**Budapest University of Technology and Economics**

Institute of Mathematics

Department of Stochastics

# Relationship Between Financial Tweets and Stock Market Data

Master thesis

Panni Homolya

**Supervisors:**

**Róbert Beck, PhD**

2023

# Acknowledgement

First and foremost, I would like to express my gratitude to Róbert Beck whose expertise, dedication and enthusiasm showed the way in numerous cases. Without his help this thesis would have never been accomplished.

I would like to thank my friends for all the years of support and help in all aspects of my life.

Furthermore, I am thankful for those who make valuable comments and remarks on my work.

Last but not least, I am very grateful to my family for providing me a stable background in my personal life, helping me throughout my studies, with patience and understanding and supporting me for long years.

# Contents

# Introduction

Nowadays, social media platforms have become widespread, where people can share their thoughts and opinion on specific topics. One such platform that everyone knows is *Twitter*[1], which is a platform where you can read a wide range of posts, some of which are about specific companies, business people or events.

In the news, you often hear the name Elon Musk, who is the Chief Executive Officer of Twitter, Tesla and other impactful companies. Year after year articles are released about him influencing the stock market movements with his tweets. It is a legitimate question whether the tweets are indeed accurate and if they have the capacity to affect the stock market.

Stock market prediction is a well-researched topic as it can offer a huge business advantage for investors. Finding a stable relationship between any social media post and the stock market, means you have found the key step to predicting the stock market. If you could predict the stock market, it would change the entire business world.

In this thesis, we are focusing on investigating the connection between tweets and the stock market, but we briefly look at prediction. We use mathematical tools to explore possible links between tweets and stock market data. Several methods have been used to get closer to a solution, such as

[1] `https://twitter.com`

Machine Learning or Neural Networks techniques.

In this thesis, we briefly overview the relevant literature in Chapter 1. The data is presented in Chapter 2, then the different data manipulation techniques and the sentiment score mapping is discussed in Chapter 3. In the Chapter 4, our methods are presented such as *Granger-causality test*[1], *Augmented Dickey-Fuller test* (ADF)[2], *Kwiatkowski-Phillips-Schmidt-Shin-test* (KPSS)[3], *Spearman rank correlation*[4] and *Random Forest*[5]. In Chapter 5 our implementation is shown and Chapter 6 summarizes our results.

# Chapter 1

# Related studies

Machine Learning techniques are common methods for stock price prediction. Shen et al.[6] predicts the next-day stock trend with the aid of *Suppor Vector Machine* (SVM). By leveraging the temporal correlation between global stock markets and various financial products, they are able to capitalize on market trends. Using SVM, US stocks can be predicted with an accuracy above 70%. Further investigation of the SVM concept in [7] presented a new approach to mitigate stock market risks involves predicting a stock's returns, using *Random Forest* to predict stock market trends. Accuracy values above 80% were obtained for long-term predictions. The [8] also uses a Random Forest to predict the next-day closing price of a stock, but also tries an *Artificial Neural Network* (ANN) technique. Based on their research, ANN has performed better, than Random Forest model on financial data from *Yahoo! finance*[2].

Further *Neural Network* implementation have been introduced by [9], which tried to predict the up and down movement of stock price using *S&P500 index* by comparing *Multi-layer Perceptron*(MLP), *Convolutional*

[2] `https://finance.yahoo.com`

*Neural Network*(CNN) and *Long Short-Term Memory* (LSTM) methods. In their research, they have found that the Neural Network cannot predict the exact future values of interest, but they can follow the trend.

Stock price prediction is a difficult subject and many perspectives have been tried to tackle it. One of the most researched fields is trying to extract emotions from texts containing financial data and looking for a relationship with the stock price value, and even applying it to predictions for better results. In [10], *Autoregressive Integrated Moving Average*(ARIMA) and combined *Recurrent Neural Nework* (RNN) and *Long Short Term Memory* (LSTM) nets are investigated, however in this article sentiment derived from financial news articles is also included in the model. The best performance is achieved by the combined RNN-LSTM model. Furthermore, they have found that there is a correlation between the textual information and stock price direction.

The relationship between emotions and stock price is also discussed in [11]. Examine correlation and Granger causality test, use emotions to predict log return values. Similarly, [12] and [13] use Granger causality to find a relationship between the stock market and social media emotions and then try to predict stock market performance based on sentiments. The novelty of their approach is that they conduct a comprehensive analysis of various mood indicators obtained from a wide range of data sources. In [12], it has been found that certain moods, such as calmness, Granger cause Dow Jones Industial Average (DJIA)[3].

Texts can be assigned emotions in different ways, for example in [12] they used *OpinionFinder* and *Google-Profile of Mood States* tracking tools. Assigning emotions to financial tweets using language models, such as a version

[3] It represents the price-weighted average performance of 30 large, publicly traded companies listed on stock exchanges in the United States.

of the BERT model[14], *FinBERT*[15], is also used.

Language models process and generate human-like text. They are trained on large datasets to understand and produce coherent and contextually relevant responses across various topics.

BERT (*Bidirectional Encoder Representations from Transformers*) is a popular language model developed by Google. It utilizes a transformer architecture and it is trained on a massive amount of text data. BERT has the ability to capture bidirectional context by considering both the preceding and following words in a sentence, allowing for a deeper understanding of language.

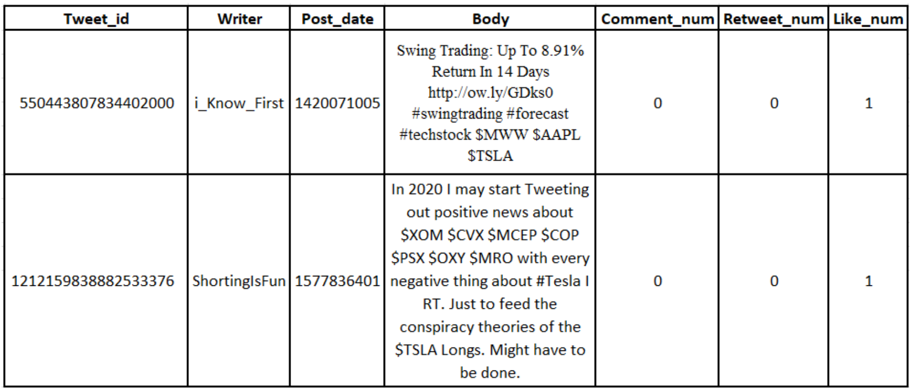
FinBERT is a specialized variant of BERT that is specifically trained for financial text analysis. It has been fine-tuned on financial domain-specific data and is designed to understand and analyze financial language, including news articles, earnings reports, and other financial documents.

The above-mentioned studies represent only a handful of the the wide ranging spectrum sentiment predection literature. We highlighted techniques beneficial for sentiment predection and the most impactful publications from the field. All of these papers have been used as a source of input for our research.

# Chapter 2

# Data

In the thesis we work with two datasets which are downloaded from *Kaggle*[4] and in the course of the thesis we will refer to them as Twitter data and stock market data.. The two datasets[5,6] were created for the *2020 IEEE International Conference on Big Data under the 6th Special Session on Intelligent Data Mining.*



| Tweet_id | Writer | Post_date | Body | Comment_num | Retweet_num | Like_num |
|---|---|---|---|---|---|---|
| 550443807834402000 | i_Know_First | 1420071005 | Swing Trading: Up To 8.91% Return In 14 Days http://ow.ly/GDks0 #swingtrading #forecast #techstock $MWW $AAPL $TSLA | 0 | 0 | 1 |
| 1212159838882533376 | ShortingIsFun | 1577836401 | In 2020 I may start Tweeting out positive news about $XOM $CVX $MCEP $COP $PSX $OXY $MRO with every negative thing about #Tesla I RT. Just to feed the conspiracy theories of the $TSLA Longs. Might have to be done. | 0 | 0 | 1 |

Figure 2.1: Representetive example from Twitter data[5]

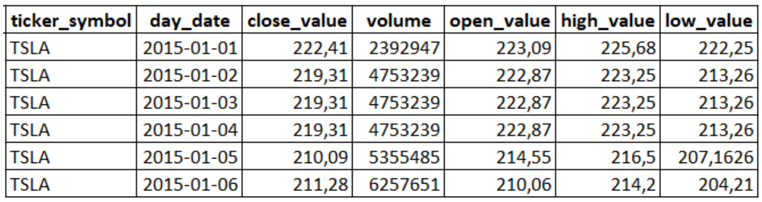
[4] https://www.kaggle.com/

[5] https://www.kaggle.com/code/tommyupton/twitter-stock-market-sentiment-analysis/data

[6] https://www.kaggle.com/datasets/omermetinn/values-of-top-nasdaq-copanies-from-2010-to-2020

We opted for Twitter datasets containing the activity record of high-end companies listed on the stock market. It is a rich and various pool of financial content rather than general tweets, which is a necessary condition for our research. As it included a dataset containing stock market data, we did not need to collect data from other sources. Throughout this thesis we only investigated the data belonging to the 6 following companies: Amazon, Apple, GOOG, GOOGL[7], Microsoft and Tesla.

The first dataset includes financial tweets which has been collected between 2015 and 2020. This dataset contains more than three million unique tweets with information such as the tweet ID(*tweet_ id*), the author of the tweet(*writer*), the date of the post(*post_ date*), the tweet (*body*), and the number of comments(*comment_ num*), retweets(*retweet_ num*)and likes (*like_ num*) of tweets associated with a given company. Figure 2.1 shows a sample of the tweet data, which contains the previously discussed columns.



| ticker_symbol | day_date | close_value | volume | open_value | high_value | low_value |
|---|---|---|---|---|---|---|
| TSLA | 2015-01-01 | 222,41 | 2392947 | 223,09 | 225,68 | 222,25 |
| TSLA | 2015-01-02 | 219,31 | 4753239 | 222,87 | 223,25 | 213,26 |
| TSLA | 2015-01-03 | 219,31 | 4753239 | 222,87 | 223,25 | 213,26 |
| TSLA | 2015-01-04 | 219,31 | 4753239 | 222,87 | 223,25 | 213,26 |
| TSLA | 2015-01-05 | 210,09 | 5355485 | 214,55 | 216,5 | 207,1626 |
| TSLA | 2015-01-06 | 211,28 | 6257651 | 210,06 | 214,2 | 204,21 |

Figure 2.2: Stock market data of Tesla

The second set of data contains stock market values. The financial data is collected from the *NASDAQ*[8] news site between 2010 and 2020. This dataset contains dates, the daily open(*open_ value*), close(*close_ value*), volume( *volume*), high(*high_ value*) and low(*low_ value*) values. Figure 2.2 shows the Tesla sample data. The *ticker_ symbol* identifies the stock listing (in this case, it is Tesla).

[7] Google has two different shares which will be referred to as GOOG and GOOGL in the thesis.

[8] `https://www.nasdaq.com/`

# Chapter 3

# Data manipulation

In the following Chapter, data processing pipelines for both datasets are described in detail.

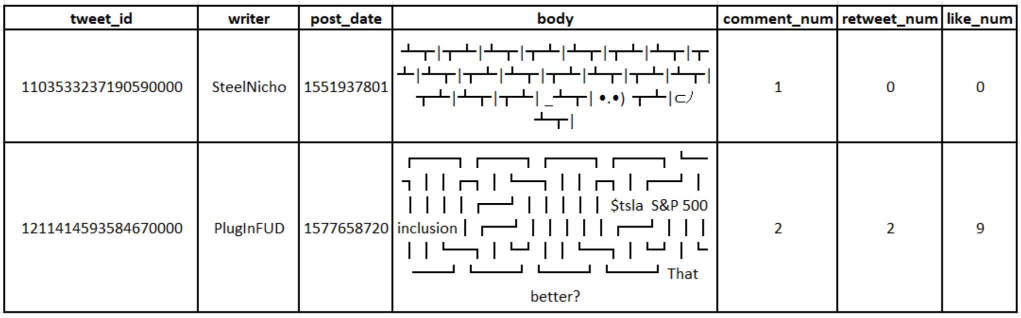
## 3.1 Twitter data processing

### 3.1.1 Data cleaning

When processing textual data, the most common data cleaning steps include emoji, link and so-called *stopwords*[9] filtering.

During the work, we only removed links, as the FinBERT model described in Section 3.1.2 contains preprocessing steps based on the model documentation, so we did not perform any other preprocessing steps.

[9] The *stopwords* of the language are a list of words that do not have any meaning or semantic relevance.



| tweet_id | writer | post_date | body | comment_num | retweet_num | like_num |
|---|---|---|---|---|---|---|
| 1103533237190590000 | SteelNicho | 1551937801 | ┴┬\|┬┴\|┴┬\|┬┴\|┴┬\|┬┴\|┴┬\|┬<br>┴\|┴┬\|┬┴\|┴┬\|┬┴\|┴┬\|┬┴\|┴┬\|<br>┬┴\|┴┬\|┬┴\| _┴┬\| •.•) ┬┴\|⊂ノ<br>┴┬\| | 1 | 0 | 0 |
| 1211414593584670000 | PlugInFUD | 1577658720 | ┌──┐ ┌──┐ ┌──┐ ┌──┐ └─<br>┐ \| \| ┌┐ \| └──┐ \| \| ┌┐ \| ┌──┘ \|<br>\| \| \| \| ┌──┘ \| \| \| \| \| $tsla S&P 500<br>inclusion \| ┌──┘ \| \| \| \| \| ┌──┘ \| \| \|<br>\| \| └──┐ \| └┘ \| \| └──┐ \| └┘ \| └<br>──┘ └──┘ └──┘ └──┘ That<br>better? | 2 | 2 | 9 |

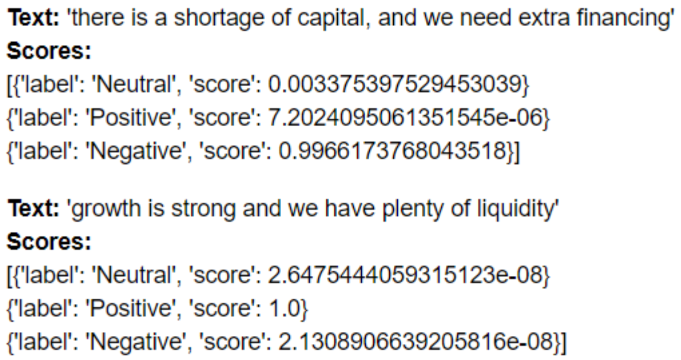
Figure 3.1: Cases, where the body of the tweets contain little text or manifests excessive use of emojies, have been excluded.

During the data discovery process, we checked the language of the tweets in the available data using *langdetect Python package*. After exploration, we found tweets that were not in English, and tweets with little or no text, as shown in Figure 3.1. Since we would not necessarily be able to interpret non-English tweets later on, and we would not be able to associate emotions to tweets without text using the model explained in Section 3.1.2, these tweets were removed.

### 3.1.2 Sentiment scores with FinBERT model

A key step in the project was to associate emotions with tweets. To do this, FinBERT(*Financial Bidirectional Encoder Representations from Transformers*) model was used. FinBERT model assigns *positive*, *negative* and *neutral* labels and values to texts. The sentiment values indicate the probability of positive, negative or neutral that a given text will be.

Figure 3.2 shows an example of the FinBERT model output. The figure shows that the first text is interpreted as mostly negative and only a small probability associates the other two emotions with it.



```
Text: 'there is a shortage of capital, and we need extra financing'
Scores:
[{'label': 'Neutral', 'score': 0.003375397529453039}
{'label': 'Positive', 'score': 7.2024095061351545e-06}
{'label': 'Negative', 'score': 0.9966173768043518}]

Text: 'growth is strong and we have plenty of liquidity'
Scores:
[{'label': 'Neutral', 'score': 2.6475444059315123e-08}
{'label': 'Positive', 'score': 1.0}
{'label': 'Negative', 'score': 2.1308906639205816e-08}]
```

Figure 3.2: Output of FinBERT model

A pre-trained FinBERT[10] model was used to tag the text data. The model was built on the following three financial communication corpora: corporate reports, earning transcripts and analyst reports. Furthermore, the model was fine-tuned with 10000 manually annotated (positive, negative, neutral) sentences from analyst reports. As the model was trained on a fairly large amount of textual data, we did not consider it necessary to further or even re-train it on our data.

### 3.1.3 Data aggregation

By performing the two data preparation steps detailed above, there are multiple tweets for a given day, and any of the three emotions can occur more than once. Since our stock market data is broken down into days rather than minutes like the Twitter dataset, it was necessary to aggregate the data.

First, for each tweet, we determined the emotion with the highest value, which we considered to be the actual tag of the tweet.

Second, the dataset is split into two parts, depending on whether the value defined above reached at least 0.8. If it does, then for those tweets we

[10] `https://huggingface.co/yiyanghkust/finbert-tone`

can say with reasonable confidence that the correct emotion was associated by the model. If the tweet has a value less than 0.8, it may be the case that the tweet contains the same proportion of multiple emotions. In that case, it is difficult to decide which is the most relevant tag for the given tweet. This uncertainty may give false results in subsequent analyses, so this data was filtered out. For the rest of the thesis, we only worked with data above 0.8 score.

As a final step, the emotions were aggregated for a given day. In aggregation, we took the positive (or negative) rate for that day, i.e. the number of tweets tagged as positive for that day, divided by the total number of tweets for that day. From now on, total number of tweets will be referred to as *tweet volume*.

For the rest of the thesis, the above detailed sentiment dataset will be used.

## 3.2 Stock market data processing

During data exploration, we checked wether the data contained unwanted values or missing elements, but the data were correct.

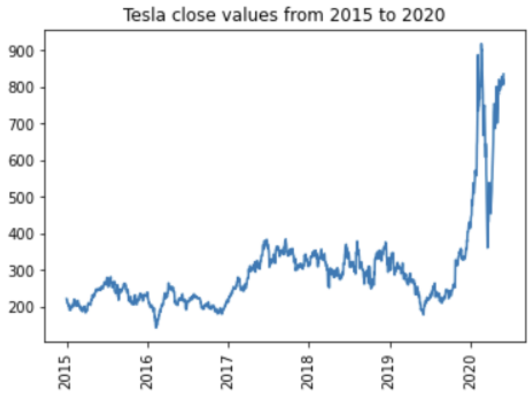


Figure 3.3: Curve of Tesla's closing values[11]

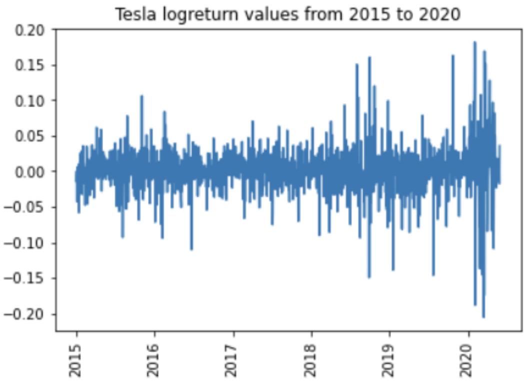


Figure 3.4: Tesla's logreturn curve

Figure 3.3 shows the curve of *Tesla*'s closing values, manifesting a clear upward trend from year to year. However, we cannot directly compare values from 2015 with 2019, as the price level in 2015 was not the same as in 2019. Therefore, we take the logarithm of the closing values and take the day by day difference[12] to obtain the curve shown in Figure 3.4. With this data manipulation we remove the price levels and focus on the relative changes, which tend to exhibit more stationary behavior. It allows for better statistical analysis[13], modeling, and comparison of assets, as it focuses on the percentage changes and captures the relative performance of investments.

[11] As there have been two share splits in the recent years, shareholders own 15 Tesla stock today for every 1 stock they owned in 2020. For this reason, the stock price nowadays is one fifteenth of what we have in the dataset visualized on Figure 3.3.

[12] From now on, we will refer to this value as logreturn

[13] Logreturns are often assumed to be normally distributed, or close to it, due to the Central Limit Theorem.

# Chapter 4

# Methodology

This Chapter describes the different methodologies we used for own work. First, we introduce time series and present the statistical tests associated with our time series analysis. Then, we discuss the correlation method used for our second approach. Finally, the model used for prediction and the metrics needed for evaluation are discussed.

## 4.1 Time series

In our case, both datasets are time series. To understand the statistical tests discussed in more detail in this Chapter, we need to introduce some definitions about time series.

Please note that basic definitions and notations, routinely used in statistics terminology, are beyond the scope of this thesis to be defined. The reader is directed to any introductory statistics textbook for further clarification.

**Definition 1.** *A time series is a sequence of data points arranged in chronological order based on their corresponding time indices, which can be written as $\{Y_t\}$ where $t = 1, 2, \ldots, T$.*

**Definition 2.** *An **autoregressive model of order p**, denoted AR(p), can be written as:*

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t$$

*where $Y_t$ is the value of the time series at time* t*, $\alpha$ is a constant term, $\phi_1, \phi_2, \ldots, \phi_p$ are the autoregressive coefficients, which represent the effect of past values on the current value and $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ i.i.d*

**Definition 3.** *The **autocovariance** of a series $Y_t$ is defined as*

$$\gamma_j = cov(Y_t, Y_{t-j})$$

**Definition 4.** *The **autocorrelation** of a series $Y_t$ is defined as*

$$\rho_j = \frac{\gamma_j}{var(Y_t)}$$

**Definition 5.** *If $Y_t$ is an AR(p) process, then it is called a process with a **unit root** if at least one root of the inverse characteristic equation $1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p = 0$ equals 1.*

**Definition 6.** *The pattern of autocorrelation as a function of lag is called **autocorrelation function**.*

Most real-life time series will contain the following patterns:

- ***Trend***: it is the overall long-term direction of the series.
- ***Seasonality***: there is repeated behavior in the data which occurs at regular intervals.
- ***Cyclicity***: a series follows an up and down pattern that is not seasonal.

The time series which does not contain the above properties is called *stationary.*

**Definition 7.** *A time series is **stationary** if it satisfies the following conditions:*

1. *The mean of the time series is constant over time.*
2. *The variance of the time series is constant over time.*
3. *The autocorrelation function of the series is constant over time*[14]*.*

**Definition 8.** *Let $Y_t$ be a time series. It is **trend-stationary** if it can be written as:*

$$Y_t = \mu_t + \epsilon_t$$

*where $\mu_t$ is a deterministic mean trend, and $\epsilon_t \sim N(0, \sigma_\epsilon^2)$.*

**Definition 9.** *If $Y_t$ is not trend-stationary, but $\Delta Y_t = Y_t - Y_{t-1}$ are stationary, then it is called **difference stationary**.*

## 4.2 Granger causality

The Granger causality test is a statistical hypothesis test to determine whether one time series is useful in predicting another time series. Instead of testing whether X causes Y, Granger causality tests whether X predicts Y.

Mathematically, it can be written in the following way.

**Definition 10.** *$X_t$ **Granger causes** $Y_t$ if $X_t$ helps to forecast $Y_t$, given past $Y_t$.*

[14] This means that the relationship between the values of the series at different time points is the same regardless of when the time points occur.

Let $X_t$ and $Y_t$ be two time series.

$$X_t = c_1 + \sum_{j=1}^{n} \alpha_j X_{t-j} + \epsilon_t \tag{4.1}$$

$$X_t = c_2 + \sum_{j=1}^{n} \alpha_j X_{t-j} + \sum_{j=1}^{n} \beta_j Y_{t-j} + \rho_t \tag{4.2}$$

where $c_1, c_2$ are constant terms, $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ and $\rho_t \sim N(0, \sigma_\rho^2)$. $Y_t$ does not Granger cause $X_t$ if $\beta = (\beta_{\mathbf{1}}, \ldots, \beta_{\mathbf{n}})^T$ is the zero vector.

The null hypothesis of the test is

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_n = 0$$

If the test statistic is greater than the 5% critical value for an $F(p, T-2p-1)$ distribution, then we reject the null hypothesis, which means $Y$ Granger causes $X$.

## 4.3 Augmented Dickey-Fuller test

In statistics, the Augmented Dickey-Fuller (ADF) test is applied to decide on the existence a unit root in the time series. If the null hypothesis fails, the alternative hypothesis of stationary time series holds.

Mathematically, it can be written as follows.

Using the Definition 2 and substituting the Definition 9, we get the below equation:

$$\Delta Y_t = \alpha + \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \ldots + \gamma_{p-1} \Delta Y_{t-p+1} + \epsilon_t \tag{4.3}$$

where $\rho = \phi_1 + \ldots + \phi_p - 1$, and $\gamma_j = -(\phi_{j+1} + \ldots + \phi_p), j = 1, \ldots, p-1$.

The null hypothesis is

$$H_0 : \rho = 0$$

We reject the null hypothesis if the *p-value* is less than 5%.

## 4.4 Kwiatkowski-Phillips-Schmidt-Shin test

In econometrics, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test describes whether a time series is stationary around a deterministic trend. If it is stationary, the null hypothesis is accepted otherwise it is rejected.

Assume that the series is decomposed into the sum of a deterministic trend, a random walk and a stationary error:

$$Y_t = \xi t + r_t + \epsilon_t \tag{4.4}$$

where $r_t$ is a random walk

$$r_t = r_{t-1} + u_t$$

where the $u_t$ are i.i.d $N(0, \sigma_u^2)$, and $\epsilon_t$ are i.i.d $N(0, \sigma_\epsilon^2)$. The intercept is $r_0$.

The null hypothesis is

$$H_0 : \sigma_u^2 = 0$$

The special case of 4.4 equation is that let $\xi = 0$. Then the null hypothesis will be that $Y_t$ is stationary around a level $r_0$.

It is important to note that the KPSS test is usually used in conjunction with the ADF test, as it may be the case that the time series is non-stationary, with no unit root, but still trend-stationary. So if we only use the ADF test, we might falsely claim that our time series is non-stationary, even though it could still be trend-stationary.

## 4.5 Spearman rank correlation

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an independent, 2-dimensional background distribution sample. The rank numbers of the sample elements $X_i$ and $Y_i$ are denoted by $R_1, \ldots, R_n$ and $S_1, \ldots, S_n$ respectively in each sample. The empirical correlation coefficient of ranks is called the **Spearman rank cor-**

**relation coefficient**, which is the following:

$$r_{sp} = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}} = \frac{\sum_{i=1}^{n}(R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2})}{\frac{n(n^2-1)}{12}}$$

where $\bar{R}$ and $\bar{S}$ are the mean of the rank numbers.

The Spearman's rank correlation coefficient used to measure the strength and direction of the relationship between two variables. It is a non-parametric measure of correlation that evaluates the monotonic relationship between two continuous or ordinal variables. The $r_{sp}$ value ranges between $-1$ and $+1$, where $-1$ indicates a perfect negative correlation, $+1$ indicates a perfect positive correlation, and 0 indicates no correlation between the variables.

The p-value associated with the Spearman rank correlation coefficient returns the probability of observing a correlation as extreme as the one calculated, assuming there is no true correlation between the variables in the population. In other words, it tells whether the observed correlation is statistically significant or not.

If the p-value is less than the chosen significance level, the null hypothesis of no correlation between the variables can be rejected, and conclude that there is a statistically significant relationship between them. If the p-value is greater than the significance level, the null hypothesis cannot be rejected, therefore there is not enough evidence to say that there is a significant relationship between the variables.

In summary, the p-value is important because it allows for determine whether the observed correlation is statistically significant or simply due to chance.

## 4.6 Random Forest

For our prediction, Random Forest is used, schematically depicted by Figure 4.1.

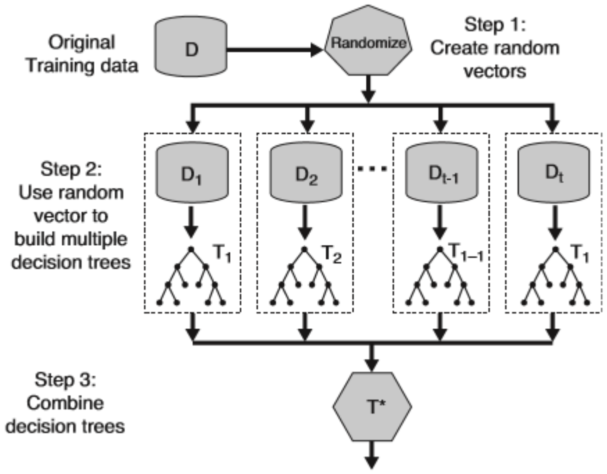


Figure 4.1: Random Forest [16, p. 278]

The original D training datsets was utilized in the so-called *bagging* method. The bagging method is the procedure itself of creating $D_t$ samples. During the bagging method, from the original $D$ dataset we draw samples with replacement and create $D_1$ dataset, which usually has the same size as $D$. We do the same method $t$ times. From the $D_1, \ldots, D_t$ sets, we build $T_1, \ldots, T_t$ decision trees in the following way. Assume that there are $n$ features of $D$. From $n$ features, randomly select $m$ input features to split at each node of the decision tree. When making a prediction, each tree in the forest independently predicts the outcome based on the input features. The predictions from all the trees are then combined by taking the average.

## 4.7 Measurement

In this section four evaluation formulas are presented. The *Mean Squared Error* (MSE) is calculated as:

$$MSE = \frac{\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}{n}$$

where $y_i$ is the $i$th observed value, $\hat{y_i}$ is the corresponding predicted value and $n$ is the number of observations.

Often a measure derived from MSE is used in the evaluation, called *Root Mean Squared Error* (RMSE). RMSE is calculated by taking the square root of the average of the squared differences between predicted and actual values:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}{n}}$$

Another formula which can be used for evaluation is *Mean Absolute Error* (MAE). MAE is calculated as the average of the absolute differences between predicted and actual values:

$$MAE = \frac{\sum_{i=1}^{n}|\hat{y_i} - y_i|}{n}$$

MAE is the simplest metric and is less sensitive to outliers than MSE and RMSE. MSE and RMSE is a commonly used metric because it penalizes larger errors more heavily than smaller ones.

The *coefficient of determination* denoted as $R^2$ is defined as:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

The $R^2$ score is a value between 0 and 1, with 1 indicating a perfect fit where the model accounts for all the variability in the target variable, and 0 indicating that the model fails to explain any of the variability. The $R^2$ score can be used to assess the importance of each feature in the model.

# Chapter 5

# Implementation

In this Chapter, our work will be explained in detail. In Section 5.1, we present a connection exploration using Spearman's rank correlation for our aggregated dataset which is written in the Subsection 3.1.3. In Section 5.2, the statistical methods are presented. In Section 5.3, using the observations of Section 5.1 and 5.2, we create a prediction model using Random Forest. If not stated otherwise, every transformation or models which will be described in this Chapter, we apply it for every company one by one. The methods will be presented in a general way, not specific to a particular company. All the results, which include company-by-company results, are described in Chapter 6.

The code used for our implementation can be found in the *GitHub repository*[15].

[15] https://github.com/homolyapanni/MSc_thesis

## 5.1 Spearman-based method (SBM)

As mentioned before, our baseline dataset will be the aggregated dataset which we merged with the stock market data described in Section 3.2.

Our main goal is to find a connection between assigned tweet sentiments and stock market data. Our first approach was to calculate Spearman's rank correlation on the merged data. We wanted to deeply explore and measure how sentiments affect the logreturn and vice versa.

To do that, we have to handle the different time resolution in our data. As it mentioned in Chapter 3, the Twitter dataset has records broken down by minutes while stock data has only with daily frequency. The stock market is open from Monday through Friday 9:30 a.m. to 4:00 p.m.. Therefore, we split the Twitter dataset into two parts. The first part is the weekend data which contains tweets in the period of Friday 4:00 p.m. through Monday 9:30 a.m.. The second dataset has records from Monday 9:30 a.m. to Friday 4:00 p.m.. After splitting, we only used the data which contains information about the weekdays.

We intended to use temporal separation to analyze both one-directional relationships of sentiments affecting logreturn values and vice versa. Thus, we manipulated our weekdays dataset in the following way. Assume that today is Monday. We aggregate the sentiment ratios only for tweets which are created after 4 p.m. and before Tuesday 9:30 a.m., let these be referred to as *tweets after closing*. We perform this for every day of the week. Let the first dataset be defined by assigning to the tweets after closing the logreturn values for the next day, in our example Tuesday. Let the second dataset be similar to the first one, but in this case we assign today's logreturn values to the tweets after closing. So Monday's tweets after closing value is assigned to the logreturn of Monday. From now on, the first dataset will be referred to as *Weekdays_ sent_ to_ stock*, the second as the *Weekdays_ stock_ to_ sent*.

We now have two datasets where, due to the separation in time, the first one expresses the sentiments affecting the logreturn and the second one expresses the other direction. After creating these two datasets, we run the Spearman's rank correlation.

## 5.2 Granger-based method(GBM)

In this Section, we discuss how the Granger causality test is run.

As in the previous Section, we use the Chapter 3 datasets. To perform the Granger causality test, we need our time series to be stationary. To check this property, the tests which were described in Sections 4.3 and 4.4 will be utilized. After testing the stationarity, we get that the sentiment ratio columns of our dataset are not stationary, which means that we have to transform them so the test could be run.

We generate a new feature by taking the daily difference of the ratio and we get the unit change day after day. We add more new ratio features for the dataset using normalization.

Let $X_t$ be our positive (or negative) ratio time series. As in [12], the new normalized $Z_t$ time series is defined as:

$$Z_t = \frac{X_t - \bar{x}(X_{t-k})}{\sigma(X_{t-k})}$$

where $\bar{x}(X_{t-k})$ and $\sigma(X_{t-k})$ represent the mean and standard deviation of the time series within the period $[t-k, t-1]$.

The variable $k$ is set to 2 weeks and $2, 6$ months back. The same $k$ is used for both ratio time series.

After computing these new features, we ran again the stationary tests and our new features became stationarity. To run Granger causality, the n lag parameter needs to be set see Section 4.2. To find the best lag, VAR (Vector Autoregression) model is used.

The VAR model is a suitable method to find the best lag for Granger causality analysis because it can handle multiple variables, account for the dynamic relationship between them and provide a framework for selecting the appropriate lag. To chose the best lag, the AIC (Akaike Information Criterion) value is used.

The AIC value is a measure of the trade-off between the goodness of fit of the model and the number of parameters used. A lower AIC value indicates that the model is either a better fit to the data, or uses fewer parameters, or both. Therefore, when comparing multiple models, the one with the lowest AIC value is generally considered the best.

Getting the best lag values based on AIC value, the Granger causality test can be run. For the VAR model and the Granger causality test, only the logreturn and the examined ratio value are given to them at the same time.

## 5.3 Prediction

In Section 5.1, we investigated the logreturn and sentiment connection, using a correlation method. Based on the results obtained, which will be described in Section 6.2.1, the greatest correlation values (around 0.3) were selected as the features for our prediction, e.g. tweet volume. Furthermore, the fundamental intuition behind Granger causality, which method is written in Section 5.2, lies in detecting the capability to forecast or anticipate from one variable the behavior of another variable. From the results of the Granger causality, which are summarized in Section 6.2.2, we obtained, which of the features we use are the ones that are actually predictive, and used them to train our prediction models. Therefore, we have decided to create a model for predicting the logreturn values, using the new stationary sentiment features, which were presented in 5.2. We have created separate models for each company, as well as a general model trained using data from all companies. As in the previous two sections, the idea is presented here, and in Section 6.2.3, the results are shown regarding the performance of the models of different companies.

In Section 5.2, we distinguished two feature groups, the ratio difference of sentiments and the normalized version of the ratio of sentiments. For prediction, one of the two feature groups, the daily number of tweets(*tweet volume*) and the financial indicators of the stock market data, such as open, low and high values are used. Based on the two feature groups, we created models which use the same transformed financial indicators. Let *features with sentiment ratio difference* and *features with normalized sentiment ratio* be the name of our two different approaches. First we introduce the transformation of financial indicators, afterwards the two approaches are presented.

Our features of sentiments are stationary as described in Section 5.2, but the open, low and high values are not. Therefore we use the same tranformation as in Section 3.2. Logreturn values are created, using open, low and high values. The new columns are *open_logreturn*, *low_logreturn* and *high_logreturn*.

In Section 5.2, we used differrent time lags for calculating the Granger causality test. Using time lags means that we are utilizing the information of the past. Similarly, for the prediction, new columns were created using the values of the *open_logreturn*, *low_logreturn* and *high_logreturn* columns, shifted from one to fifteen days. Because of this change, a given record contains the information from the past fifteen days.

For the approach of features with sentiment ratio difference, the same ratio difference was used as in Section 5.2. The sentiment ratio difference of a given record includes information of the given day and the previous day, because of the difference calculation. If we gave this record to the model, then our model should predict the logreturn of the given day based on some information of the given day. Basically, we would predict logreturns based on data that would become available at the same time as (or later than) the target. To avoid this, we use the same shifted method as above. Fifteen new columns are created from the ratio difference value. So, the features of the sentiment ratio approach are the fifteen shifted ratio columns, the fifteen shifted open, low and high logreturn columns and the tweet volume.

For the approach of features with normalized sentiment ratio, the normalized features were used which are described in Section 5.2. They include information about the previous periods of a given column and in Section 5.2 we created them with a 1 day shift, so now we do not have to shift these features. For this model, we used the six normalized ratio values, the fifteen shifted open, low and high logreturn columns and the tweet volume as features.

As customary for every Machine Learning approach, our data is split into three sets: train, validation and test set. Because we work with time series, therefore we split our data based on time. The train set contains data from 2015 to 2017. The validation set contains data from 2018 and the test set contains data from 2019. Our models are trained on the train set, and evaluated on the validation set. Our models have been improved by tuning hyperparameters based on the previous two sets, and the final evaluation was on the test set. All the results are summarized in Section 6.2.3.

# Chapter 6

# Results

## 6.1 Qualitative results

Before we get into actual mathematical tests, first we explore, through visualisation, whether it is possible to see a change between a tweet posted by an important person and a connection in the stock price. In this Section, we show our visualization results.

We screened for real events between 2015 and 2020 when the value of stock price changed dramatically following a Twitter post. After having some claims about stock price increase (or decrease) because of a tweet, we examined the claims using our datasets. In Section 5.1, we split our dataset into weekdays and weekends data. We use the weekdays data in our visualization, because the stock market is closed in the weekends. But the *Weekdays_sent_to_stock* and *Weekdays_stock_to_sent* mentioned in Section 5.1 are derived from the data set used here. In the data set used here, tweets were not filtered to the time after stock market close only.

| *In July 2017, Tesla's stock dropped after a tweet from Elon Musk suggesting that the company's new Model 3 sedan was experiencing production delays.* |
|---|

Table 6.1: Statement about Tesla in July 2017

Reading the statement in Table 6.1, we expect that the ratio of negative tweets will increase due to production delays, and therefore the stock price will decrease. Figure 6.1 shows the corresponding values of our dataset. The period in the statement is indicated by the red shaded area. The green curve which is the time series of logreturn dropped at the beginning of July. In the same time, the negative ratio curve, which is the blue curve, suddenly soared. But the red positive ratio curve also increased. It seems that, as more tweets were posted, the sentiment ratios changed and the stock price dropped.

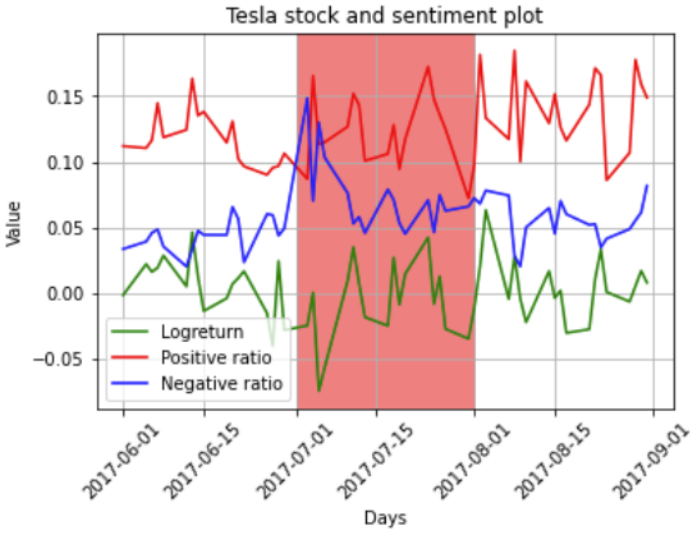


Figure 6.1: Tesla's logreturn and sentiments ratios in July 2017

| |
|---|
| *In September 2019, Tesla's stock dropped after a tweet from Elon Musk suggesting that the company's new "cybertruck" had received 146,000 pre-orders, which was lower than many analysts had expected.* |

Table 6.2: Statement about Tesla in September 2019

After reading Table 6.2, we expect a similar Figure as before. Examining Figure 6.2, we observe that the logreturn curve suddenly decreased and the negative ratio curve increased. But in the same time, the positive ratio curve also dropped. It looks like a perfect example that we expect when the stock price drops.

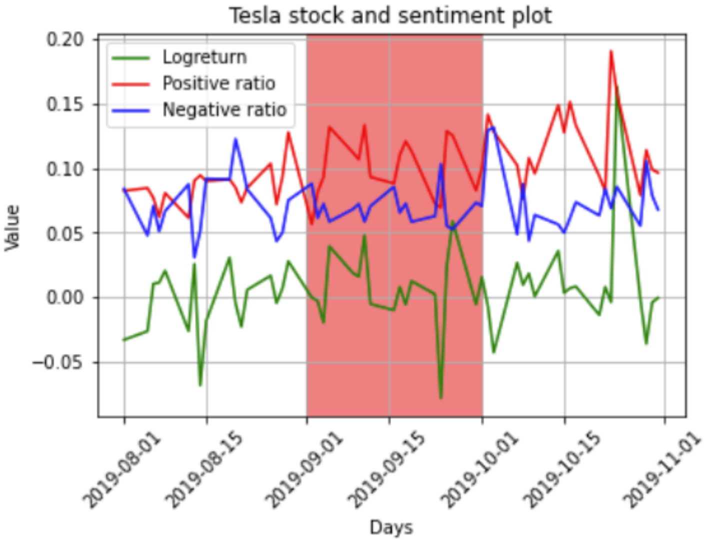


Figure 6.2: Tesla's logreturn and sentiment ratios in September 2019

The text of Table 6.3 is similar to the previous two cases, therefore we are waiting for the same output. But in this case, Figure 6.3 looks significantly different from the previous examples.

*In June 2017, Amazon's stock dropped after the company announced plans to acquire Whole Foods Market, which led to concerns about the impact on the grocery industry and potential antitrust issues.*

Table 6.3: Statement about Amazon in June 2017

We would expected a negative ratio jump, but the positive ratio curve has several peaks. On the plot, the green and blue curves appear mostly featureless. There are no peaks where they are expected. It seems there is no apparent connection between stock price and sentiment scores.

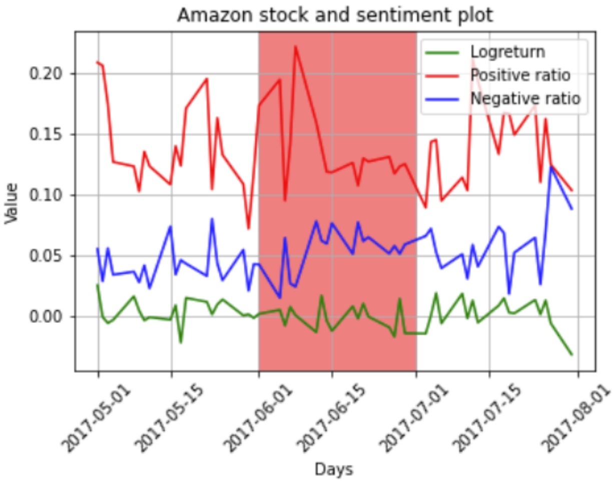


Figure 6.3: Amazon's logreturn and sentiment ratios in June 2017

In the first two Figures, we saw an example of the relationship between stock and tweets, but in the last Figure, we observed the opposite, as if there was no relationship between the time series used in the thesis. In the next Section, we will use statistical tools, not just visualisation, for a deeper understanding of whether the examples presented in this Section were just coincidence or whether there is an actual connection between the stock market and sentiments of tweets.

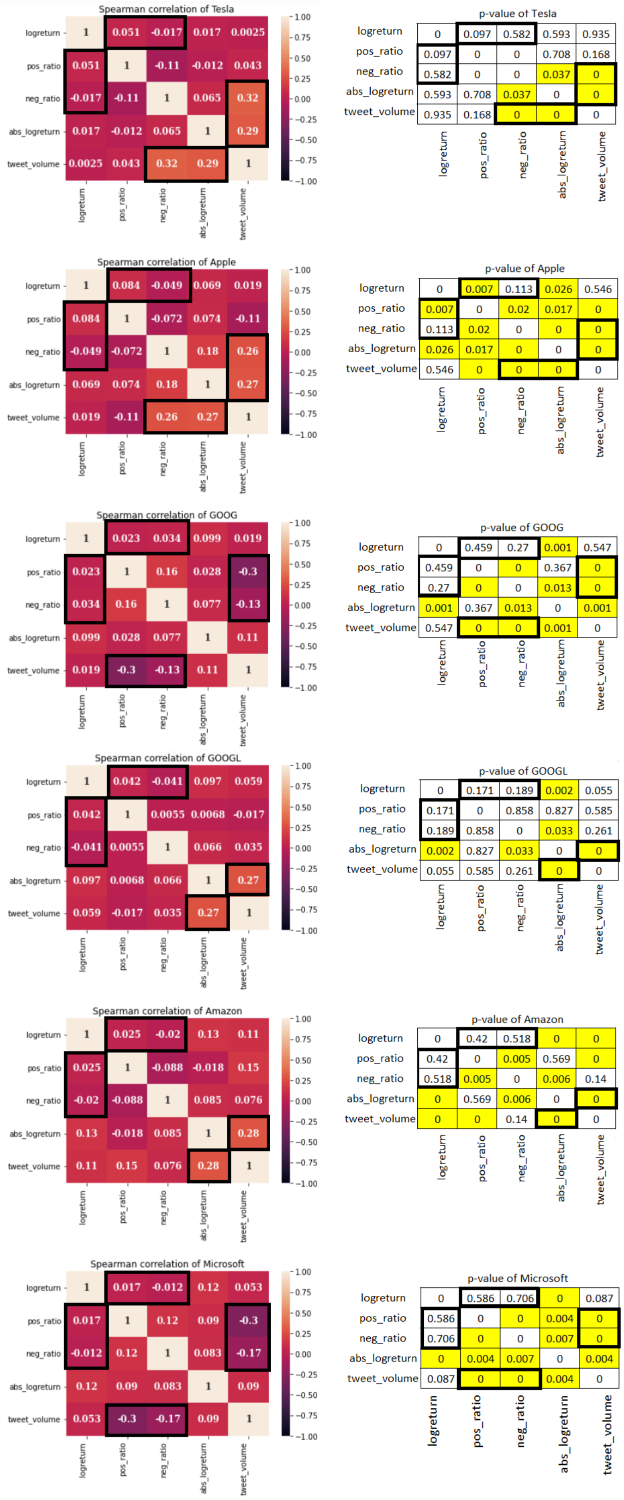
## 6.2 Quantitative results

### 6.2.1 Result of SBM

In Section 5.1, we introduced two datasets: *Weekdays_sent_to_stock* and *Weekdays_stock_to_sent*. The *Weekdays_sent_to_stock* dataset expresses that we know the sentiments of the tweets on a given day and we compare them to the next day logreturn value. The *Weekdays_stock_to_sent* expresses the reverse direction, we know a given day's logreturn value, and we compare it to the tweets which have been created after stock market closing. In this Section, we show our results which were evaluated using Spearman's rank correlation.

Figure 6.4 shows the results of the different companies' Spearman correlation and the corresponding p-values for the *Weekdays_sent_to_stock dataset*. First, we examine the different heatmaps and afterwards the tables of the p-values. There are no outstanding correlation values on the left column of Figure 6.4. Most of the values are around 0. Potential relevant connections are denoted with a black frame in the correlation tables, we detail these below.

On the plot of Tesla and Apple, there is weak positive correlation between negative ratio and tweet volume. There is a weak negative correlation between tweet volume and positive ratio, and tweet volume and negative ratio on GOOG and Microsoft heatmaps. The tweet volume correlates with the absolute logreturn on the heatmap of Tesla, Apple, GOOGL and Amazon. However, examining the potential impact of sentiment on logreturns, we observe that logreturns exhibit a correlation around 0 regardless of sentiment. This pattern holds true across all companies, indicating a lack of evidence supporting a relationship between sentiment and logreturns.



p-value of Tesla

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 0 | 0.097 | 0.582 | 0.593 | 0.935 |
| pos_ratio | 0.097 | 0 | 0 | 0.708 | 0.168 |
| neg_ratio | 0.582 | 0 | 0 | 0.037 | 0 |
| abs_logreturn | 0.593 | 0.708 | 0.037 | 0 | 0 |
| tweet_volume | 0.935 | 0.168 | 0 | 0 | 0 |

p-value of Apple

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 0 | 0.007 | 0.113 | 0.026 | 0.546 |
| pos_ratio | 0.007 | 0 | 0.02 | 0.017 | 0 |
| neg_ratio | 0.113 | 0.02 | 0 | 0 | 0 |
| abs_logreturn | 0.026 | 0.017 | 0 | 0 | 0 |
| tweet_volume | 0.546 | 0 | 0 | 0 | 0 |

Spearman correlation of GOOG
logreturn 1 0.023 0.034 0.099 0.019
pos_ratio 0.023 1 0.16 0.028 -0.3
neg_ratio 0.034 0.16 1 0.077 -0.13
abs_logreturn 0.099 0.028 0.077 1 0.11
tweet_volume 0.019 -0.3 -0.13 0.11 1
logreturn pos_ratio neg_ratio abs_logreturn tweet_volume
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00

p-value of GOOG

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 0 | 0.459 | 0.27 | 0.001 | 0.547 |
| pos_ratio | 0.459 | 0 | 0 | 0.367 | 0 |
| neg_ratio | 0.27 | 0 | 0 | 0.013 | 0 |
| abs_logreturn | 0.001 | 0.367 | 0.013 | 0 | 0.001 |
| tweet_volume | 0.547 | 0 | 0 | 0.001 | 0 |

Spearman correlation of GOOGL
logreturn 1 0.042 -0.041 0.097 0.059
pos_ratio 0.042 1 0.0055 0.0068 -0.017
neg_ratio -0.041 0.0055 1 0.066 0.035
abs_logreturn 0.097 0.0068 0.066 1 0.27
tweet_volume 0.059 -0.017 0.035 0.27 1
logreturn pos_ratio neg_ratio abs_logreturn tweet_volume
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00

p-value of GOOGL

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 0 | 0.171 | 0.189 | 0.002 | 0.055 |
| pos_ratio | 0.171 | 0 | 0.858 | 0.827 | 0.585 |
| neg_ratio | 0.189 | 0.858 | 0 | 0.033 | 0.261 |
| abs_logreturn | 0.002 | 0.827 | 0.033 | 0 | 0 |
| tweet_volume | 0.055 | 0.585 | 0.261 | 0 | 0 |

Spearman correlation of Amazon
logreturn 1 0.025 -0.02 0.13 0.11
pos_ratio 0.025 1 -0.088 -0.018 0.15
neg_ratio -0.02 -0.088 1 0.085 0.076
abs_logreturn 0.13 -0.018 0.085 1 0.28
tweet_volume 0.11 0.15 0.076 0.28 1
logreturn pos_ratio neg_ratio abs_logreturn tweet_volume
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00

p-value of Amazon

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 0 | 0.42 | 0.518 | 0 | 0 |
| pos_ratio | 0.42 | 0 | 0.005 | 0.569 | 0 |
| neg_ratio | 0.518 | 0.005 | 0 | 0.006 | 0.14 |
| abs_logreturn | 0 | 0.569 | 0.006 | 0 | 0 |
| tweet_volume | 0 | 0 | 0.14 | 0 | 0 |

Spearman correlation of Microsoft
logreturn 1 0.017 -0.012 0.12 0.053
pos_ratio 0.017 1 0.12 0.09 -0.3
neg_ratio -0.012 0.12 1 0.083 -0.17
abs_logreturn 0.12 0.09 0.083 1 0.09
tweet_volume 0.053 -0.3 -0.17 0.09 1
logreturn pos_ratio neg_ratio abs_logreturn tweet_volume
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00

p-value of Microsoft

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 0 | 0.586 | 0.706 | 0 | 0.087 |
| pos_ratio | 0.586 | 0 | 0 | 0.004 | 0 |
| neg_ratio | 0.706 | 0 | 0 | 0.007 | 0 |
| abs_logreturn | 0 | 0.004 | 0.007 | 0 | 0.004 |
| tweet_volume | 0.087 | 0 | 0 | 0.004 | 0 |

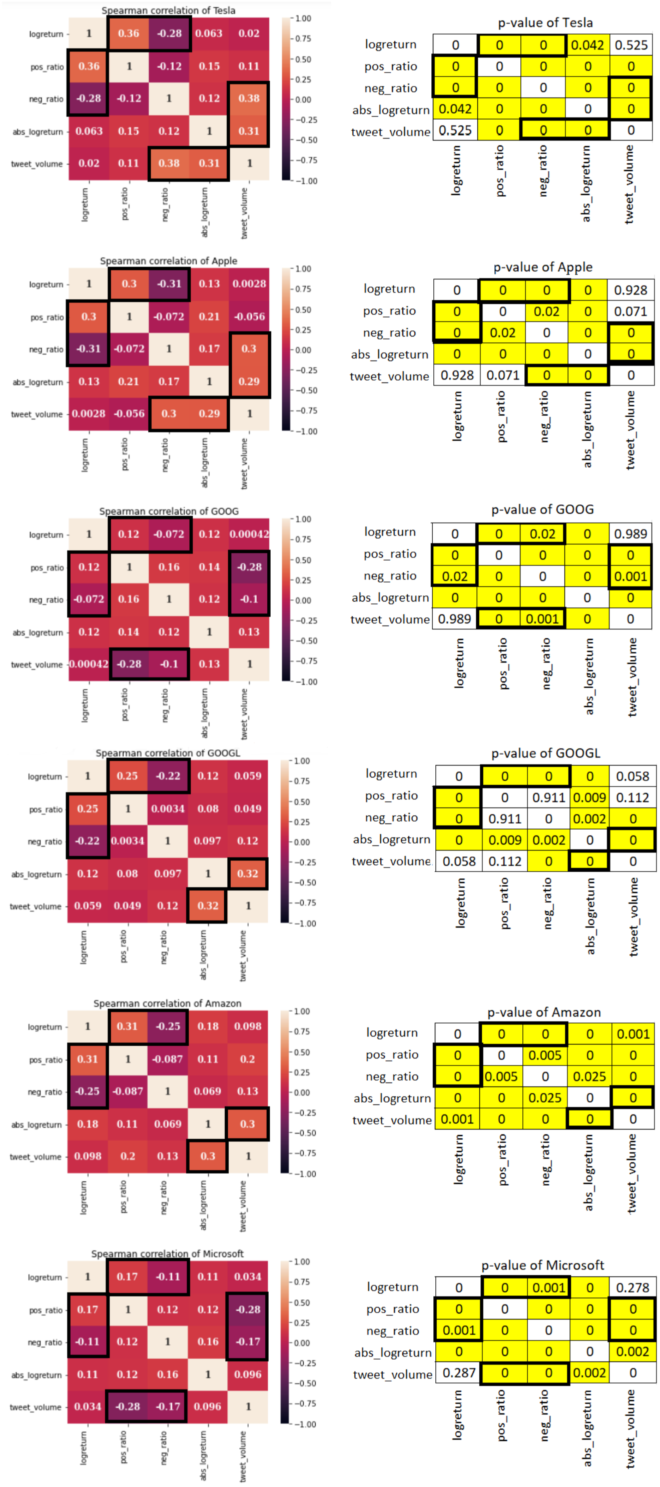
Figure 6.4: Spearman correlation and the corresponding p-values of the Weekdays_sent_to_stock dataset[16]

To quantify the robustness of these correlations, we show their p-values on the right column of Figure 6.4. The null hypothesis is that there is no correlation between the variables. If it is rejected, then the correlation is statistically significant. In Figure 6.4 right column, the cells of the companies, tables are yellow, if the p-value is less than 0.05, which means the null hypothesis is rejected. The framed yellow cells correspond to the framed correlation values of a given company, showing that the above discussed correlations are significant. For identically framed cells in a given company, where a weak correlation is assumed, is supported by the p-values in the hypothesis. Furthermore, this observation confirms that the entries within the logreturn row and column in the matrix pass the test, indicating the absence of correlation between the sentiment and logreturn variables. Only in the case of Apple's figure do we encounter conflicting results. When comparing the heatmap and the table of p-value, it becomes difficult to determine whether there is a correlation between the positive ratio and logreturn variables or not.

Figure 6.5 contains the *Weekdays_stock_to_sent* dataset correlation values and p-values. After careful visual assessment, higher correlations can be detected than on Figure 6.4. The statements derived from the previous measures on Figure 6.4 still hold, however other correlations also manifest. On the heatmap of Tesla, Apple, GOOGL and Amazon, there is a positive correlation between logreturn and positive ratio, and negative correlation between logreturn and negative ratio, which are marked with a frame.

Comparing the corresponding framed and yellow cells between the left and right columns of Figure 6.5, it can be seen that the correlations discussed above are statistically significant.

[16] The values in the tables are rounded to three decimal



Spearman correlation of Tesla

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 1 | 0.36 | -0.28 | 0.063 | 0.02 |
| pos_ratio | 0.36 | 1 | -0.12 | 0.15 | 0.11 |
| neg_ratio | -0.28 | -0.12 | 1 | 0.12 | 0.38 |
| abs_logreturn | 0.063 | 0.15 | 0.12 | 1 | 0.31 |
| tweet_volume | 0.02 | 0.11 | 0.38 | 0.31 | 1 |

p-value of Tesla

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 0 | 0 | 0 | 0.042 | 0.525 |
| pos_ratio | 0 | 0 | 0 | 0 | 0 |
| neg_ratio | 0 | 0 | 0 | 0 | 0 |
| abs_logreturn | 0.042 | 0 | 0 | 0 | 0 |
| tweet_volume | 0.525 | 0 | 0 | 0 | 0 |

Spearman correlation of Apple

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 1 | 0.3 | -0.31 | 0.13 | 0.0028 |
| pos_ratio | 0.3 | 1 | -0.072 | 0.21 | -0.056 |
| neg_ratio | -0.31 | -0.072 | 1 | 0.17 | 0.3 |
| abs_logreturn | 0.13 | 0.21 | 0.17 | 1 | 0.29 |
| tweet_volume | 0.0028 | -0.056 | 0.3 | 0.29 | 1 |

p-value of Apple

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 0 | 0 | 0 | 0 | 0.928 |
| pos_ratio | 0 | 0 | 0.02 | 0 | 0.071 |
| neg_ratio | 0 | 0.02 | 0 | 0 | 0 |
| abs_logreturn | 0 | 0 | 0 | 0 | 0 |
| tweet_volume | 0.928 | 0.071 | 0 | 0 | 0 |

Spearman correlation of GOOG

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 1 | 0.12 | -0.072 | 0.12 | 0.00042 |
| pos_ratio | 0.12 | 1 | 0.16 | 0.14 | -0.28 |
| neg_ratio | -0.072 | 0.16 | 1 | 0.12 | -0.1 |
| abs_logreturn | 0.12 | 0.14 | 0.12 | 1 | 0.13 |
| tweet_volume | 0.00042 | -0.28 | -0.1 | 0.13 | 1 |

p-value of GOOG

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 0 | 0 | 0.02 | 0 | 0.989 |
| pos_ratio | 0 | 0 | 0 | 0 | 0 |
| neg_ratio | 0.02 | 0 | 0 | 0 | 0.001 |
| abs_logreturn | 0 | 0 | 0 | 0 | 0 |
| tweet_volume | 0.989 | 0 | 0.001 | 0 | 0 |

Spearman correlation of GOOGL

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 1 | 0.25 | -0.22 | 0.12 | 0.059 |
| pos_ratio | 0.25 | 1 | 0.0034 | 0.08 | 0.049 |
| neg_ratio | -0.22 | 0.0034 | 1 | 0.097 | 0.12 |
| abs_logreturn | 0.12 | 0.08 | 0.097 | 1 | 0.32 |
| tweet_volume | 0.059 | 0.049 | 0.12 | 0.32 | 1 |

p-value of GOOGL

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 0 | 0 | 0 | 0 | 0.058 |
| pos_ratio | 0 | 0 | 0.911 | 0.009 | 0.112 |
| neg_ratio | 0 | 0.911 | 0 | 0.002 | 0 |
| abs_logreturn | 0 | 0.009 | 0.002 | 0 | 0 |
| tweet_volume | 0.058 | 0.112 | 0 | 0 | 0 |

Spearman correlation of Amazon

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 1 | 0.31 | -0.25 | 0.18 | 0.098 |
| pos_ratio | 0.31 | 1 | -0.087 | 0.11 | 0.2 |
| neg_ratio | -0.25 | -0.087 | 1 | 0.069 | 0.13 |
| abs_logreturn | 0.18 | 0.11 | 0.069 | 1 | 0.3 |
| tweet_volume | 0.098 | 0.2 | 0.13 | 0.3 | 1 |

p-value of Amazon

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 0 | 0 | 0 | 0 | 0.001 |
| pos_ratio | 0 | 0 | 0.005 | 0 | 0 |
| neg_ratio | 0 | 0.005 | 0 | 0.025 | 0 |
| abs_logreturn | 0 | 0 | 0.025 | 0 | 0 |
| tweet_volume | 0.001 | 0 | 0 | 0 | 0 |

Spearman correlation of Microsoft

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 1 | 0.17 | -0.11 | 0.11 | 0.034 |
| pos_ratio | 0.17 | 1 | 0.12 | 0.12 | -0.28 |
| neg_ratio | -0.11 | 0.12 | 1 | 0.16 | -0.17 |
| abs_logreturn | 0.11 | 0.12 | 0.16 | 1 | 0.096 |
| tweet_volume | 0.034 | -0.28 | -0.17 | 0.096 | 1 |

p-value of Microsoft

| | logreturn | pos_ratio | neg_ratio | abs_logreturn | tweet_volume |
|---|---|---|---|---|---|
| logreturn | 0 | 0 | 0.001 | 0 | 0.278 |
| pos_ratio | 0 | 0 | 0 | 0 | 0 |
| neg_ratio | 0.001 | 0 | 0 | 0 | 0 |
| abs_logreturn | 0 | 0 | 0 | 0 | 0.002 |
| tweet_volume | 0.287 | 0 | 0 | 0.002 | 0 |

Figure 6.5: Spearman correlation and the corresponding p-value of the Weekdays_stock_to_sent dataset[17]

### 6.2.2 Result of GBM

In this Section, the results of the Granger causality test will be shown. In Section 5.2, we introduced two different feature groups. One of them was the daily difference of the sentiment ratio and the other one was the $k$ days normalized sentiment ratio. With the Granger causality test, we examined if the given ratio feature predicts the logreturn and vice versa. The null hypothesis of the test is that $X$ does not Granger cause $Y$.

| **Models** | **Hypothesis test with positive ratio** | | **Hypothesis test with negative ratio** | |
|---|---|---|---|---|
| | **Lags** | **p-value** | **Lags** | **p-value** |
| Tesla | 13 | 0.312 | 15 | 0.639 |
| Apple | 14 | 0.099 | 15 | 0 |
| GOOG | 15 | 0.032 | 13 | 0.166 |
| GOOGL | 7 | 0.097 | 13 | 0.017 |
| Amazon | 14 | 0.111 | 13 | 0.160 |
| Microsoft | 13 | 0.004 | 13 | 0.011 |

Table 6.4: The p-values resulting from the hypothesis testing of whether the given sentiment ratio Granger causes the logreturn[18]

On the Table 6.4, the results of the ratio difference are summarized. The tables contain the p-values of the different companies resulting from the hypothesis test. As we mentioned in Section 5.2, in order to get the best lag, we trained a VAR model and evaluated it with AIC. We have chosen the lag of the model with the best AIC value for the Granger causality, the value of which is given in the *Lags* column in the tables. The yellow cells indicate p-values below the 0.05 significance level where our null hypothesis is rejected.

[17] The values in the tables are rounded to three decimals

[18] The values in the table are rounded to three decimals

On Table 6.4, we find that the positive ratio of the GOOG and Microsoft tweets Granger cause the logreturn. The negative ratio tweets of Apple, GOOGL and Microsoft also Granger cause the logreturn. It seems that the daily difference ratio of Tesla and Amazon with almost 2 weeks lags do not affect the logreturn.

| **Company** | **2 weeks normalization** | | **2 months normalization** | | **6 months normalization** | |
|---|---|---|---|---|---|---|
| | **Lags** | **p-value** | **Lags** | **p-value** | **Lags** | **p-value** |
| Tesla | 12 | 0.117 | 2 | 0.111 | 7 | 0.029 |
| Apple | 1 | 0.001 | 7 | 0.026 | 13 | 0.04 |
| GOOG | 10 | 0.018 | 7 | 0.003 | 7 | 0.005 |
| GOOGL | 12 | 0.552 | 7 | 0.132 | 7 | 0.058 |
| Amazon | 3 | 0.003 | 7 | 0.067 | 7 | 0.048 |
| Microsoft | 14 | 0.001 | 9 | 0 | 8 | 0 |

Table 6.5: The p-values resulting from the hypothesis testing of whether the normalized positive ratio Granger causes the logreturn[19]

Comparing the previous table and the Table 6.5, it can be detected that, thanks to several weeks and months of normalization, the logreturns of many more companies can be predicted from sentiment. In general, because the features contain information about the more distant past, the lag values are on average 40% smaller than for the ratio differences. On Table 6.5 you can see that for GOOGL all $k$ days normalized positive ratios do not Granger cause the logreturn value. Tesla is only under the significance threshold at 6 months normalization.

[19] The values in the table are rounded to three decimals

| Company | 2 weeks normalization | | 2 months normalization | | 6 months normalization | |
|---|---|---|---|---|---|---|
| | Lags | p-value | Lags | p-value | Lags | p-value |
| Tesla | 1 | 0.045 | 1 | 0.041 | 7 | 0.272 |
| Apple | 15 | 0 | 14 | 0 | 14 | 0 |
| GOOG | 1 | 0.061 | 3 | 0.023 | 5 | 0.066 |
| GOOGL | 13 | 0.003 | 3 | 0.005 | 3 | 0.004 |
| Amazon | 1 | 0.086 | 3 | 0.009 | 6 | 0 |
| Microsoft | 12 | 0.746 | 8 | 0.51 | 8 | 0.579 |

Table 6.6: The p-values resulting from the hypothesis testing of whether the normalized negative ratio Granger causes the logreturn[19]

Comparing Table 6.5 and Table 6.6, we get yellow cells in almost the opposite places. The p-values of Tesla and GOOGL are less than the given significance level on the normalized negative ratio Table 6.6, but on the previous table, the null hypothesis is not rejected. It seems that for these companies the causality relationship depends on what sentiment we look at in what period. GOOG and Amazon have values where the two tables' sentiment Granger cause logreturn in the same period. From the results, the value of the Apple logreturn seems to be well predicted from the sentiments.

| Models | Hypothesis test with positive ratio | | Hypothesis test with negative ratio | |
|---|---|---|---|---|
| | Lags | p-value | Lags | p-value |
| Tesla | 13 | 0.243 | 15 | 0.033 |
| Apple | 14 | 0.272 | 15 | 0.048 |
| GOOG | 15 | 0.024 | 13 | 0.139 |
| GOOGL | 7 | 0.331 | 13 | 0 |
| Amazon | 14 | 0.106 | 13 | 0.006 |
| Microsoft | 13 | 0.384 | 13 | 0.745 |

Table 6.7: The p-values resulting from the hypothesis testing whether the logreturn Granger causes the given sentiment ratio[20]

[20] The values in the table are rounded to three decimals

On Table 6.7, we indicate the results of the hypothesis test of the opposite direction, i.e. whether the logreturn Granger causes sentiment daily difference. The logreturn predicts the positive ratio only in the case of GOOG. On the other hand, the logreturn does not Granger cause the negative ratio for GOOG and Microsoft, but for the other companies it does. It seems that in a short period, the logreturn values mainly affect the tweets with negative tone.

| **Company** | **2 weeks normalization** | | **2 months normalization** | | **6 months normalization** | |
|---|---|---|---|---|---|---|
| | **Lags** | **p-value** | **Lags** | **p-value** | **Lags** | **p-value** |
| Tesla | 12 | 0.138 | 2 | 0.570 | 7 | 0.740 |
| Apple | 1 | 0.971 | 7 | 0.763 | 13 | 0.725 |
| GOOG | 10 | 0.055 | 7 | 0.472 | 7 | 0.476 |
| GOOGL | 12 | 0.552 | 7 | 0.361 | 7 | 0.384 |
| Amazon | 3 | 0.243 | 7 | 0.395 | 7 | 0.251 |
| Microsoft | 14 | 0.825 | 9 | 0.620 | 8 | 0.788 |

Table 6.8: The p-values resulting from the hypothesis testing of whether the logreturn Granger causes the normalized positive ratio[21]

In contrast to Table 6.7, where there was only one case, when we rejected that the logreturn does not Granger cause positive ratio, on Table 6.8, we do not discover any connection between logreturn and normalized positive ratio. The normalization contains information about the past, and with larger $k$ values, it points to more distant periods. The logreturn cannot predict long-period positive sentiments.

[21] The values in the table are rounded to three decimals

| Company | 2 weeks normalization | | 2 months normalization | | 6 months normalization | |
|---|---|---|---|---|---|---|
| | Lags | p-value | Lags | p-value | Lags | p-value |
| Tesla | 1 | 0.086 | 1 | 0.201 | 7 | 0.247 |
| Apple | 15 | 0.14 | 14 | 0.644 | 14 | 0.505 |
| GOOG | 1 | 0.664 | 3 | 0.513 | 5 | 0.463 |
| GOOGL | 13 | 0.008 | 3 | 0.155 | 3 | 0.056 |
| Amazon | 1 | 0.022 | 3 | 0.113 | 6 | 0.451 |
| Microsoft | 12 | 0.821 | 8 | 0.759 | 8 | 0.729 |

Table 6.9: The p-values resulting from the hypothesis testing of whether the logreturn Granger causes the normalized negative ratio[21]

On Table 6.7, we could see that with 2 weeks of lags, the logreturn Granger causes negative ratio. On Table 6.9, the logreturn shows Granger causality with 2 weeks normalized negative ratio values for some companies. However, as with Table 6.8, the longer period we consider in the normalized negative ratio, the less the logreturn is able to predict them.

### 6.2.3 Result of the prediction

In this Section, the results of the prediction are presented. The Table 6.10 summarizes the results of our models on the validation data set. In the Table, the results are separeted based on the two input feature approaches detailed in Section 5.3, the normalized and difference sentiment ratio. For models in a given row, the value with the better metric is highlighted in bold. The metrics used in the following tables are described in detail in Section 4.7.

| **Models** | **Features with sentiment ratio** | | | | **Features with normalized sentiment ratio** | | | |
|---|---|---|---|---|---|---|---|---|
| | **MAE** | **MSE** | **RMSE** | $R^2$ | **MAE** | **MSE** | **RMSE** | $R^2$ |
| Tesla | 0.0202 | 0.0011 | 0.0331 | -0.167 | **0.0187** | **0.0008** | **0.0297** | **0.1696** |
| Apple | 0.0096 | 0.0002 | 0.0152 | 0.0058 | **0.0082** | 0.0002 | **0.0126** | **0.3203** |
| GOOG | 0.0095 | 0.0002 | 0.0148 | 0.0273 | **0.0094** | 0.0002 | **0.0145** | **0.0633** |
| GOOGL | 0.0094 | 0.0002 | 0.0149 | 0.0222 | **0.0087** | 0.0002 | **0.0134** | **0.2020** |
| Amazon | 0.0118 | 0.0004 | 0.0193 | -0.0180 | **0.0109** | **0.0003** | **0.0174** | **0.1745** |
| Microsoft | 0.0091 | 0.0002 | 0.0147 | 0.0335 | 0.0091 | 0.0002 | **0.0145** | **0.0684** |
| General model | 0.0113 | 0.0004 | 0.0192 | 0.0047 | **0.0107** | **0.0003** | **0.0175** | **0.1735** |

Table 6.10: Validation results of the Random Forest models [22]

Examining Table 6.10, it can be seen that the models with normalised sentiment ratio features perform better. Based on the $R^2$ score of the model, which were trained on features with sentiment ratio, they fail to explain any of the variability. The $R^2$ values are higher for the other group of models. The model of Apple and GOOGL can learn from the normalized sentiment data, and can predict the logreturn values. To a somewhat lesser extent, the same is true for Amazon and Tesla. However, GOOG and Microsoft models do not perfom well with normalized sentiment data either. These two stocks have previously shown behaviour opposite to the other listings, for example the correlation heatmaps of features in Section 6.2.1, where between sentiment ratios and logreturn they gave a correlation value around 0. In contrast, the other companies gave values between 0.2 and 0.3. Using the general model gives a weaker prediction, but it can be useful when one has little information about a company and still would like to make a prediction. For instance, rather than relying on the GOOG and Microsoft seperately models, which exhibit poor prediction performance, the more effective approach is to utilize the general model.

[22] The values in the table are rounded to four decimals

| **Models** | **Features with sentiment ratio** | | | | **Features with normalized sentiment ratio** | | | |
|---|---|---|---|---|---|---|---|---|
| | **MAE** | **MSE** | **RMSE** | $R^2$ | **MAE** | **MSE** | **RMSE** | $R^2$ |
| Tesla | 0.0162 | 0.0007 | 0.0263 | -0.0555 | **0.0140** | **0.0005** | **0.0218** | **0.2796** |
| Apple | 0.0078 | 0.0002 | 0.0125 | 0.0049 | **0.0069** | **0.0001** | **0.0107** | **0.2699** |
| GOOG | 0.0072 | 0.0002 | 0.0124 | -0.0216 | **0.0072** | **0.0001** | **0.0121** | **0.0387** |
| GOOGL | 0.007 | 0.0001 | 0.0122 | -0.0237 | **0.0066** | 0.0001 | **0.0108** | **0.1977** |
| Amazon | 0.0075 | 0.0001 | 0.0117 | -0.0434 | **0.0068** | 0.0001 | **0.0103** | **0.1859** |
| Microsoft | 0.0065 | 0.0001 | 0.001 | 0.0027 | **0.0063** | 0.0001 | **0.0095** | **0.1035** |
| General model | 0.0084 | 0.0002 | 0.015 | 0.0001 | **0.0078** | 0.0002 | **0.0134** | **0.2048** |

Table 6.11: Test results of the Random Forest models [23]

Table 6.11 contains the results of the final runs, during which our model is evaluated on the unseen test set. For the features with normalized sentiment ratio, $R^2$ score performs almost the same on the test set as it does on the validation set, which suggests that the models' performance is consistent and they are not significantly overfitting or underfitting the data. On Table 6.11, it can be shown that the models which use the features with sentiment ratio do not perform as well as on Table 6.10. The features with normalized sentiment ratio are significantly better features than sentiment ratio in terms of prediction.

[23] The values in the table are rounded to four decimals

# Conclusions

In this thesis, we sought to find out whether a connection can be established between Twitter messages and the stock market. We have investigated and implemented several different tools to explore the possible links.

Our ideas focused on predictability, starting with correlation analysis. After correct processing and construction of the data, ensuring temporal separation, Spearman's rank correlation was used to determine the connection between logreturn and sentiment ratio. We examined the connection from two directions: sentiments forecasting logreturn (correlation results of *Weekdays_sent_to_stock_dataset*) and vice versa (correlation results of *Weekdays_stock_to_sent_dataset*). Investigating *Weekdays_sent_to_stock_dataset*, we found weak correlations between logreturn and tweet volume. Therefore, the tweet volume feature was used for the prediction models. Furthermore, we identified a correlation between logreturn and sentiments ratio for some companies, exploring *Weekdays_stock_to_sent_dataset*. From a correlation point of view, unlike the other companies, GOOG and Microsoft generally showed weaker and less significant correlations, considering the results of both datasets.

To better capture the predictive links between logreturn and Twitter sentiment, we used a Granger causality statistical test for deeper analysis. In general, the normalised sentiment variables performed better than the sentiment ratio differences. In most cases, both normalized positive and negative sentiment ratio values predict the logreturn value from a 2 weeks normalization interval to 6 months. It is also important to distinguish between companies, as they behave differently from each other. The logreturn of Apple and Amazon can be forecasted from both of the normalized sentiment ratios. On the other hand, at the normalization intervals where the normalized negative ratio of GOOGL, Tesla and Microsoft Granger cause logreturn, the normalized positive ratio does not, and the reverse is true as well. Moreover, the results of GOOG and GOOGL are usually in stark contrast, even though they are different listings of the same company.

Conversely, the logreturn typically does not Granger cause sentiment ratios with long term normalization. However, in some cases logreturn is able to predict negative sentiment normalized in the 2 week time interval. Additionally, the logreturn exhibits Granger causality towards the sentiment ratio differences, which are short-term measures. In the Spearman's rank correlation analysis, we similarly compared short-term (daily) events and obtained weak correlation values within *Weekdays_stock_to_sent_dataset*, which supports the Granger causality finding.

Since we obtained results via correlation and causality that suggested the predictability of logreturn from emotions, we created a prediction model. Since we had previously examined companies separately and observed differences in behaviour, we created company-by-company Random Forest models from the features we had used previously. However, we were curious about how well a general model could predict, so we experimented with that as well.

When attempting to predict logreturns based on sentiment, learning with sentiment ratio difference features did not work at all, as we could expect from the Granger causality test results. When using the features with normalized sentiment ratios, we obtained model, which are relatively weak in terms of predictive power but do explain a non-negligible fraction of the variance. It seems that the normalized features are useful, but the Random Forest did not learn as well as expected.

Comparing the separate company model and the general model, we can conclude that both have their advantages and disadvantages. Using a company-by-company model, we can better predict the logreturn of a given company, but the company must have the necessary amount of data and must be well-known enough to tweet about, otherwise we cannot associate aggregated sentiments with them. The general model may not accurately predict all companies, but it might be used to predict logreturns of small impact companies as well, given sufficient data to run the model.

In the thesis, we focused on exploring the relationship rather than the actual prediction, so as future work, the prediction methodology could be explored in more detail. Such investigation could include comparing other Machine Learning and Neural Network models using our normalized sentiment features.

# Bibliography

[1] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[2] David A Dickey and Wayne A Fuller. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica: journal of the Econometric Society*, pages 1057–1072, 1981.

[3] Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, 54(1-3):159–178, 1992.

[4] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[5] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[6] Shunrong Shen, Haomiao Jiang, and Tongda Zhang. Stock market forecasting using machine learning algorithms. *Department of Electrical Engineering, Stanford University, Stanford, CA*, pages 1–5, 2012.

[7] Luckyson Khaidem, Snehanshu Saha, and Sudeepa Roy Dey. Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*, 2016.

[8] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, and Arun Kumar. Stock closing price prediction using machine learning techniques. *Procedia computer science*, 167:599–606, 2020.

[9] Luca Di Persio and Oleksandr Honchar. Artificial neural networks architectures for stock price prediction: Comparisons and applications. *International journal of circuits, systems and signal processing*, 10(2016):403–413, 2016.

[10] Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia, and David C Anastasiu. Stock price prediction using news sentiment analysis. In *2019 IEEE fifth international conference on big data computing service and applications (BigDataService)*, pages 205–208. IEEE, 2019.

[11] Huina Mao, Scott Counts, and Johan Bollen. Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv:1112.1051*, 2011.

[12] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[13] Eric Gilbert and Karrie Karahalios. Widespread worry and the stock market. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, pages 58–65, 2010.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[15] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[16] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison Wesley, us ed edition, May 2005.

[17] Bolla Marianna and Krámli András. *Statisztikai következtetések elmélete*. Typotex Kiadó, 2008.