doi: https://doi.org/10.1002/aris.1440370103.
30. J. Tobin, “Estimation of relationship for limited dependent variables,” Econometrica, vol. 26, no. 1, pp. 24–36, 1958, doi: https://doi.org/10.2307/1907382.
31. S. J. Brown, P. Lajbcygier, and B. Li, “Going Negative: What to Do with Negative Book Equity Stocks,” J. Portf. Manag., vol. 35, no. 1, pp. 95–102, 2008, doi: https://doi.org/10.3905/JPM.2008.35.1.95.
32. T. C. Ang, “Are Firms with Negative Book Equity in Financial Distress?,” Finance Corp. Gov. Conf., 2010, doi: http://dx.doi.org/10.2139/ssrn.1533964.

33. T. Reyes and N. Waissbluth, “Saddled with Attention: Overreaction to Bankruptcy Filings,” Int. Rev. Finance, vol. 19, no. 4, pp. 787–819, 2019, doi: https://doi.org/10.1111/irfi.12199.
34. R. C. Merton, “Option pricing when underlying stock returns are discontinuous,” J. Financ. Econ., vol. 3, no. 1–2, pp. 125–144, 1976, doi: https://doi.org/10.1016/0304-405X(76)90022-2.
35. A. Câmara, I. Popova, and B. J. Simkins, “Options on Troubled Stock,” J. Future Mark., vol. 34, no. 7, pp. 637–657, 2014, doi: 10.1002/fut.21616.
36. M. Rubinstein, “Displaced Diffusion Option Pricing,” J. Finance, vol. 38, no. 1, pp. 213–217, 1983, doi: https://doi.org/10.1111/j.1540-6261.1983.tb03636.x.
37. M. Dong, “Option pricing with a non-zero lower bound on stock price,” J. Futur. Mark., vol. 25, no. 8, pp. 775–794, 2005, doi: 10.1002/fut.20159.
38. F. Black and M. Scholes, “The Pricing of Options and Corporate Liabilities,” J. Polit. Econ., vol. 81, no. 3, pp. 637–654, 1973, doi: 10.1086/260062.
39. E. Fama, “Efficient capital market: A review of theory and empirical work,” J. Finance, vol. 25, no. 2, pp. 383–417, 1970.

Chapter 4

1. DeGennaro, R.P.; Shrieves, R.E. Public information releases, private information arrival and volatility in the foreign exchange market. *J. Empir. Financ.* 1997, *4*, 295–315. https://doi.org/10.1016/S0927-5398(97)00012-1.
2. Atkins, A.; Niranjan, M.; Gerding, E. Financial news predicts stock market volatility better than close price. *J. Financ. Data Sci.* 2018, *4*, 120–137.
3. Audrino, F.; Sigrist, F.; Ballinaria, D. The impact of sentiment and attention measures on stock market volatility. *Int. J. Forecast.* 2020, *36*, 334–357.
4. Ren, Y.; Liao, F.; Gong, Y. Impact of News on the Trend of Stock Price Change: An Analysis based on the Deep Bidirectiona LSTM. *Procedia Comput. Sci.* 2020, *174*, 128–140.
5. Li, X.; Chen, L.; Wang, J.; Deng, X. News impact on stock price return via sentiment analysis. *Knowl.-Based Syst.* 2014, *69*, 14–23.
6. Vega, C. Stock price reaction to public and private information. *J. Financ. Econ.* 2006, *82*, 103–133. https://doi.org/10.1016/j.jfineco.2005.07.011.
7. Chan, W.S. Stock price reaction to news and no-news: Drift and reversal after headlines. *J. Financ. Econ.* 2003, *70*, 223–260. https://doi.org/10.1016/S0304-405X(03)00146-6.
8. Derakhshan, A.; Beigy, H. Sentiment analysis on stock social media for stock price movement prediction. *Eng. Appl. Artif. Intell.* 2019, *85*, 569–578. https://doi.org/10.1016/j.engappai.2019.07.002.
9. Antweiler, W.; Frank, M.Z. Is all that talk just noise? the information content of internet stock message boards. *J. Financ.* 2004, *59*, 1259–1294. https://doi.org/10.1111/j.1540-6261.2004.00662.x.

10. Tumarkin, R.; Whitelaw, R.F. News or Noise? Internet Postings and Stock Prices. *Financ. Anal. J.* 2001, *57*, 41–51. https://doi.org/10.2469/faj.v57.n3.2449.
11. Steinert, L.; Herff, C. Predicting altcoin returns using social media. *PLoS ONE* 2018, *13*, 2018. https://doi.org/10.1371/journal.pone.0208119.
12. Fischer, B. Studies of stock price volatility changes. In Proceedings of the American Statistical Association Business and Economic Statistics Section, Washington, DC, USA, 1976; pp. 177–181.
13. Schwert, G.W. The Adjustment of Stock Prices to Information About Inflation. *J. Financ.* 1981, *36*, 15–29. https://doi.org/10.1111/j.1540-6261.1981.tb03531.x.
14. Pearce, D.K.; Roley, V.V. Stock Prices and Economic News. *J. Bus.* 1985, *58*, 49–67. https://doi.org/10.1086/296282.
15. Cutler, D.M.; Poterba, J.M.; Summers, L.H. What moves stock prices? *J. Portf. Manag.* 1989, *15*, 4–12. https://doi.org/10.3905/jpm.1989.409212.
16. McQueen, G.; Roley, V. Stock Prices, News and Business Conditions. *Rev. Financ. Stud.* 1993, *6*, 683–708. https://doi.org/10.1093/rfs/6.3.683.
17. Kahneman, D.; Tversky, A. Prospect Theory: An Analysis of Decision Under Risk. *Econometrica* 1979, *47*, 263–291. https://doi.org/10.2307/1914185.
18. Al Nasseri, A.; Ali, F.M.; Tucker, A. Good news and bad news: Do online investor sentiments reaction to return news asymmetric? *MIDAS* 2016, *1774*, 55–66.
19. Ruan, Y.; Durresi, A.; Alfantoukh, L. Using Twitter trust network for stock market analysis. *Knowl.-Based Syst.* 2018, *145*, 207–218. https://doi.org/10.1016/j.knosys.2018.01.016.
20. Sun, W.; Cui, K. Linking corporate social responsibility to firm default risk. *Eur. Manag. J.* 2014, *32*, 275–287. https://doi.org/10.1016/j.emj.2013.04.003.
21. Zhang, X.; Fuehres, H.; Gloor, P.A. Predicting Stock Market Indicators Through Twitter 'I hope it is not as bad as I fear'. *Procedia Soc. Behav. Sci.* 2011, *26*, 55–62. https://doi.org/10.1016/j.sbspro.2011.10.562.
22. Yang, S.Y.; Mo, S.Y.K.; Liu, A. Twitter financial community sentiment and its predictive relationship to stock market movement. *Quant. Financ.* 2015, *15*, 1–27. https://doi.org/10.1080/14697688.2015.1071078.
23. Nisar, T.M.; Yeung, M. Twitter as a tool for forecasting stock market movements: A short-window event study. *J. Financ. Data Sci.* 2018, *4*, 101–119. https://doi.org/10.1016/j.jfds.2017.11.002.
24. Pagol, V.S.; Challa, K.N.R.; Panda, G. Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. *Int. Conf. Signal Process. Commun. Power Embed. Syst.* 2016. Available online: https://arxiv.org/pdf/1610.09225.pdf (accessed on 25 Feb 2021).
25. García-Medina, A.; Junior, L.S.; Bañuelos, E.U.; Martínez-Argüello, A.M. Correlations and flow of information between the New York Times and stock markets. *Phys. Stat. Mech. Its Appl.* 2018, *502*, 403–415. https://doi.org/10.1016/j.physa.2018.02.154.
26. Daradkeh, M.K. A Hybrid Data Analytics Framework with Sentiment Convergence and Multi-Feature Fusion for Stock Trend Prediction. *Electronics* 2022, *11*, 250. https://doi.org/10.3390/electronics11020250.

27. Koukaras, P.; Nousi, C.; Tjortjis, C. Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning. *Telecom* 2022, *3*, 358–378. https://doi.org/10.3390/telecom3020019.
28. Maqsood, H.; Maqsood, M.; Yasmin, S.; Mehmood, I.; Moon, J.; Rho, S. Analyzing the Stock Exchange Markets of EU Nations: A Case Study of Brexit Social Media Sentiment. *Systems* 2022, *10*, 24. https://doi.org/10.3390/systems10020024.
29. Yang, H.; Ryu, D. Investor Sentiment and Price Discrepancies between Common and Preferred Stocks in Korea. *Sustainability* 2021, *13*, 5539. https://doi.org/10.3390/su13105539.
30. Guijarro, F.; Moya-Clemente, I.; Saleemi, J. Liquidity Risk and Investors' Mood: Linking the Financial Market Liquidity to Sentiment Analysis through Twitter in the S&P500 Index. *Sustainability* 2019, *11*, 7048. https://doi.org/10.3390/su11247048.
31. López-Cabarcos, M.Á.; Pérez-Pico, A.M.; López-Pérez, M.L. Does Social Network Sentiment Influence S&P 500 Environmental & Socially Responsible Index? *Sustainability* 2019, *11*, 320. https://doi.org/10.3390/su11020320.
32. Dogra, V.; Singh, A.; Verma, S.; Alharbi, A.; Alosaimi, W. Event Study: Advanced Machine Learning and Statistical Technique for Analyzing Sustainability in Banking Stocks. *Mathematics* 2021, *9*, 3319. https://doi.org/10.3390/math9243319.
33. Siev, S. The Rich Get Richer and the Poor Get Poorer: Social Media and the Post-IPO Behavior of Investors in Biotechnology Firms: The Relationship with Twitter Volume. *J. Risk Financ. Manag.* 2021, *14*, 456. https://doi.org/10.3390/jrfm14100456.
34. Thomas, D.; Julia, P.F. Using transfer entropy to measure information flows between financial markets. *Stud. Nonlinear Dyn. Econom.* 2013, *17*, 85–102.
35. Fama, E. Efficient capital market: A review of theory and empirical work. *J. Financ.* 1970, *25*, 383–417.
36. Scaramozzino, R.; Cerchiello, P.; Aste, T. Information Theoretic Causality Detection between Financial and Sentiment Data. *Entropy* 2021, *23*, 621. https://doi.org/10.3390/e23050621.
37. Olbryś, J.; Ostrowski, K. An Entropy-Based Approach to Measurement of Stock Market Depth. *Entropy* 2021, *23*, 568. https://doi.org/10.3390/e23050568.
38. Zhang, W.-G.; Liu, Y.-J.; Xu, W.-J. A possibilistic mean-semivariance-entropy model for multi-period portfolio selection with transaction costs. *Eur. J. Oper. Res.* 2012, *222*, 341–349.
39. Usta, I.; Kantar, Y.M. Mean-variance-skewness-entropy measures: A multi-objective approach for portfolio selection. *Entropy* 2011, *13*, 117–133.
40. Xu, Y.; Wu, Z.; Jiang, L.; Song, X. A maximum entropy method for a robust portfolio problem. *Entropy* 2014, *16*, 3401–3415.
41. Yue, P.; Fan, Y.; Batten, J.A.; Zhou, W.-X. Information Transfer between Stock Market Sectors: A Comparison between the USA and China. *Entropy* 2020, *22*, 194. https://doi.org/10.3390/e22020194.
42. Balasudarsun, N.L.; Ghosh, B.; Mahendran, S. Impact of Negative Tweets on Diverse Assets during Stressful Events: An Investigation through Time-Varying Connectedness. *J. Risk Financ. Manag.* 2022, *15*, 260. https://doi.org/10.3390/ jrfm15060260.

43. García-Medina, A.; Luu Duc Huynh, T. What Drives Bitcoin? An Approach from Continuous Local Transfer Entropy and Deep Learning Classification Models. Entropy 2021, 23, 1582. https://doi.org/10.3390/e23121582
44. Mendoza Urdiales, R.A.; García-Medina, A.; Nuñez Mora, J.A. Measuring information flux between social media and stock prices with Transfer Entropy. *PLoS ONE* 2021, *16*, e0257686. https://doi.org/10.1371/journal.pone.0257686.