LUND UNIVERSITY
School of Economics and Management

Master's Programme in Economics

Predicting Tesla Stock Return Using Twitter Data

An Intraday View on the Relation between Twitter Dimensions and the Tesla Stock Return

by

Gustav Edman & Martin Weishaupt

Abstract In this thesis, Twitter data is used to predict the intraday stock return for Tesla, Inc. We present two different methods to extract the tweets' sentiment: A dictionary-based approach (VADER) and a machine learning approach (SVM). Additionally, we control for other dimensions as the user and discussion dimension. Then a Granger causality test and a lasso regression are conducted on a one- and five-minute interval. The results suggest that there is no predictive power in the information of the tweets for the dictionary data set and the machine learning data set. Using a subset of the dictionary data set with only the cashtag does not alter the results. The reason for this may be that we employ two linear models on a possible non-linear problem.

NEKP01
Master's Thesis (15 credits ECTS)
May 2020
Supervisor: Anders Vilhelmsson

Contents

List of Figures

List of Tables

1 Introduction

Since the development of social media platforms in the last two decades, new sources of information and platforms for communication have opened up, in which billions of people can stay connected, share opinions, and express feelings regarding individuals, companies, and specific events (Smailović et al., 2014). Facebook, Twitter, and other new media platforms have been booming, where Twitter grew by 8 per cent between 2015 and 2019 (Oritz-Ospina, 2019). The growing use of the aforementioned social media platforms is affecting the stock price movements through the use of tweets, posts, as information is now readily accessible for the public and may affect investors decision-making. One good example of this is the impact of tweets made by Elon Musk regarding his company, Tesla Inc. On May 1, 2020, Elon Musk tweeted "*Tesla stock price is too high imo*" stating he thinks Tesla's stock is overvalued. The tweet resulted in a sharp fall in the stock price, a 4.6 per cent fall within the first ten minutes after the tweet was published (Korosec, 2020). Another tweet from Elon Musk on August 7, 2018, where he tweeted about taking Tesla, Inc. private and stated he had secured fundings to buy the company, resulted in an instantaneous price jump and Tesla's shares increased by 10 per cent (Ferris, 2018). Although these are two extreme examples, it demonstrates the impact of social media platforms and its users can have on the stock market.

The growing influence of social media platforms has consequently brought a vast amount of attention from researchers within the field of behavioural finance. Besides the interest to investigate the role of new media platforms has on the stock market and the possibility to predict stock price movements by the sentiment of the shared online posts, the social media development has brought the legitimacy of the Efficient Market Hypothesis (EMH) and its thesis on unpredictable stock prices into questioning (Smailović et al., 2014; Lo, 2004; Benthaus and Beck, 2015; Bollen et al., 2011). The majority of previous studies conducted on the predictability for the sentiment on stock price movements have mainly focused on stock market indices, whereas there is a limited number of studies examining the effect on individual stocks. Further, almost exclusively, all studies have applied their sentiment analysis on daily data. Bollen et al. (2011) were one of the first authors to use Twitter-based data to predict future stock price movements. They use a dictionary-based approach to obtain sentiment based on the expressed opinions, emotions and moods in tweets on daily data. Their results suggest that Twitter sentiment have an 88 per cent predictive power on the movements on the stock market index, DJIA. Additionally, one of the earliest works examining the relationship between the microblogging, Yahoo!, platform and the stock market was Das and Chen (2007), employing daily data on both an index level and individual level. Results show that there is weak evidence of the predictive power of sentiment on individual stocks, while the authors find a significant effect on the index level. Smailović et al. (2014) and Sul et al. (2017) explore this relationship further by investigating the impact Twitter sentiment has on individual stock returns. Smailović et al. (2014) apply two different sentiment classifications (positive vs negative, and positive vs negative vs neutral) on eight individual stocks over nine months in 2011. By running Granger causality tests, the authors find a correlation in the three-classification sentiment, but no correlation in the positive vs negative classified sentiment. Sul et al. (2017) takes, besides, the Twitter sentiment to predict future stock returns, the user dimensions into consideration. They look at individual stocks listed on S&P 500 and perform sentiment analysis using a dictionary-based approach over two years. The results from Sul et al. (2017) suggest that a high number of followers and retweets of a Twitter user have a significant and immediate effect on the stock price.

This thesis will also make use of the dictionary-based approach, as well as a machine learning approach when conducting the textual sentiment analysis. The reason for this is to examine whether there exists a causal relationship between Twitter sentiment related to Tesla, Inc. and the Tesla stock return. To accomplish this, we apply two different methods, a Granger causality test and a lasso regression. We divided the data up into three different sets. The first data set includes our full sample where we extract the sentiment with a dictionary-based tool called VADER. The second data set is a subset of the first data set, only including tweets with the cashtag $\$TSLA$.[1] Our third data set is based on the machine learning approach, where we manually classify a training set on sentiment and then employ a Support Vector Machine (SVM) to predict the sentiment for all other tweets. We then divide these up to check if the trading interval has an impact. Such that we have data on a one- and five-minute interval.

This paper aims to contribute to the field of predicting stock price movements by using microblogging platforms. We are contributing to the research by using intraday stock returns, as to our knowledge, the majority uses daily data. Such that our main objective is the following: *Is Twitter data related to Tesla Inc. a predictor for the Tesla stock return?* Therefore, we broaden the understanding by controlling for the dimensions of a tweet, for example, incorporating sentiment, username, number of retweets, tags and more into the prediction. As a result, we investigate the different effects of the dimensions.

In the first step of analysing the data, we conduct a Granger causality test, where we do not find a clear pattern in the results, indicating that there is most likely no real causal relationship between Twitter information – sentiment, replies, retweets, favourites – and the Tesla stock return, even though we do find significant results. When conducting the Granger causality test on the subset with the cashtag, we do find that the coefficients for the first ten lags are significant, which could hint that using a smaller trimmed sample for our financial problem is superior to the bigger sample. In the second step of the process, we applied a lasso regression on the same data sets for the first ten lags. For the specifications with the manually classified tweets and the dictionary-based model, we do not find any predictability, indicating

[1] A cashtag is similar to a hashtag, with writing an $ in front of the companies ticker symbol. The user wants to point out that the tweets include investment information.

that the expected return is the most appropriate model to use. When applying the lasso on the cashtag sample, we find that the number of tweets with two lags are possible predictors for the stock return on the one-minute interval and retweets with two lags can be used to predict the return on the five-minutes interval. Nevertheless, we cannot find any result that is strong enough to answer the research questions, and we conclude that, for our setting, Twitter cannot be used to predict the Tesla stock return on an intraday level.

The structure of this thesis is as follows; Section 2 introduces and discusses the theory and previous literature of sentiment analysis and stock market predictions. The hypotheses are presented at the end of Section 2. Section 3 introduces a data description of the collected data from Twitter and the Tesla stock price. Also, we present the descriptive statistics for the used data. In Section 4, two time-series methods are described, and we present a thorough description of the sentiment analysis. Further, Section 4 describes the weighting process and the machine learning in two parts; Support Vector Machine and lasso regression. Section 5 presents the main results from the different tests on the data sets. Lastly, Section 6 concludes the results and discuss further research improvements.

2 Theory and Previous Literature

2.1 Theoretical Framework

Following the renowned efficient market hypothesis (EMH) from Fama (1965), the stock price today incorporates all relevant information available, and thus the price is the best forecast of the future development of the stock (Shiller, 2003). Then the assumption of complete information and rationality indicates there is no room for predictability. Changes in the price are merely random, and hence the stock follows a random walk, implying that any exchange of information on social media platforms is irrelevant since all investors already have all the relevant information, as they react instantaneously, and thereby, the stock price incorporates the information. However, as investors do not have access to all relevant information and they are not acting rational, the assumptions cannot be fulfilled (Lo, 2004). As Lo (2004) states, the investors are not rational under uncertainty and instead have different biases, for example, overconfidence, overreaction, loss aversion, herding, calibration of probabilities and more. Hence, the market is not efficient and gives opportunities for market participants to realise abnormal returns. Gathering information about a stock can, therefore, improve the understanding of its future development, giving an investment advantage (Benthaus and Beck, 2015). The additional information can be gathered from the traditional media, like TV stations and newspapers, but also from the new media platforms (e.g. Twitter, Facebook, StockTwits). The new media platforms have the advantage that the majority of the traditional media provide their information on these platforms. Additionally, regular users can also exchange information or share their opinions. Another advantage is that the shared information is directly available worldwide for everybody in the social network, such that the information can be used for example to reevaluate the price of a stock market (Benthaus and Beck, 2015). There are different ways of analysing the information on social media, and we focus on a sentiment analysis incorporated into a dimensional model.

2.1.1 Sentiment Analysis and Stock market

Lawrence et al. (2007) created an asset pricing model where they include investor sentiment to capture the individual beliefs of future performance of a stock. Taking into account individual preferences such as risk awareness, and underlying factors (wealth, educational background, age, gender, and culture) results in different levels of sentiment between individuals. Additionally, an individual investor's sentiment level can be changed, for example, by new information (Lawrence et al., 2007), such as news regarding a listed company as Tesla, Inc. Therefore, Lawrence et al. (2007)

state that an investor with a high sentiment expects a high growth rate for a stock, which results in that the expected growth and discount rate given the investor's sentiment determines the stock price. Such that the stock price depends on the future dividend (DIV_1), the expected discount rate (r^s) and the expected growth rate (g^s):

$$P_0 = \frac{DIV_1}{r^s - g^s} \tag{2.1}$$

For an investor with a high sentiment the expected growth rate will be high, whereas the relationship between sentiment and expected discount rate is the contrary (high sentiment implies low expected discount rate), which results in a high stock price. The opposite is true for an investor with a low sentiment; thus, the growth rate is expected to be low and the discount rate high, yielding a lower perceived stock price compared to the market price. Therefore, investors with low sentiment will sell the stock when they perceive that the market price is higher than what they consider it should be. The investors' with a high sentiment, on the other hand, value the stock price to be higher than the market price and are willing to buy the stock (Lawrence et al., 2007).

2.1.2 Social media dimensions and its relation to the stock market

As seen in section 2.1 the EMH does not hold, and thus there exist opportunities for abnormal returns. Investors can use social media data to exploit these abnormal return possibilities, as explained in a model from Benthaus and Beck (2015). They argue that three different dimensions of social media data can influence movements on the stock market. The model is visualised in figure 2.1.

The first dimension is called the *user dimension*, as it depends on the publishers' expertise and popularity. Benthaus and Beck (2015) argue that the quality of a post from an expert contains beneficial information to other users. Since the user mostly shares information that is new to the network. As a result, the post is shared more often on the social platform. Additionally, experts share more often their opinion on the social network than others, and if the information is related to a stock, this should increase the trading activity. Also, popular users have a higher impact on the stock market, than normal users, as the popularity may indicate a higher competence. Which results in a higher impact on the trading activity of a stock. An indicator of a user's popularity may be the number of connections a user has on the social platform (Benthaus and Beck, 2015).

The second dimension is the *message dimension* and contains two different parts. The first is the information richness of a post, which is increasing by mentioning other users, for example, experts. Doing so, the user intends to increase the validity of the post according to Benthaus and Beck (2015). Another way to increase the richness is to include certain words in a post, which can be marked with a hashtag or cashtag to make it easier to follow a discussion or find valuable information. Additionally, linking to some external website with follow-up information in the post is increasing the information richness. All these three points will increase the

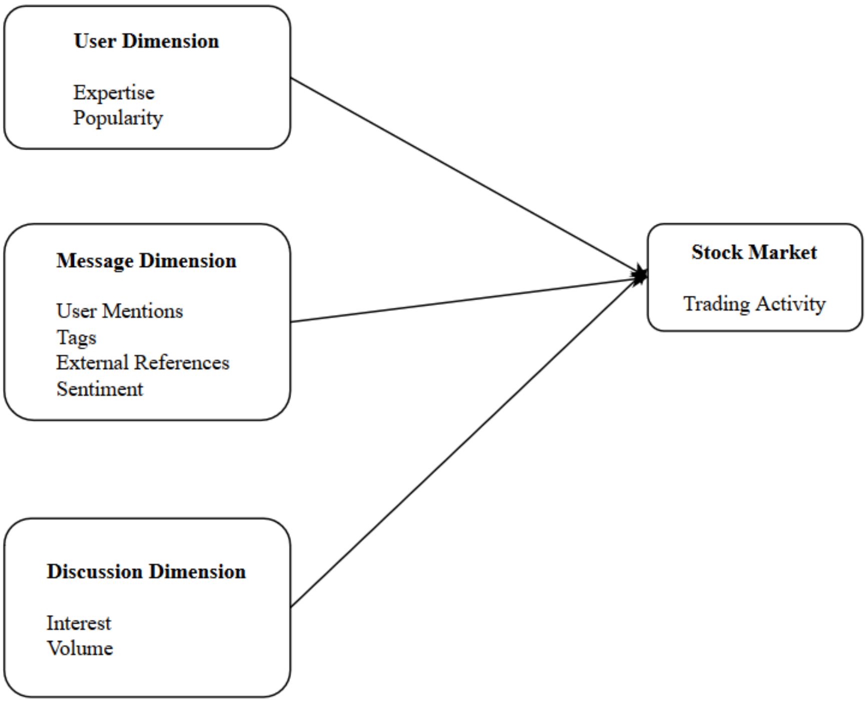

Figure 2.1: Impact of social media on the stock market according to Benthaus and Beck (2015)

information richness of a post on a social media platform. The second part is the sentiment analysis of the message, which we described in more detail in section 2.1.1, but the main conclusion for the sentiment analysis from Benthaus and Beck (2015) is that a message with positive sentiment should increase the trading activity of the stock, which are then possible to forecast the stock price.

The third and last dimension is the *discussion dimension*, which we divide up in two points. The first is that microblogging websites allow users to retweet posts so they can show their interest which can have a positive effect on the trading activity. If a post gets a significant number of retweets, a herd behaviour may exist, which can have a positive effect on the trading activity on the stock market (Lo, 2004). Secondly, Benthaus and Beck (2015) argue with the help of Rui et al. (2013) that the number of posts can have an impact on the economy, as Rui et al. (2013) find that the volume of tweets has a positive effect on movie sales. With this result and other (Sprenger et al., 2014; Tirunillai and Tellis, 2012), Benthaus and Beck (2015) conclude that the volume of tweets is positively correlated with the trading activity. Thus, the volume can be used to predict the stock return, along with the retweets.

2.2 Previous Literature

Throughout the last two decades, researchers within behavioural finance have been trying to answer the question of how the stock markets are affected by the public's mood and sentiment. A recently common approach is to use the textual sentiment analysis, where a posted text from a user on social media is extracted into sentiment information to measure people's opinions (Li et al., 2019) and the posts together represent the public's mood to predict the stock market (Bollen et al., 2011). The sentiment can be measured as positive-negative or other classification effects such as strong-weak (Kearney and Liu, 2014). The increased focus on social media when employing a sentiment analysis on the stock market is attributed to the user-generated content, which is considered to be more credible and trustworthy than the traditional media content (Yu et al., 2013). Even though modern media sources are regarded as more volatile, since it is on an individual level including individual traders compared to traditional media, the advantage of using individual messages is that they account for people's instant reactions to new information available and can have an impact on the stock market (Das and Chen, 2007). Therefore, collecting social media sentiments can be seen as obtaining small investors' behaviour which may affect the stock market (Kearney and Liu, 2014). Das and Chen (2007) conduct one of the earliest works investigating the relationship between social media and stock predictions. They use message board postings from Yahoo!, a microblogging platform, to analyse investor sentiment effects on stock returns. Further, Das and Chen (2007) use the index MSH of the chosen stocks and aggregated the sentiments across all stocks into an index. Employing daily data, the authors find that there exists an effect between sentiment index and the MSH index while looking at an individual level the sentiment effect on an individual stock seems to be weak. Other researchers did not use sentiment analysis to analyse microblogging data. For example, Mao et al. (2012) found that the number of tweets is correlated with the closing price of S&P 500. Similarly, Asur and Huberman (2010) find that the rate of movie tweets are better predictors for movie revenue than other forecasting models. Furthermore, the authors combine the prediction model with a sentiment analysis of the tweets, which improves the prediction model even further.

More recent studies have used Twitter data when predicting public sentiment effects and correlation with stock returns. Bollen et al. (2011) use a dictionary-based sentiment analysis on approximately 9.9 million tweets to predict the future stock market (DJIA) daily closing prices between February and December in 2008. The sentiment analysis is conducted by using two different types of mood tracking tools and employing two different methods (Granger Causality Analysis and a Neural Network approach) on a time-series regression. The first tool measures the mood of the tweets as "positive" or "negative". The second one includes more moods (Calm, Alert, Sure, Vital, Kind, and Happy) to predict the public mood and sentiment on DJIA returns. The authors conclude that there is some Granger causality between moods a couple of days ago and today's value of DJIA. As a result, there is a predictive power of public sentiment on DJIA values in terms of changes of public moods and especially in the mood term "Calm". However, they do not find effects between the mood terms "positive" or "negative" and changes in the DJIA. Moreover, the authors find that with the Neural Network method, their sentiment on different moods has an 88 per cent direction accuracy at predicting future daily up and down movements on the closing price of the DJIA.

Smailović et al. (2014) use the dimensional approach to examine if Twitter sentiment can predict stock price changes. The sample data contains the closing prices of eight individual stocks and 153,000 financial tweets connected to the stocks between March 2011 and December 2011. They use two different sentiment classifications, a two-class setting (positive, negative) and a three-class setting (positive, neutral, negative). A Granger causality test shows that for the first setting, there is no significant evidence between sentiment and stock price movements. For the three-class setting, the results indicate that there is Granger causality correlation between the positive sentiment probability and the stocks closing prices. Further, results show that daily changes in the positive sentiment have a significant effect on future trading days individual stock prices. Additionally, tests showed that there is no evidence of reverse causality, i.e. stock price movements affect the sentiment (Smailović et al., 2014).

Another study also using the dimensional approach and Twitter sentiment to predict future stock returns is Sul et al. (2017), who collect over 2.5 million tweets over two years. They focus on individual firms from the S&P 500 stock index. For the sentiment analysis, the authors use a dictionary-based approach employing the Harvard-IV dictionary. Beyond examining the predictability of Twitter sentiment on stock returns, Sul et al. (2017) further investigates the difference in the effect on stock returns depending on the number of followers and retweets from a user. Sul et al. (2017) find that tweets from users with many followers have an immediate effect on the stock prices because the tweet is spread more quickly to many users, similar to news media, and this is true for both positive and negative sentiments. Further, tweets from users with many followers seem to have little to no effect on future stock returns. Contrary, tweets from users with few followers (less than the median of 171 followers) and with retweets has no significant instant effect on stock prices but has a more substantial impact on future trading days. The reason for this, the authors argues, is because the tweets from users with few followers take a longer time to spread, and so does the impact on the stock returns. Moreover, tweets from users with few followers have the most significant effect on future stock returns, as they do not have many retweets (Sul et al., 2017).

2.3 Summary and Research Question

From the theory and previous literature section, we formalise our research questions and hypotheses. First, our approach differs from previous research as we focus on one specific stock (Tesla, Inc.) and not on a stock index. Second, we use intraday stock prices and Tweets related to Tesla, Inc. Third, we do not solely focus on a sentiment analysis of the tweets. Fourth, we also try to analyse other factors of the tweets as described in section 2.1.2. Leading us to our research questions:
Can Twitter data about Tesla, Inc. predict the intraday stock price of Tesla, Inc.? If yes, what dimensions are the most important for predicting the Tesla, Inc. stock return?

From our research questions, we formulated the following four hypotheses:

Hypothesis 1. *A high sentiment in the tweets has a positive effect on the future stock price of Tesla, Inc. Whereas a low sentiment in the tweets has a negative effect on the future stock price of Tesla, Inc.*

This has been studied in previous literature before, however not for intraday stock prices. By exploring this new avenue such that we try to broaden the knowledge of the effect, twitter data has on the stock market. However, we anticipate that the results are similar to previous literature (Sul et al., 2017; Bollen et al., 2011) since the change in sentiment of investors should respond similarly to new information no matter the time interval.

Hypothesis 2. *A high volume of tweets related to Tesla, Inc. increases the predictability of the Tesla, Inc. stock return.*

Hypothesis 3. *An expert or popular user for Tesla, Inc. will increase the predictability of the Tesla, Inc. stock return.*

Hypothesis 4. *A tweet with high information richness, for example, using a cashtag, increases the predictability of the Tesla, Inc. stock return.*

Hypotheses 2 - 4 are related to the theoretical framework form Benthaus and Beck (2015) and the results from Sul et al. (2017). We find previous literature for hypothesis 2 (Mao et al., 2012) and 3 (Sul et al., 2017) but not regarding the stock price of Tesla, Inc. For hypothesis 4 there exists no previous literature, at least to our knowledge. Therefore we will try to broaden the understanding of the information richness and its effect on the Tesla, Inc. stock price.

3 Data

In this section, we present the data cleaning part and the descriptive statistics of the used data. First, we present the procedure of the data cleaning for the Twitter data and second, for the Tesla Stock data, and third, we display and discuss the descriptive statistics for both data sets.

As we retrieve the Tesla and Twitter data sets from different sources, we need to take care of time zone differences, as the Twitter data is in GMT and the Tesla data is in CET/CEST time. As a result, we adjust for differences in the daylight saving time changes for both data sets. Even though the two data sets are adjusted to be in the same time zone, they are not following the same calendar. The stock data follows the Swedish calendar, while the Twitter data follows the US calendar. Therefore, the daylight savings time (DST) are different between our two data sets. We took this into account and adjusted for the differences where the Swedish DST ended on October 27, 2019, and the US DST ended on November 3, 2019.

3.1 Twitter Data

The Twitter data set used in this paper contains publicly available tweets from October 1, 2019, to December 30, 2019, related to Tesla, Inc. retrieved from Twitter.com. To collect the tweets the search query consists of the following words: "Tesla", "Elon Musk", "@Tesla", "@elonmusk", "$TSLA", "#Tesla". We choose to include more than only the Tesla cashtag, contrary to previous research (Smailović et al., 2014; Sul et al., 2017). The reasoning for this is that we want to capture a bigger picture of the sentiment to get a more general attitude towards the company and not only tweets directly related to the Tesla stock. Furthermore, only one-fifth of the collected tweets have a cashtag, which means that we could rely on single tweets for a time interval. Thus, we may not capture the “actual” sentiment and instead only capture an extreme value of the sentiment. Nevertheless, we are checking our results with only the cashtag too, as there may be much noise in the full data set and not a lot of relevant information for the Tesla stock. A reason why there is noise is that Tesla is a much-discussed company on Twitter comparing to other companies. Additionally, we are interested if experts have a bigger impact on the stock than other users, for example, shown in the introduction for Elon Musk.

To calculate the sentiment of each tweet, we use two different approaches. The first method uses the Python package VADER, where every tweet gets assigned a number between minus one and one (more in section 4.2). The second approach applies machine learning techniques. Where we manually classify 1,605 tweets by ourselves into three categories (negative, neutral and positive). Then we fit a Support Vector Machine model to predict all other tweets. With both approaches, we classified nearly 610,000 tweets, and for each tweet the following variables are available: date[1], username, replies, retweets, favourites, text, mentions, hashtag, VADER sentiment, and manually classified tweet sentiment.

As the stock market is not traded every day and hour, we removed all non-trading days and hours, except for hours from 07:00 to 21:59 for the one-minute interval and from 07:00 to 21:55 for the five-minute interval to still be able to capture a lag in the sentiment.[2] Furthermore, to be able to match the Twitter data with the stock data, as the used stock data is on a one- and five-minute interval, we aggregate the Twitter data on a one-minute and a five-minute interval. Where the sentiment is the average for all tweets in the interval. Thus, the data set contains 54.000 observations for the one-minute interval and 11.505 observations for the five-minute interval. An example output of the Twitter data is presented in table 3.1. To show our data cleaning for the Twitter data, see figure 3.1.

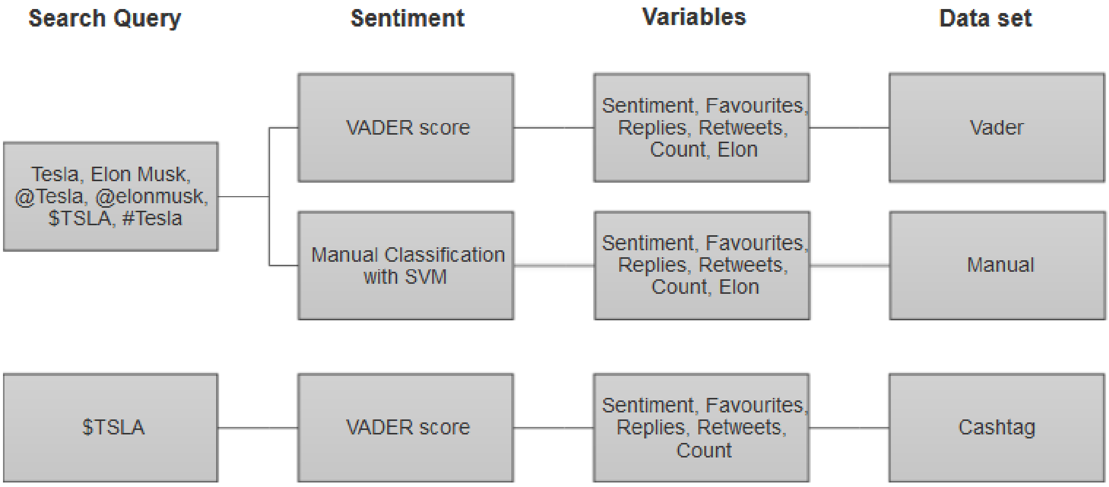
Search Query
Sentiment
Variables
Data set
Tesla, Elon Musk, @Tesla, @elonmusk, $TSLA, #Tesla
VADER score
Sentiment, Favourites, Replies, Retweets, Count, Elon
Vader
Manual Classification with SVM
Sentiment, Favourites, Replies, Retweets, Count, Elon
Manual
$TSLA
VADER score
Sentiment, Favourites, Replies, Retweets, Count
Cashtag

Figure 3.1: Data cleaning procedure of the Twitter data; the procedure is done for the one- and five-minute interval.

3.2 Tesla Stock Data

The Tesla stock data is obtained from Bloomberg, by a Bloomberg Terminal, and contains the opening and closing price on a one- and five-minute interval. For the one-minute interval, we have data from October 7, 2019, to December 31, 2019, and for the five-minute interval, we have the stock prices from October 1, 2019, to December 31, 2019. For predicting the stock price movements, we subtract the opening price from the closing price to calculate the return. As the stock data set only includes the trading hours, but the sentiment data set starts at 07:00 for every trading day, we added empty observations to the data set to match the number of observations of the Twitter data.

[1] Timestamp: “YYYY-MM-DD hh:mm:ss”

[2] For some special days, like December 24, the trading day is shorter which we account for.

Table 3.1: Example output for the Twitter data set

	date	username	text	VADER	retweets	replies	favorites
1	2019-10-25 16:57:26	infactjack	That's me!	0	0	0	0
2	2019-11-06 01:54:27	mathwam	@tesla_raj an answer to the old age question!	0	0	1	1
3	2019-10-18 20:35:15	ElenaChudomirov	Long lost twin @elonmusk #[r.l.]	-0.3182	0	0	0
4	2019-11-24 11:10:24	DesignWorlds	Elon Musk boasts of nearly 150,000 Tesla Cybertruck orders despite launch gaffe [r.l.]	0	0	0	0
5	2019-12-05 08:08:29	westernshores	Elon Musk trial: Vernon Unsworth says entrepreneur's tweets 'humiliated' him [r.l.]	0	0	0	0
6	2019-11-29 05:20:36	yikes_itsaut	Is this Elon Musk	0	0	0	7
7	2019-10-29 17:53:54	httpJunkie	Excellent article by #Bloomberg @business about the #TeslaModel3Survey! 5,000 Tesla Model 3 Owners Tell Us What Elon Musk Got Right and Wrong. As a former Tesla employee, I find these articles that are largely unbiased very refreshing and interesting! [r.l.]	0.6184	0	0	0
8	2019-11-15 01:36:17	formgram	Can you please remove Advertisement from the Tesla touchscreen chromium browser? Since you hate advertisement?	-0.4137	0	0	0
9	2019-11-24 02:31:50	John_Gardi	Why the #Tesla #Cybertruck Looks So Weird. The pickup truck @ElonMusk unveiled Thursday night features sharp angles, but no side mirrors or "crumple zone" to absorb the force of a collision: [r.l.]	-0.7612	2	5	7
10	2019-11-29 19:53:43	PlainSite	Fast forward eight years. Who is in the $TSLA hot seat now? Robyn M. Denholm, one of the same defendants named in all of those Juniper suits. For permitting Tesla and SolarCity to do the exact same thing. [r.l.]	0	13	1	39

3.3 Descriptive Statistics

In this section, we present the descriptive statistics for the *VADER* sentiment, the manual classified sentiment, the return and the weighted retweets, replies and favourites. The descriptive statistics can be seen in table 3.2. There are differences in the number of observations between the two sentiments and the return which is expected. As observations have been added before the trading days the return has a lot of missing values. Whereas for the sentiment, missing values only exist when for example Twitter is not available, like on October 2, 2019 when Twitter was down for a ten minute period. It could also be that there have been no tweets during a specific time interval. When the missing observations are included we have the same amount of observations as explained in the previous section 3.1. When comparing the mean for the two sentiments, we can see they are nearly the same between the intervals. Whereas for the return, we can see that the difference between the intervals for the return is seven times higher for the five-minute interval, which is close to the expected five times higher mean for the five-minute interval. When looking at the weighted variables, we can see that the mean and median change a bit, but not as much as the maximum and minimum, which results in a higher standard deviation. This is especially true for the one-minute interval where we have high values for minimum and maximum. If the prediction with the weighted variables will increase the accuracy, then probably more extreme values are better suited to predict the Tesla stock return.

To see if we can find a relationship between the explanatory variables and the return, we plotted the correlation for the full data set on the one-minute interval, which is presented in figure 3.2. We can see that there is nearly zero correlation in the variables, as all correlation is between -0.1 and 0.1. This should be kept in mind, as any regression will have problems, finding good results, when there is little to zero correlation between the independent variables and the dependent variable.

Table 3.2: Descriptive statistics for VADER sentiment, manual classification sentiment, return, weighted retweets, weighted replies and weighted favourites on a one- and five-minute interval.

Interval	**Variable**	**N**	**Mean**	**Median**	**Std**	**Min**	**Max**
1 Minute	*VADER*	48 414	0.141	0.136	0.270	-0.970	0.993
	Manual	48 414	0.039	0.000	0.264	-1.000	1.000
	Return	23 050	0.003	0.001	0.324	-3.700	4.017
	Retweets	48 414	0.038	0.000	3.995	-720	386
	Replies	48 414	0.004	0.000	1.185	-195	74
	Favourites	48 414	0.250	0.000	18.742	-2517	2 012
5 Minute	*VADER*	11 505	0.143	0.145	0.114	-0.935	0.855
	Manual	11 505	0.039	0.000	0.133	-1.000	1.000
	Return	4 921	0.021	0.025	0.722	-7.890	5.913
	Retweets	11 505	0.034	0.000	1.009	-65.500	27.800
	Replies	11 505	0.004	0.000	0.344	-21.638	10.250
	Favourites	11 505	0.216	0.000	6.06	-229.409	209.205

This is odd as we would expect a higher correlation between the variables and the return, as explained in the theory and previous literature part. Additionally, we plotted the correlation for all other data sets too with the same outcome.

The most frequently used words in the tweets are presented in the word cloud, see Appendix A.[3] It can be seen that "elon", "musk", and "tesla" are one of the most used words, which is not surprising as they are all included in the search query.

[3] Stopwords have been removed by the R package stopwords, such that 174 words have been removed

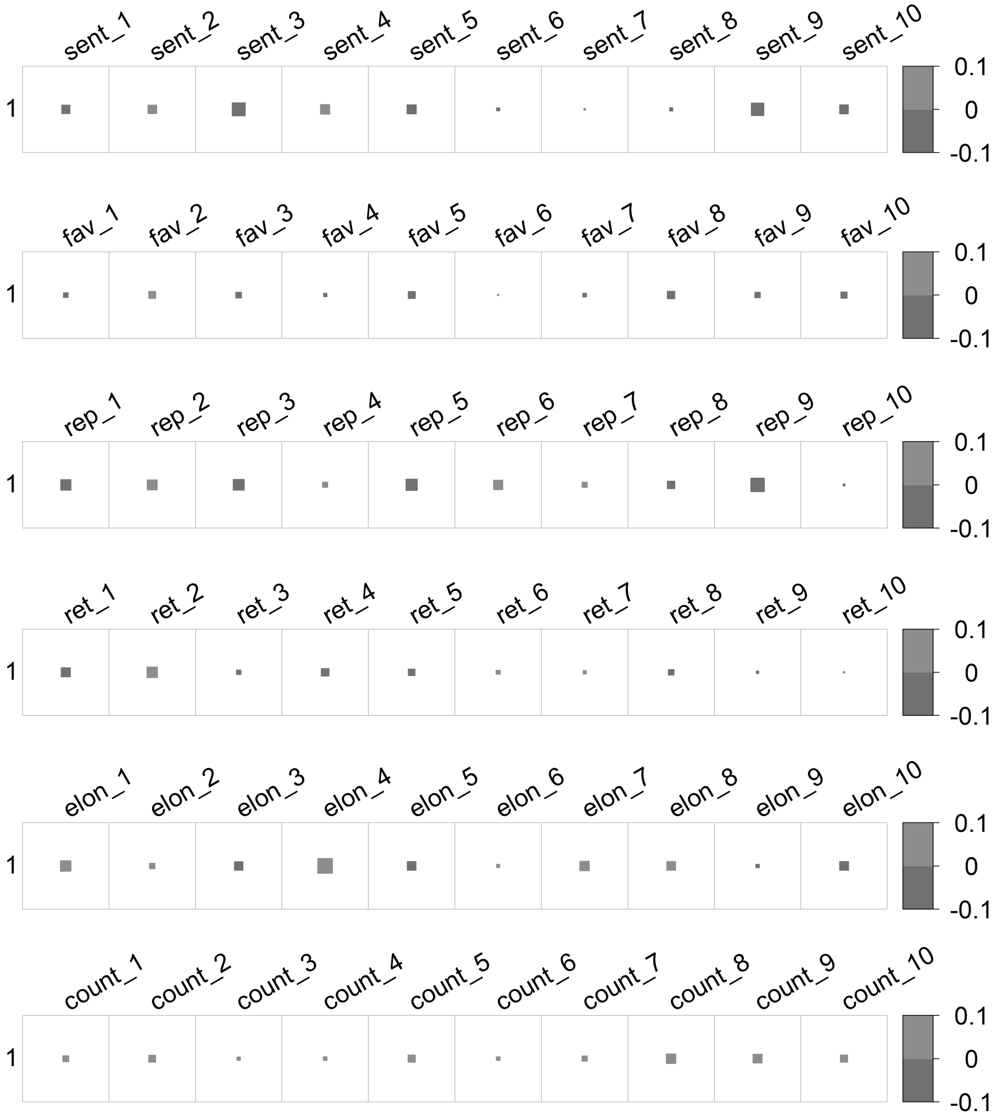

Figure 3.2: Correlation between return and the explanatory variables for full sample (Vader 1) for the one-minute interval.

4 Methods

In this section, we explain different empirical methods to predict the Tesla stock return from tweets. In the first part, we present the time series methods to analyse the relationship between the explanatory variables and the return. In the second part, we introduce the concepts of sentiment. In the third part, we explain a weighting scheme for the variables and lastly, we introduce two machine learning techniques. We use the first method to predict the sentiment from tweets, and with the second, we analyse the effects of the tweets on the return. We perform all computations in R, except for the estimation of the VADER sentiment which we execute in Python.

4.1 Time Series Analysis

In what follows, we present two methods to check if we can use sentiment to predict stock return. A condition for being able to test if sentiment can predict stock return is that the data needs to be stationary. Therefore, we apply an Augmented Dickey-Fuller test which tests for stationarity. The second method examines if there are any causal effects between the Twitter sentiment and the stock return; for this, we conduct by running a linear Granger causality test.

4.1.1 Testing for Stationarity – Augmented Dickey-Fuller Test

To conduct the stationarity test the regressions take the following forms:

$$\Delta y_t = \gamma y_{t-1} + \epsilon_t \tag{4.1}$$

$$\Delta y_t = \alpha_0 + \gamma y_{t-1} + \epsilon_t \tag{4.2}$$

$$\Delta y_t = \alpha_0 + \gamma y_{t-1} + \alpha_2 t + \epsilon_t \tag{4.3}$$

Equation 4.1 is with a random walk, equation 4.2 is with an intercept (α_0) and equation 4.3 additionally contains a drift term ($\alpha_2 t$). On these regressions, an Augmented Dickey–Fuller test can be conducted. The null hypothesis of the test states that the data is non-stationary, such that $\gamma = 0$. If we can reject the null hypothesis, the underlying data is stationary, such that we can perform a Granger causality test (Enders, 2014).

4.1.2 Granger Causality test

To test if past values of a variable (x_{t-m}) can affect the current value of another variable (y_t) a test for Granger causality can be conducted. In other words, can past values of x_t (for example sentiment) be used to predict the value of y_t (return)? If this is true, then it can be said that x_t (sentiment) Granger causes y_t (return). To conduct the test the variables y_t and x_t need to be stationary (Enders, 2014). Then the regression to test if x_t is Granger causing y_t takes the following form:

$$y_t = \sum_{j=1}^{m} \alpha_j y_{t-j} + \sum_{j=1}^{m} \beta_j x_{t-j} + \epsilon_j \tag{4.4}$$

Where ϵ_t is uncorrelated and white noise. m needs to be smaller than the length of the time series. The null hypothesis for the Granger causality is that β_j is equal to zero (Granger, 1969). We decided only to conduct a linear Granger causality test and not a non-linear Granger causality test as we later employ a linear regression model.

4.2 Sentiment Analysis

Sentiment analysis, focusing on social media and internet postings, are well suited for analysing the short-term effects of the sentiment on stock returns, and other variables connected to the financial market, due to the increased number of users and the increased amount of time spent on the different platforms (Kearney and Liu, 2014). The individual messages posted on social media can have an impact on other people's opinions and behaviour, creating an essential alternative source other than traditional news and media when it comes to explaining the changes in the financial markets as stock prices, returns, and volatility (Das and Chen, 2007). Kearney and Liu (2014) states that social media messages are small investor sentiments. However, there is a couple of challenges that one should be aware of when especially extracting sentiment from social media platforms, which are that the sentiment will be noisier since it is on an individual level and at a larger volume. Furthermore, the shortness of the text in each message, length limitation on Twitter for example, and the use of abbreviations, acronyms and slang in the social media content make it even more difficult to practically apply sentiment analysis (Hutto and Gilbert, 2014).

The most common way to extract the collected textual sentiments is to classify words and messages into different moods in the form of negative and positive sentiment (Li et al., 2019), through two different types of methods, either using a dictionary-based approach or a machine learning approach (Kearney and Liu, 2014). The dictionary-based approach predicts sentiment based on a chosen dictionary, where a text is read by a computer program which classifies the message into a group based on the written words (Liu, 2012). The classifier is based on a word-list of pre-defined dictionary categories and employed with a weight on each word. Usually, equal weight is put on each word and therefore giving the same label of importance to all words. Although, there are exceptions in the study by Loughran and McDonald (2011) words are assigned different weights depending on the frequency of appearance in the data set. Normally a textual sentiment analysis employing a dictionary-based approach process looks as follows; firstly, select and collect text from a platform, for example, Twitter. Lastly, select a dictionary tool to compute the sentiment categories and then compute the text files in the computer program to obtain sentiment scores. The sentiment scores are later used as a measure to create a model and test a hypothesis (Kearney and Liu, 2014).

An advantage of employing a dictionary-based approach in textual sentiment analysis instead of a machine-learning approach is the "non-need" of a training set, which is difficult to find in a social media sentiment content and is time-consuming to create manually (Li et al., 2019). However, there are several issues related to the usage of a dictionary-based approach, the two most difficult issues are the choice of a dictionary (word-list) where there is a difference between dictionaries in the categorisation of words, and the other issue is how a researcher assigns the weight between words in the dictionary. The two most widely used dictionary-based methods as benchmarks are the GI and the LIWC, both classifying words into binary classes (positive or negative). The GI, or Harvard GI lexicon, is a popular conventional tool from 1966 containing more than eleven thousand classified words into more than 180 categories. It is developed for content and sentiment analysis. Likewise, the LIWC analysis tool is also well-known in the field of content analysis, and both of them are appropriate in extracting sentiment in the context of social media. Regardless, there are setbacks with both of the models as they do not account for slang, acronyms and emoticons, which is important in the extraction of social media sentiment (Hutto and Gilbert, 2014). Moreover, they are unable to account for the sentiment differences by the intensity in the text, for example, "Tesla is a good company" and "Tesla is a great company". They will only register "good" and "great" as positive sentiments but not distinguish the intensity between them. ANEW, SentiWordNet and SenticNet are three dictionaries that do take intensity into account and, thus, also assign different weight depending on the sentiment intensity of the word. Similar to the above-mentioned conventional models, ANEW is not taking the lexical features in the social media sentiment into account, which applies for the SentiWordNet model as well. The SenticNet are based on human-curated data but also supervised techniques. It classifies its sentiment scores within a continuous range between -1 and 1 (Hutto and Gilbert, 2014).

In order to extract and classify the Twitter sentiment and apply our sentiment analysis, we have chosen to use the dictionary-based tool Valence Aware Dictionary and sEntiment Reasoner (VADER). The reason is that the VADER tool performs very well in sentiment analysis in terms of extracting sentiment from social media, better than the conventional methods using the LIWC and GI dictionaries (Hutto and Gilbert, 2014). Hutto and Gilbert (2014) find that the VADER tool outperforms the manually human classified sentiment. VADER classifies the sentiment in valence, i.e. categorise the sentiment into different moods (positive, negative and neutral) including slang, abbreviations, emoticons and acronyms. The VADER accounts for the sentiment intensity by assigning weight on the use of punctuation's, capitalisation's, degree modifiers, the contrastive conjunction "but" to allow for mixed sentiments in a sentence, and lastly examining the tri-gram preceding a sentiment-laden lexical feature (Hutto and Gilbert, 2014). The VADER method yields a normalised and weighted VADER score between -1 and 1, similar to the SenticNet, where the former one is very negative, and the latter one is very positive. We will use the normalised and weighted VADER score in our paper to estimate the potential impact Twitter sentiment has on the Tesla stock return and not the three-class VADER sentiment.

Table 4.1: Demonstration of weighting scheme

	VADER	retweets	replies	favorites	count	weighted replies	weighted favourites	weighted retweets
1	0	0	0	0	2	0	0	0
2	-0.681	6	1	46	1	-0.681	-31.326	-4.086
3	0.5719	0	0	0	1	0	0	0
4	0	0	0	0	5	0	0	0
5	0.3612	3	1	6	1	0.3612	2.1672	1.0836

4.3 Weighting

As stated in hypothesis 3, we are interested in if an expert or a popular Twitter user increases the predictability in the stock return, i.e. examining the user dimension effects on stock return. In a way to capture the user dimension in the form of popularity, we have chosen to define popularity as the number of retweets, replies, and favourites from a user's posted tweet. Where a higher number of retweets most likely will have a bigger impact on the stock return of Tesla, while a tweet with no retweets will have a lower effect. Therefore, tweets inside a one-minute and a five-minute interval have been weighed against each other and then multiplied with the sentiment score, see equation 4.5 and table 4.1. The higher number of retweets a tweet has relative to the total number of tweets in the interval, the larger weight the tweet will be assigned. Therefore, we also partly account for hypothesis 2, as we include the number of tweets in the weighting. We perform the same weighting procedure for both replies and favourites. Thereby, more popular tweets have a higher effect on the return, still accounting for the quality (the sentiment) of a tweet.

$$sw_{i,j} = \frac{x_{i,j}}{x_{n,j}} \cdot s_{i,j} \tag{4.5}$$

Where $sw_{i,j}$ is the weighted sentiment for a tweet, i, based on j which is either retweets, replies, or favourites. The $x_{i,j}$ is $i's$ amount of $j's$ in the interval n, and $x_{n,j}$ is the total number of $i's$ in the interval n (one or five) for $j's$. $s_{i,j}$ is the score from the VADER sentiment or the manually classified sentiment. An example of how this is done is presented in table 4.1.

4.4 Machine Learning Techniques

Another well-known and commonly used approach in textual sentiment analysis, besides the dictionary-based approach, is the machine learning approach (Kearney and Liu, 2014). The concept of using machine learning in sentiment analysis is to connect statistical inference with the textual content (Li, 2010). Where, similar to the dictionary-based approach, classifying words into positive and negative sentiment using statistical inference is the key (Liu, 2012). Normally, the process is to start off sentiment classifying parts of the pre-processed text data, tweets, into negative, positive, or neutral. Either a pre-defined training set is used or the sentiment is manually classified, such that it becomes the training set. The training set is then trained onto a machine learning sentiment analysis algorithm, Naive Bayesian and Support Vector Machine algorithms are the ones most commonly used.

Further, when the algorithms have been trained on the training set it has learned the sentiment classification rules which are then applied onto the rest of the text data and finally derive sentiment scores on the full data set (Kearney and Liu, 2014).

There are some drawbacks with using a machine learning approach, such as the SVM algorithm. One drawback is the necessity to have a training set. Either one needs to get access to a "pre-defined" training set which is difficult. Alternatively, one needs to manually classify a training set which is very time-consuming (Hutto and Gilbert, 2014). Further, the quality of the classification depends on the people reading and classifying the text, which can have mixed results (Kearney and Liu, 2014). Despite these difficulties that may arise when conducting a machine learning approach, Li (2010) finds that the accuracy rate for the machine learning approach is most often higher compared to the dictionary-based approach, contrary to the findings of Hutto and Gilbert (2014).

The approach we use in this paper to predict the sentiment is the following. First, 1.605 tweets were randomly selected and manually classified by ourselves. The manual classification depends on the expressed attitude, opinion, and feeling of a tweet which could affect the stock price in different ways. Therefore, we checked the text and the attached pictures and links to get a better understanding of what the user wants to achieve with the tweet. We then classified the tweet into three valence-categories (negative (-1), neutral (0), positive (1)). Second, 70% of the manually classified tweets were used to fit a Support Vector Machine and tested on the other 30%. Third, all manually classified tweets are fitted with the best model to predict all other tweets.

4.4.1 Support Vector Machine

Support Vector Machines (SVM) are one of the most popular algorithms for classifying tweets. The SVM classifiers are distinctively different from the Naive Bayes classifier and other machine learning methods by using hyperplanes to separate data points, and not probability to classify the data (Hutto and Gilbert, 2014). Similar to the previous work by Smailović et al. (2014) we use the SVM to categorise texts of tweets into three different sentiment classes (positive, negative, and neutral). Their results suggest that including a neutral category improves the predictive power between tweets and the stock price. The above-mentioned positive results when employing the SVM instead of another machine learning approach is the reason for our choice of method.

The SVM classifies the class labels (sentiment) by the attributes (Tweet text) of a data set with the help of a separating hyperplane.[1] In this section, we explain the SVM for two classes, but SVM can be used for more than two classes, either by one-vs-one classification or one-vs-all classification. When applying the method, the data set with the manually classified tweets (class labels) needs to be divided up into a training and testing data set. Then the goal of the SVM is to classify all class labels in the training set correctly. The SVM does this by drawing a sep-

[1] A hyperplane in a p-dimensional space is a flat subspace with dimension p-1 that goes through the origin. For example in a two-dimensional space a hyperplane is a line.

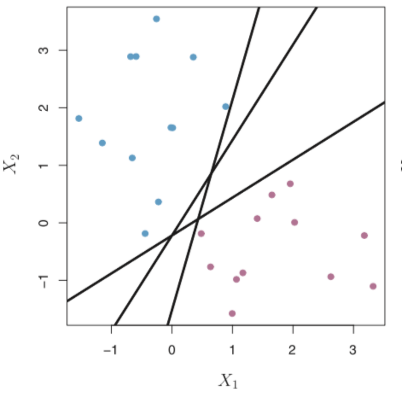

Figure 4.1: Visualisation of three different separating hyperplanes from James et al. (2013)

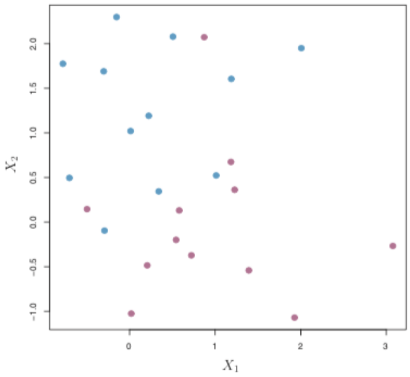

Figure 4.2: Visualisation of a data set without a separating hyperplane added from James et al. (2013)

arating hyperplane between the classes. Mostly there are infinite options to draw a separating hyperplane between the classes (see figure 4.1). Due to visualisation purposes, the figure only shows an example of two classes. To choose one of the infinite separating hyperplanes, one can use the hyperplane that is farthest away for the observations in the training set. Such that the distance between the data points and the hyperplane is minimised, which is called the margin. Then the maximal margin hyperplane is the hyperplane that has the farthest minimum distance to the observations. This classifier is called the *maximal margin* classifier and this classifier has the problem that it tends to overfit when p (the number of attributes) is large, and it can be sensitive to single observations. Another problem could be that a separating hyperplane may not exist as visualised in figure 4.3, again only for two classes. Alternatively, sometimes it would be desirable to misclassify an observation in the training set, to get a better result in the testing set, as seen in figure 4.3. Then it would be preferable to have the dashed line and not the solid line as the decision boundary, even with one misclassification observation. The following maximisation problem leads to the desired decision boundary and is called the Support Vector Classifier (James et al., 2013):

$$\max_{\beta_0,\beta_{11},\beta_{12}\ldots,\beta_{p1},\beta_{p2},\epsilon_1,\ldots,\epsilon_n,M} M \quad \text{subject to} \tag{4.6}$$

$$\sum_{j=1}^{p} \beta_j^2 = 1, \tag{4.7}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \geq M(1 - \epsilon_i), \tag{4.8}$$

$$\epsilon_i \geq 0, \tag{4.9}$$

$$\epsilon_i \leq C \tag{4.10}$$

Where M is the width of the margin. The $\beta' s$ specify the slope of the separating hyperplane for the two classes. C is the cost tuning parameter because it allows for misclassification. If C is high, we allow for a higher degree of misclassification. The other tuning parameter is ϵ, also called the slack parameter, as ϵ_i indicates where the i'th observation is located relative to the hyperplane and the margin. This specification only allows for a linear relationship. To allow for quadratic, polynomial,

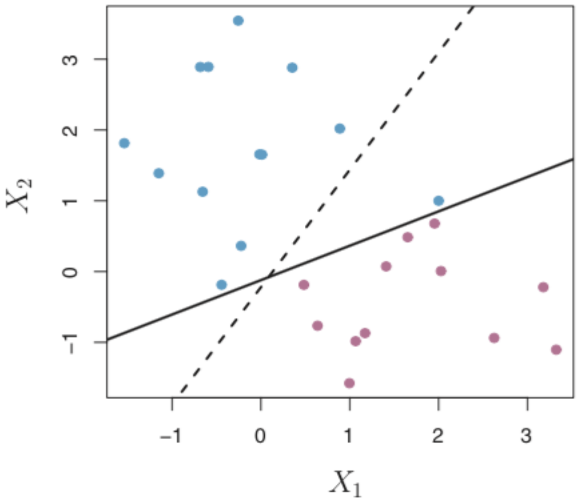

Figure 4.3: Visualisation of two different maximal margin hyperplanes. The solid line is the hyperplane with a misclassified observation, and the dashed line indicates a hyperplane without the misclassified observation. Added from James et al. (2013)

cubic or any other non-linear boundaries between the classes, equation 4.7 and 4.8 need to be changed, which results in the Support Vector Machine. Another solution for the Support Vector Classifier is to compute the hyperplane by using kernels[2], and for the Support Vector Classifier (equation 4.6 - 4.10) this is just the inner product (James et al., 2013):

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{n} x_{ij} x_{i'j} \tag{4.11}$$

Another form of a kernel is the polynomial that allows for non-linear boundaries, where d is a positive number:

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d \tag{4.12}$$

For matching other patterns, a radial kernel can also be used, formulated as:

$$K(x_i, x_{i'}) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2) \tag{4.13}$$

The basic procedure to select the best tuning parameters is described by Hsu et al. (2003). First, transform the data readable for the SVM package. Second, scale the data. Third, use a radial kernel. Fourth, use cross-validation to find the best value for the two tuning parameters. Fifth, use the best parameters to fit the model on the whole training set. We deviated from these steps a bit, firstly because our main focus lies not in the machine learning part but instead on the economic outcome. Secondly, our R package does not allow for some of these tests. Nevertheless, we tested for different kernels, classification methods and values for C. In our case, the radial kernel with $C = 1$ produced the highest accuracy with around 77%. We

[2] A kernel quantifies the similarities between two observations

are aware of that we could have improved the accuracy of our results by manually classifying more observations using special stopwords for our data set, or following more closely the steps suggested by Hsu et al. (2003). Although this may result in less precise results than desired as well as a misclassification bias, the method and approach chosen are still appropriate given the focus of our thesis.

4.4.2 Lasso Regression

The choice of using the least absolute shrinkage and selection operator (lasso) method rather than a similar ridge regression, or a VAR model was simply because of the lasso's variable selection allowing coefficients to be equal to zero and not only close to zero as in the case of the ridge regression. Therefore, the lasso would provide a better model since it excludes variables with no predictability from the model, and only the variables with predictability remain. Employing a VAR model was also considered, but due to correlated independent variables, it was not a plausible method, whereas the lasso deals with a correlation problem by shrinking some of the correlated variables equal to zero. Additionally, the VAR model falls short in selecting the number of lags, compared to the lasso, as the number of lags is selected for the whole model and not for each variable. As a result, a lasso regression model is the model of choice.

The major difference between the lasso and a standard OLS regression is that it does not only minimises the residual square error. Instead, it minimises the following equation for β:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_o - \sum_{j=1}^{p} {}_j x_{ij} \right)^2 \quad \text{subject to} \quad \lambda \sum_{j=1}^{p} |\beta_j| \tag{4.14}$$

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4.15}$$

The constraint of the equation - also called the penalty term - allows the model to shrink variables equal to zero, which have no impact on y, depending on the choice of λ,[3] and so the selection of a good λ is crucial. Therefore, the algorithm tries different λ's and picks the λ that either produces the smallest measurement error or yields the lowest information criteria (cross-validation). If for example, λ is equal to zero, then the model will just be an OLS, without any shrinkage. Whereas for a $\lambda = \infty$ all coefficients will be equal to zero, also called a null model. The idea of the lasso is to sacrifice a small bias to reduce the variance of the predicted values to improve the overall prediction accuracy (Tibshirani, 1996). An essential condition for the lasso is that the data has a zero mean and variance one. Therefore, the data is standardised.[4] Otherwise, a parameter with a bigger scale will not shrink towards zero as fast as a predictor with a lower scale (James et al., 2013).

To apply the lasso we need to modify the data sets. The main concern with the

[3] Positive number

[4] The R package glmnet standardises the data before the cross-validation and fitting, but rescales the coefficients afterwards back.

lasso is that it cannot deal with missing values, which means we need to shorten the data set. Resulting in that we need to remove all observations that are not within a trading day or outside of the trading hours. To still be able to account for lags, the starting frame has to be moved later in time by the number of lags included. Furthermore, all observations with missing values, for example, explained by Twitter being down, have been set equal to zero (James et al., 2013). Otherwise, there will be a shift in the lags, which could change the results, such that, we decided, setting them equal to zero is the better trade-off.

4.5 Limitations

Our results of the SVM should be viewed with caution as we only coded 1605 tweets, which results in a bad prediction, and we will have a misclassification bias later when applying the lasso. When running the lasso we decided to apply a linear regression on a highly complex problem, for example, Oh (2002) states that stock indices have a non-linear behaviour, and thus the author is suggesting using an artificial neural network. We decided not to do this because neural networks are "black boxes", such that it is not clear how the technique is computing. Additionally, this method would be above the scope of this thesis. To account for lags in the lasso regression, we have to shorten the data sets, and correspondingly cannot test for lags further in the past. As a result, we cannot say anything about lags further in the past. Another limitation with the data set is that the variables are not highly correlated, and thus a regression will most likely not find any realistic results. A comparison with other companies is not applicable, as mentioned already in the data section since Tesla is a much-discussed topic and even though controlling with the cashtag, may not reduce the noise.

5 Results and Discussion

In this section we present the empirical results. We split the section into three parts, where the first part regards our basic model with the VADER sentiment and the return on two different data sets. While in the second part we use the SVM to predict the sentiment and run a lasso regression. We finalise this section with a comparison of both approaches.

5.1 Basic model – Results

5.1.1 Full sample

Granger Causality – Results

First, we run an Augmented Dickey-Fuller test to check if all variables are stationary to later run a Granger causality test. The results are presented in table 5.1. We present the results for the VADER sentiment and return for certain lags. Nevertheless, we test for all lags between 1 and 60 where the p-value is smaller than 0.01 for all lags, indicating that we reject the null hypothesis and conclude that the two variables, VADER and return, are stationary in both intervals. We assume that if the variable is stationary up until lag 60 this will also be the case for all lags bigger than 60. Similarly, we conduct the test for the weighted variables and the *count* variable, too, which are all stationary. We do not test for the dummy variable, as it is by default stationary.

Next, we test if there exists a causal relationship between the VADER sentiment and the return on an intraday level. From the theory and previous literature section we know that sentiment can Granger cause stock returns with a lag of a few

Table 5.1: Results of the Augmented Dickey–Fuller test

	1-minute				**5-minute**			
Variable	*VADER*		*Return*		*VADER*		*Return*	
Lag	Dickey-Fuller	p-value	Dickey-Fuller	p-value	Dickey-Fuller	p-value	Dickey-Fuller	p-value
1	-63.187	0.01***	-65.653	0.01***	-26.794	0.01***	-31.292	0.01***
5	-63.187	0.01***	-65.653	0.01***	-26.794	0.01***	-31.292	0.01***
10	-63.187	0.01***	-65.653	0.01***	-26.794	0.01***	-31.292	0.01***
20	-63.187	0.01***	-65.653	0.01***	-26.794	0.01***	-31.292	0.01***
40	-63.187	0.01***	-65.653	0.01***	-26.794	0.01***	-31.292	0.01***
60	-63.187	0.01***	-65.653	0.01***	-26.794	0.01***	-31.292	0.01***

Notes: *** p<0.01, ** p<0.05, * p<0.1

Results from the Augmented Dickey-Fuller test for the one- and five-minute interval to check for stationarity. We also tested for all lags between 1 and 60 without any differences in the results. Both variables are stationary in the one- and five-minute interval.

days (Smailović et al., 2014; Sul et al., 2017), this leads us to our stated hypothesis 1. Therefore, the following null hypothesis is tested for the Granger causality test: *Sentiment does not Granger causes returns*. When conducting the Granger causality test we find that sentiment does not Granger cause return in the interval between 1 and 300 minutes (five hours) on the one- and five-minute interval data, and thus we cannot reject the null hypothesis. As a result, we need to reject hypothesis 1. However, we find significant effects for reverse causality on the one- and five-minute interval, meaning that return Granger causes sentiment.[1] This is contradicting our theory and the previous literature, as the authors find that sentiment Granger causes stock returns and not the other way around. This means that for the data set with the full search query, Twitter user react to sudden changes in the return and then tweet about it. This indicates that a Twitter user may be interested in the development of the company, which results in her tweeting about it but does not buy or sell the Tesla stock. Another outcome from these results could be that there exist no relationship between Twitter and the Tesla stock return on an intraday level, which means that the people investing in Tesla are taking their time before going to action and do not base their investment decisions on sudden changes in the Twitter sentiment. Further, it could be that there is a lot of noise in the data set with the full search query and this does not capture anything related to the Tesla stock. This is why we conduct the same tests in the next chapter where we only use the cashtag.

An interesting side note is that we only find significant results for lag five to nine for the one-minute interval, whereas, we do not find a significant effect for lag one (5 minutes) in the five-minute interval. A possible explanation to why this may be, is that lag six to nine belongs to lag two and lag five to lag one in the five-minute interval. Although, we do not find a significant effect for the second lag in the five-minute interval. A conclusion that could be drawn from this is that in the five-minute interval we lose information about the sentiment and the one-minute interval is more accurate. Another thing to keep in mind with these results is that we test for a lot of lags and only find some significant lags. From basic statistics, it is not unlikely to find some significance just by randomness. We try to improve our empirical analysis by conducting a lasso regression.

Lasso Regression– Results

As the Granger causality test implies that we should not find any predictive power in the VADER sentiment for the return, we conduct a lasso regression to confirm this, such that we expect that the coefficients are zero for all lags. The main reason for us to run the lasso regression is, however, that we want to test for hypothesis 1 to 3. Therefore, we weighted the replies, retweets, and favourites with the procedure explained in the method section. We then added a dummy variable stating if Elon Musk (hypothesis 3) tweeted in the observed interval or not. Further, we include a count variable which counts the number of tweets in the interval (hypothesis 2). We decided not to include a variable accounting for tweets with the cashtag as we use an extra data set to check for hypothesis 4. To conduct a lasso regression, we modified the data, as explained in section 4.4.2. We choose to check for the first ten minutes

[1] We changed the data set, such that we add observations after the closing hour and do not have the sentiment starting until at 07:00 the next day

Table 5.2: Results of the Granger Causality Test full data set

	1-Minute					**5-Minute**			
Lag	*VADER Granger Return*		*Return Granger VADER*		*Lag*	*VADER Granger Return*		*Return Granger VADER*	
	F	p-value	F	p-value		F	p-value	F	p-value
1	0.0023	0.9617	0.1462	0.7022	1	0.9334	0.334	0.9585	0.3276
5	0.4868	0.7864	2.5664	0.0250*	6	1.0009	0.4227	0.4326	0.8576
10	0.8125	0.6166	1.3994	0.9681	12	0.307	0.9884	0.3888	0.2390
50	0.6689	0.9650	1.1086	0.2781	24	0.6878	0.8684	0.5153	0.9752
100	1.0463	0.3573	0.8499	0.8566	36	0.7253	0.8863	0.5331	0.9899
200	1.0472	0.3156	0.9262	0.7582	48	0.7498	0.8969	1.6793	0.0027**
300	1.1029	0.3156	0.9302	0.7594	60	0.7699	0.8999	1.4386	0.0181*

Notes: *** p<0.01, ** p<0.05, * p<0.1

Results from the Granger causality test for the one- and five-minute interval. Also tested for all lags between 1 and 300 for the one-minute interval and lags 1 to 60 for the five-minute interval, without any big changes in the results. We only find significant results that return Granger causes sentiment, for the one-minute and five-minute interval, but for different times (one-minute interval: five minutes; five-minute interval: between five and six hours.

as we still want to use a sufficient amount of lags but do not want to lose too many observations. For example, investors already use Twitter to predict stock returns, and they react instantaneously to new tweets.[2] For the lasso regression on the one-minute interval, we find that all of the estimated coefficients are equal to zero (see table B.1 column (1)). Meaning that the lasso has shrunk them all to zero, stating they have no predictive power. The development of the coefficients is visualised in figure A.3. We can see that the *elon* variables are the last variables to shrink to zero, indicating that they have the highest impact of all variables. We are therefore able to reject all three hypotheses, and merely taking the expected return would be the best model to predict the return. As explained earlier, these results are not surprising since we have already found that sentiment does not have any impact on the return in the Granger causality test. Further, weighting favourites, replies, and retweets with the VADER sentiment did not improve the predictive power for the model since the coefficients are equal to zero. Similar, the coefficients for the *count* variables, as well as for the *elon* variables, are zero and have no predictive power. There could be several reasons for this; for example, Elon Musk only tweeted 66 times during the observed period, and it may not be sufficient for the *elon* variables to predict the return. Furthermore, it could be that all variables suffer from too much noise in the Twitter data, i.e. information that is not related to the Tesla stock price, and the used data is, therefore, unsuitable to use for predicting Tesla stock returns.

We run the same lasso regression on the five-minute interval, where we are only estimating two lags to get the same time frame as for the one-minute interval. From the results we can again see that all coefficients are equal to zero and hypotheses 1 to 3 are rejected for the five-minute interval, too (for visualisation see figure A.5). Again, this is not surprising since we do not find any Granger causality for the VADER sentiment on the return.

[2] This means our data set starts 15:40 every day and ends at 21:59 CET/CEST

5.1.2 Cashtag sample

Granger Causality – Results

As mentioned in the data section 3.1, contrary to the previous research (Smailović et al., 2014) we do not only include the cashtag of Tesla to create the VADER sentiment and instead include a broader search query. A small visualisation in the differences in the wordclouds is presented in the Appendix A. To test if the use of a different data set will affect our results, we conducted the same data cleaning procedure and tested for stationarity[3] with only tweets containing the cashtag "$TSLA". Furthermore, we also want to test for hypothesis 4, i.e. if information richness in the tweets can increase the predictability for the stock returns. The results from the Granger causality test are presented in table 5.3 and shows that VADER sentiment Granger causes return for the first ten lags, and after lag ten there is no significance on the one-minute interval. This is in line with hypothesis 4, only including observations with a cashtag increases the models predictability when comparing the results to table 5.2. Additionally, the results from table 5.2 indicate that the whole data set with the full search query presented in chapter 3.1 contains a lot of noise and is thereby not well-suited to predict the Tesla stock return. Based on previous literature, it is not unexpected that we find that the VADER sentiment can Granger cause return when only using tweets with the cashtag. It is also of no surprise that we find significant results for the first ten lags, but none after lag ten, because traders partly already work with Twitter to predict stock market movements by using algorithms to react to tweets. This is also why we find the lowest p-value in the one-minute lag since these algorithms react almost instantaneously. The conclusion that can be made from these results is that a data set only including cashtags is probably more suitable for predicting stock price movements compared to tweets related to the company, as these seems to contain a lot of noise unrelated to the stock. Another take on this could be that the used algorithms only read tweets with a cashtag and not all tweets related to Tesla as a company. The mutual Granger causality on the first lag might indicate that both variables have an effect on each other. But since the p-value is on a higher significance level for VADER sentiment Granger causing return, we conclude that this is the direction of the Granger causality, as well as only lag one for return Granger cause VADER sentiment is significant.

When comparing the results for the one- and five-minute interval, it is surprising that we do not find any significant results on the first and second lag on the five-minute interval since we find significant results in that time span for the one-minute interval. One reason may be that we lose information by taking the average of the VADER sentiment and thereby getting rid of a lot of information to predict stock returns. Hence, we presume that the shorter time interval the more information is included to predict stock returns. When comparing our results with previous literature, Twitter data has mostly been used to predict daily stock price movements. One should therefore keep in mind that by using daily data they may lose a lot of information to predict stock price movements, as we seem to be affected by it when only switching from a one-minute to a five-minute interval. A more reasonable outcome for the differences could be that we are only able to capture the effect from

[3] We could again reject the null hypothesis and confirm that we have stationary data.

Table 5.3: Results of the Granger causality test for the cashtag sample

	1-Minute					**5-Minute**			
Lag	*VADER Granger Return*		*Return Granger VADER*		*Lag*	*VADER Granger Return*		*Return Granger VADER*	
	F	p-value	F	p-value		F	p-value	F	p-value
1	9.675	0.002***	3.249	0.071*	1	0.040	0.842	1.157	0.282
2	4.943	0.007***	2.271	0.103	6	1.348	0.232	1.404	0.209
3	3.527	0.014**	1.493	0.214	12	0.794	0.657	1.408	0.154
4	2.687	0.030**	1.313	0.263	24	0.745	0.809	0.969	0.505
5	2.228	0.049**	1.076	0.371	36	0.842	0.735	0.907	0.629
6	2.173	0.043*	1.109	0.354	48	0.729	0.919	1.027	0.422
7	2.069	0.043*	1.120	0.347	60	0.726	0.945	0.917	0.657
8	1.822	0.068*	0.996	0.437					
9	1.800	0.063*	0.973	0.460					
10	1.711	0.072*	0.921	0.512					

Notes: *** p<0.01, ** p<0.05, * p<0.1

Results from the Granger causality test when only using the cashtag for the one- and five-minute interval. Also tested for all lags between 1 and 300 for the one-minute interval and between 1 and 60 for the five-minute interval. For the one-minute interval there is no significance after the tenth lag when VADER Granger causes return. For the five-minute interval there is no significant results at all.

the institutional investors which uses algorithms to buy and sell stocks. Whereas, previous literature found the effects Twitter has on non-institutional investors, who read a tweet on Twitter and then think about it for a couple of days and then make the decision to invest.

Lasso Regression – Results

As already explained in the full sample above, we conduct the same lasso analysis. The only difference is that we do not include a dummy for Elon Musk, as none of his tweets includes ”$TSLA” in the observed period. We present the results for the one-minute interval in table B.1 column (3). The obtained results are similar to the ones in the full data set, except for the variable *count_2* since it is now non-zero. The sign has the expected positive sign, such that it is in line with hypothesis 2. To get the variable development depending on λ in more detail, see figure 5.1. We can see that the variable *count_2* only reaches zero for a bigger λ than the optimal given one. The optimal λ is displayed by the vertical line (s=0.0052) and the *count_2* variable being equal to 0.0014. The y-axis is displaying the logarithm of 10^{λ}, and thus the axis is inverted. We can see that some of the *vader* variables are close to being non-zero. However, the results are creating a surprising outcome, as we find significant coefficients for the first ten lags of the Granger causality test for the *VADER* variable, indicating a predictive power for the *VADER* variable on return. Nonetheless, neither the lags of the *vader* variable nor the other weighted variables, have any predictive power in the lasso regression. As a result, we cannot confidently confirm hypothesis 4, stating that the inclusion of a cashtag will increase the predictive power for the Tesla stock return. The results get even more spurious, as Hecq et al. (2019) states that the lasso tends to select too many variables, but it still cannot confirm the results from the Granger causality test. A reason for why we get different results could be that we did not weight the lags of the *vader* variable,

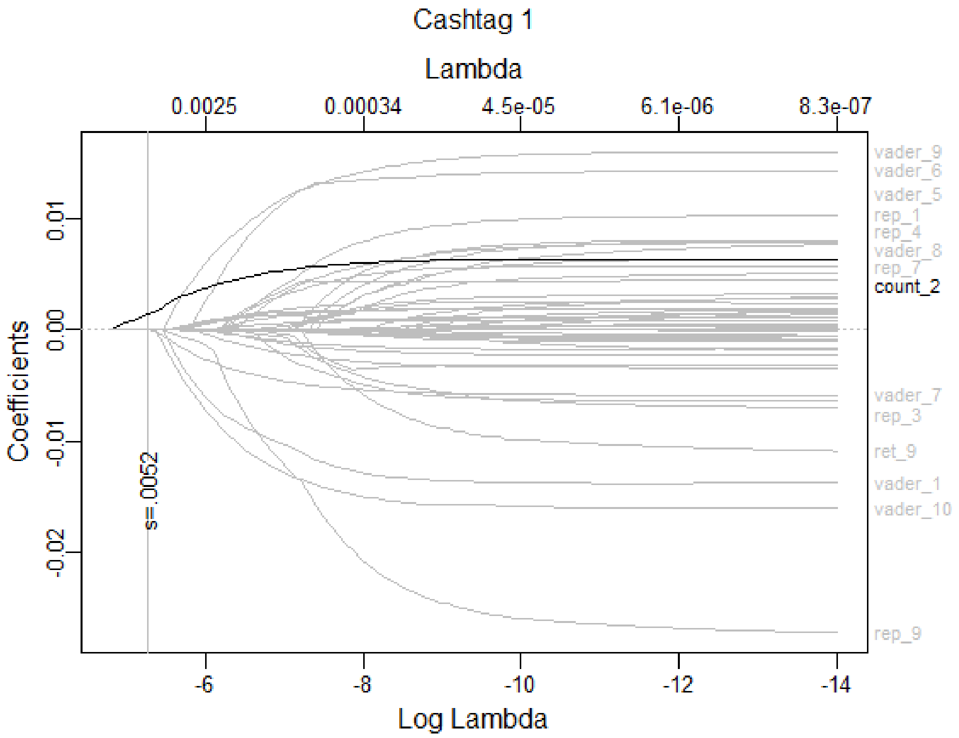

Figure 5.1: Value of the coefficients depending on the value of log(10^{λ}) *for the cashtag sample on the one-minute interval.*

such that the regression would be called an adaptive lasso regression, named after Zou (2006).[4] We could base the weighting either on the results from the Granger causality test or from the shrinkage graphs presented in the Appendix A. So we could have put more weight on the significant variables from the Granger causality test as the test implies that there exists a relationship between the lags of *vader* and *return*. Another reason for why we find different results may be that we tested the Granger causality on a bigger data set, where we did not remove the first ten observations of the return. And it could be that the first ten observations of a day are important for Twitter sentiment to predict the return, as the trading volume is high during the first minutes of market opening.

We present the results for the five-minute interval in table B.2 column (3). We can see that variable *ret_2* is non-zero and negative (visualised in figure A.6). Meaning that the weighted retweets with two lags have a negative impact on the return. The results are partly contradicting hypothesis 1 since we multiply the variable *ret_2* with the VADER sentiment. However, it is strengthening hypothesis 3 because the more retweets a tweet has – assuming the number of tweets and VADER is constant – the more negative the return will be. Implying a negative relationship, which we did not expect. A reason for this could be that negative tweets get retweeted more often and so investors react to these. Nevertheless, we cannot explain the difference between the one- and five-minute interval.

[4] By weighting certain variables in the lasso regression, it takes longer for these variables to shrink. Thus they will not necessarily become equal to zero.

5.2 Machine Learning model

In this section, we present the results of the manually classified sentiment, as explained in chapter 4.4.1. Here we want to verify if manual classification provides superior results, as there is no consensus in the literature (Hutto and Gilbert, 2014; Li, 2010). We do not manually classify our data set only including the cashtag into three valences because of time constraints. A comparison between the cashtag data set and the manually classified data set is therefore not suitable.

Granger Causality – Results

First, we run an Augmented Dickey-Fuller test, which shows that we have stationary data. Second, we conduct the Granger causality test, to test if the variables *elon, fav, sent, rep, ret* or *count* Granger causes *return*, and we also test for the other way around . The results for the one-minute interval with only sentiment and return are presented in table 5.4, nevertheless we also tested for the other variables. Looking at the one-minute interval, we see that for lag 200 the coefficient is significant and also for some lags around 200. But we also find significant results of reversed causality, for nearly all the same lags, which means we have mutual causality. We can conclude that both variables are highly interconnected and influence each other. When looking at the other variables, especially *elon* produces spurious results where around two thirds of the lags are significant, which is unrealistic. A possible explanation for this could be that the dummy variable only has a few observations of ones. Hence, we assume that these results do not represent the true relationship between tweets from Elon Musk and the return of Tesla. Regarding the other variables, we do not find substantial results to state that we have a clear Granger causality, we run the tests but they are not presented in tables.[5] For the five-minute interval we find that sentiment Granger causes return for all lags, and we find no evidence of reversed causality. To conclude, we have found Granger causality for both intervals and in the expected direction, but the results are also spurious as we find significant results for all lags on the five-minute interval. Additionally, we find mutual causality on the one-minute interval, such that we cannot confidently state that sentiment can be used to predict stock returns.

Lasso Regression – Results

As the Granger causality test does find meaningful results, we again run a lasso regression to test for the same hypothesis as in the previous section. The results for the one-minute interval are presented in table B.1 column (2) and visualised in figure A.4. As for the two other data sets, we have to reject hypotheses 1 to 3, as all coefficients shrink to zero and the expected return would be the best model. A reason for why we do not find any predictive power could be that our manual classification is not good enough, and we were not able to classify the tweets sufficient to filter the noise out of the data set. The results are the same for the five-minute interval, where the lasso shrinks all the coefficients to zero as they do not have any predictive power presented in table B.2 column (2) and visualised in figure A.7. The reason for these results could be that there is no valuable information in the data set to predict the

[5] All tables are available upon request.

Table 5.4: Results of the Granger Causality Test for the manually classified data set.

	1-Minute					**5-Minute**			
Lag	*Sentiment Granger Return*		*Return Granger Sentiment*		*Lag*	*Sentiment Granger Return*		*Return Granger Sentiment*	
	F	p-value	F	p-value		F	p-value	F	p-value
1	1.106	0.293	0.5553	0.4562	1	15.425	0.0001***	0.1741	0.6765
5	0.7951	0.553	1.6581	0.141	6	3.1241	0.0047***	1.0659	0.3805
10	0.8437	0.5862	1.4031	0.1717	12	2.5331	0.0025**	1.3137	0.2027
50	1.1731	0.1881	1.1382	0.2337	24	2.0221	0.0023***	1.1843	0.2433
100	1.131	0.1754	1.0716	0.2945	36	1.5916	0.0139**	1.0989	0.3149
200	1.2087	0.0239**	1.2119	0.0223**	48	1.5889	0.0061***	1.0152	0.444
300	1.0859	0.1485	1.2066	0.0086***	60	1.4548	0.0129**	1.0415	0.3888

Notes: *** p<0.01, ** p<0.05, * p<0.1

Results from the Granger causality test on the manual classified data set for the one- and five-minute interval. Also tested for all lags between 1 and 300 for the one-minute interval and between 1 and 60 for the five-minute interval. For the one-minute interval we find that for the lags around 200 we have mutual significance, whereas for the five-minute interval all lags for sentiment Granger causes return are significant.

stock return. Additionally, the results from the lasso regression are implying that the results from the Granger causality test are not correct. Otherwise, the variables *sent_1* and *sent_2* would not be zero. We cannot say something about the one-minute interval, as we did not test for lag 200, but as we do not find results in the lasso regression that confirms the results of the five-minute interval, we assume that the outcomes from the one-minute and five-minute interval of the Granger causality test are spurious.

5.3 Summary of the results

As seen in the previous sections 5.1 and 5.2, we do not find any predictive power in the variables when conducting the lasso regression. With two exceptions for *count_2* on the one-minute interval and *ret_2* on the five-minute interval in the cashtag sample. An explanation for the lack of predictive power could be the missing correlation brought up in section 3, visualised in figure 3.2. The lasso regression is unsuccessful in capturing the effects, as there is a lack of correlation between the return and the explanatory variables. Including the cashtag increases the predictive power, but we believe that two different variables that are non-zero on different intervals, is not enough to confirm hypothesis 4. Therefore, we do not put too much weight on this finding, as all other variables are non-zero in the lasso regression. In the Granger causality test, we have seen that the cashtag sample produces significant results that do match with theory and also with real-life, whereas we do not find the same results for the full sample with the VADER sentiment nor with the manual classified sentiment. Which let us conclude that, as mentioned earlier, the full sample probably contains too much noise, not related to the stock return. The manual classification could have filtered that noise out, but as we only classified 1605 tweets, this filtering was not successful. We cannot explain the differences between the lasso and the Granger tests, but as the Granger produce inconsistent results, we believe that the lasso regression generates better outcomes for our setup using linear models. As mentioned in the method section, we only conduct a linear

Granger causality test, and as Tank et al. (2018) state when applying linear Granger on non-linear variables, the results will be inconsistent. In the end, we conclude that we cannot predict the Tesla stock return with the used Twitter data on an intraday level for the given period, and we reject all four hypotheses. Additionally, we have shown that the EMH holds, as the Twitter information is already incorporated in the Tesla stock price.

6 Conclusion

Previous studies have used Twitter data to predict the stock returns of several companies. This study sets out to investigate the impact of different dimensions that tweets have on the return of Tesla, Inc. Therefore, we laid out our thesis to apply a dictionary-based approach (VADER) and a machine learning approach (SVM) on Twitter data to obtain a sentiment score. After this process, we match the Twitter data with the Tesla stock price on a one-minute and a five-minute interval. Also, we perform the same procedure as for the dictionary-based approach on a smaller Twitter data set, including only the cashtag "$\$TSLA$". With the setup we choose, we believe in broadening the knowledge in the relationship between Twitter and stock returns, as we evaluate the relationship on two different intraday levels. We also examine how the dimensions of a tweet affect the relationship. We analyse the data sets employing two different methods to get the results: A Granger causality test and a lasso regression. With the use of these methods, we increased the validity in the results as we do not solely rely on one method for our results. We analysed over 600 000 tweets from October 1, 2019, to December 30, 2019, and removed all non-trading days.

From theory and previous research, presented in section 2, we assumed to find that we can use sentiment to predict stock return. Additionally, we expected that including retweets, favourites, and replies as indicators for the popularity of a user will increase the accuracy of the prediction, along with controlling for experts of the company. Also, we thought, checking for the number of tweets will help to predict the stock return. As a result, we formulate four different hypotheses. In contradiction to our hypotheses, this study does not find enough evidence to confirm either of these hypotheses as the lasso regression finds that no variable is useful in predicting the stock return, except for two specifications in the cashtag data set. However, we believe that these results are not strong enough to confirm hypothesis 4 and especially none of the others. The results may hint that a cashtag is increasing the information richness. For the Granger causality test, we find some significant results, but they do not show any clear pattern in the results, such that we conclude that this is not the true relationship. As a result, we reject all hypotheses, and we cannot predict the stock return.

For future research, we recommend using a non-linear model to predict the stock returns with the used variables, as this is most likely the reason for why we do not find enough non-negative coefficients in the result section for the lasso. Plus, a recommendation is only to use tweets with a cashtag, as they tend to contain a slightly higher richness of information and we believe that there should be a relationship between the tweets and the return.

Bibliography

Asur, S. and Huberman, B. A. (2010), Predicting the future with social media, *in* '2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology–Workshops', IEEE Computer Society, [Los Alamitos, Calif.], pp. 492–499.

Benthaus, J. and Beck, R. (2015), 'It's more about the content than the users! the influence of social broadcasting on stock markets', *ECIS 2015 Completed Research Papers* .

Bollen, J., Mao, H. and Zeng, X. (2011), 'Twitter mood predicts the stock market', *Journal of Computational Science* **2**(1), 1–8.

Das, S. R. and Chen, M. Y. (2007), 'Yahoo! for amazon: Sentiment extraction from small talk on the web', *Management Science* **53**(9), 1375–1388.

Enders, W. (2014), *Applied econometric time series*, fourth edn, Wiley, Hoboken, NJ.

Fama, E. F. (1965), 'The behavior of stock-market prices', *The Journal of Business* **38**(1), 34–105.
URL: *www.jstor.org/stable/2350752*

Ferris, R. (2018), 'Tesla shares surge 10% after elon musk shocks market with tweet about going private'. https://www.cnbc.com/2018/08/07/tesla-says-no-final-decision-has-been-made-to-take-company-private.html?fbclid=IwAR3si-r4k_BBQz_I-64sPN0u4TIScfr5rdVDj2ARsT95nb3J2-JrlCpT9_E [Accessed:20.05.2020].

Granger, C. W. J. (1969), 'Investigating causal relations by econometric models and cross-spectral methods', *Econometrica* **37**(3), 424.

Hecq, A., Margaritella, L. and Smeekes, S. (2019), 'Granger causality testing in high-dimensional vars: a post-double-selection procedure'.
URL: *http://arxiv.org/pdf/1902.10991v3*

Hsu, C.-W., Chang, C.-C. and Lin, C.-J. (2003), A practical guide to support vector classification, Technical report, Department of Computer Science, National Taiwan University.
URL: *http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf*

Hutto, C. J. and Gilbert, E. (2014), Vader: A parsimonious rule-based model for sentiment analysis of social media text., *in* E. Adar, P. Resnick, M. D. Choudhury,

B. Hogan and A. H. Oh, eds, 'ICWSM', The AAAI Press.
URL: *http://dblp.uni-trier.de/db/conf/icwsm/icwsm2014.htmlHuttoG14*

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, Vol. 103, Springer New York, New York, NY.

Kearney, C. and Liu, S. (2014), 'Textual sentiment in finance: A survey of methods and models', *International Review of Financial Analysis* **33**, 171–185.

Korosec, K. (2020), 'Tesla shares fall on elon musk 'stock price too high' tweet'. `https://tcrn.ch/2Ba771Q` [Accessed: 28.05.2020].

Lawrence, E. R., McCabe, G. and Prakash, A. J. (2007), 'Answering financial anomalies: Sentiment-based stock pricing', *Journal of Behavioral Finance* **8**(3), 161–171.

Li, F. (2010), 'The information content of forward-looking statements in corporate filings-a naïve bayesian machine learning approach', *Journal of Accounting Research* **48**(5), 1049–1102.

Li, Z., Fan, Y., Jiang, B., Lei, T. and Liu, W. (2019), 'A survey on sentiment analysis and opinion mining for social multimedia', *Multimedia Tools and Applications* **78**(6), 6939–6967.

Liu, B. (2012), 'Sentiment analysis and opinion mining', *Synthesis Lectures on Human Language Technologies* **5**(1), 1–167.

Lo, A. W. (2004), 'The adaptive markets hypothesis', *The Journal of Portfolio Management* **30**(5), 15–29.

Loughran, T. and McDonald, B. (2011), 'When is a liability not a liability? textual analysis, dictionaries, and 10-ks', *The Journal of Finance* **66**(1), 35–65.

Mao, Y., Wei, W., Wang, B. and Liu, B. (2012), Correlating s&p 500 stocks with twitter data, *in* X. Fu, P. Gloor and J. Tang, eds, 'Proceedings of the first ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research', ACM, New York, NY, pp. 69–72.

Oh, K. (2002), 'Analyzing stock market tick data using piecewise nonlinear model', *Expert Systems with Applications* **22**(3), 249–255.

Oritz-Ospina, E. (2019), 'The rise of social media'. `https://ourworldindata.org/rise-of-social-media` [Accessed: 28.05.2020].

Rui, H., Liu, Y. and Whinston, A. (2013), 'Whose and what chatter matters? the effect of tweets on movie sales', *Decision Support Systems* **55**(4), 863–870.

Shiller, R. J. (2003), 'From efficient markets theory to behavioral finance', *The Journal of Economic Perspectives* **17**(1), 83–104.
URL: *www.jstor.org/stable/3216841*

Smailović, J., Grčar, M., Lavrač, N. and Žnidaršič, M. (2014), 'Stream-based active learning for sentiment analysis in the financial domain', *Information Sciences* **285**, 181–203.

Sprenger, T. O., Tumasjan, A., Sandner, P. G. and Welpe, I. M. (2014), 'Tweets and trades: the information content of stock microblogs', *European Financial Management* **20**(5), 926–957.

Sul, H. K., Dennis, A. R. and Yuan, L. I. (2017), 'Trading on twitter: Using social media sentiment to predict stock returns', *Decision Sciences* **48**(3), 454–488.

Tank, A., Covert, I., Foti, N., Shojaie, A. and Fox, E. (2018), 'Neural granger causality for nonlinear time series', *arXiv: Machine Learning* .
URL: *http://arxiv.org/pdf/1802.05842v1*

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Tirunillai, S. and Tellis, G. J. (2012), 'Does chatter really matter? dynamics of user-generated content and stock performance', *Marketing Science* **31**(2), 198–215.

Yu, Y., Duan, W. and Cao, Q. (2013), 'The impact of social and conventional media on firm equity value: A sentiment analysis approach', *Decision Support Systems* **55**(4), 919–926.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**(476), 1418–1429.

Appendix A Additonal Figures

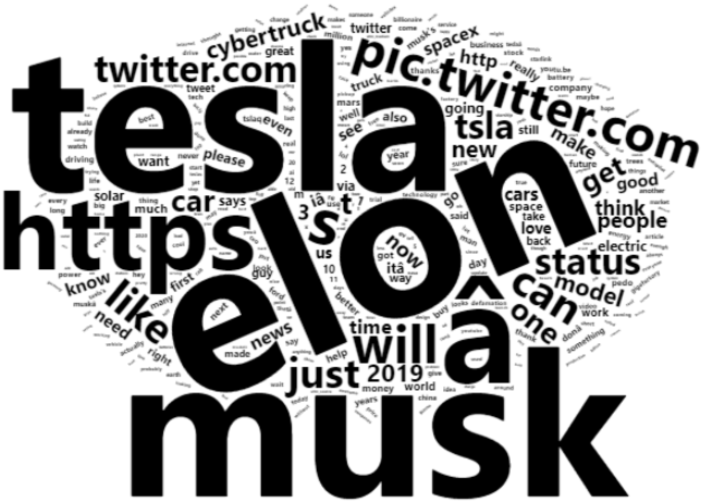

Figure A.1: Most used words for the whole data set

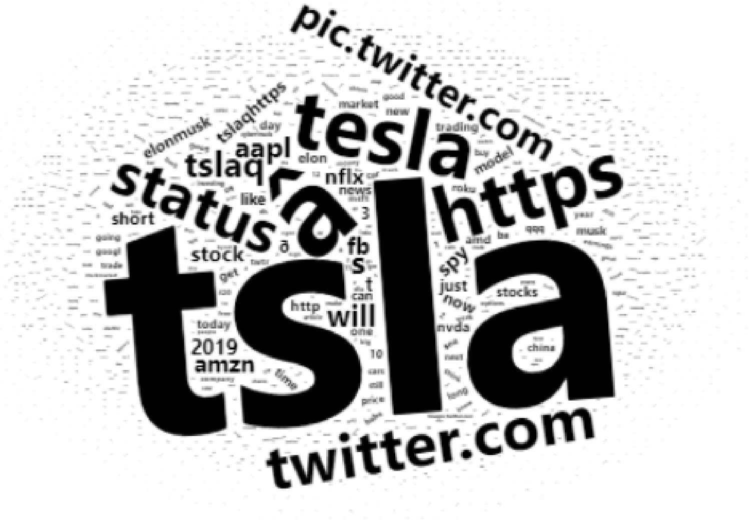

Figure A.2: Most used words for the cashtag data set

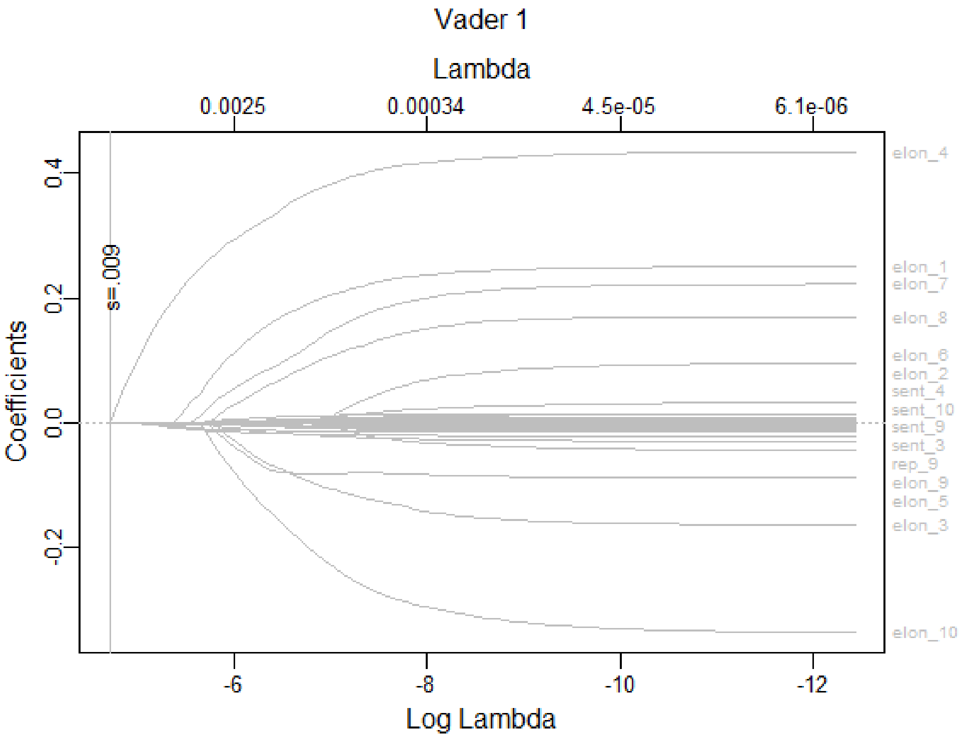

Figure A.3: Value of the coefficients depending on the value of log(10^{λ}) for the full data set on the one-minute interval.

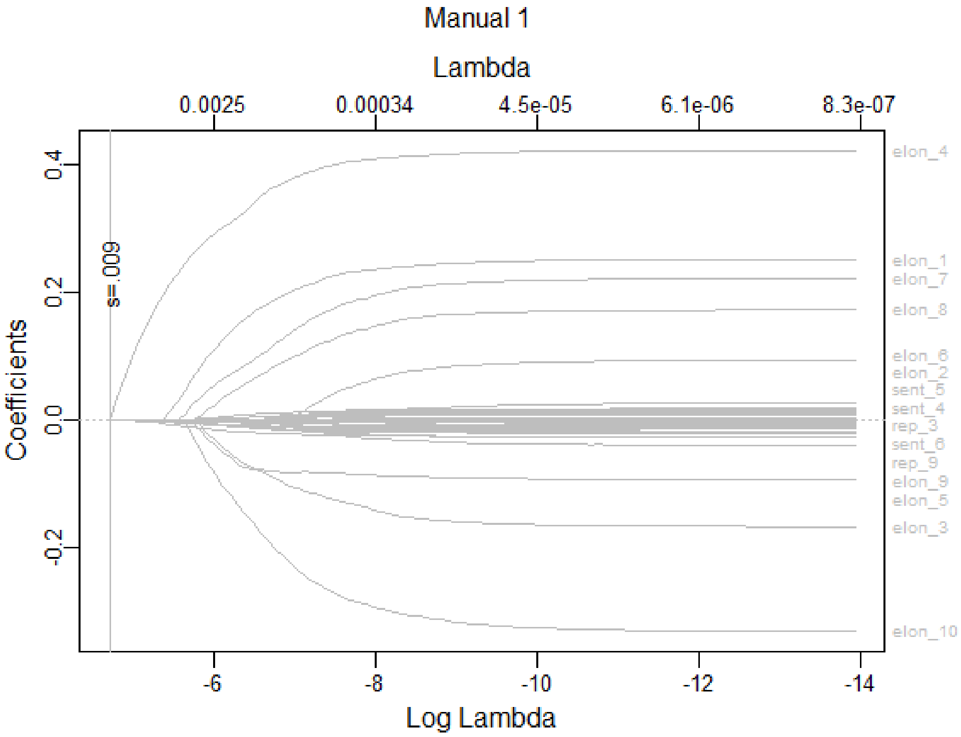

Figure A.4: Value of the coefficients depending on the value of log(10^{λ}) for the manually classified sample on the one-minute interval.

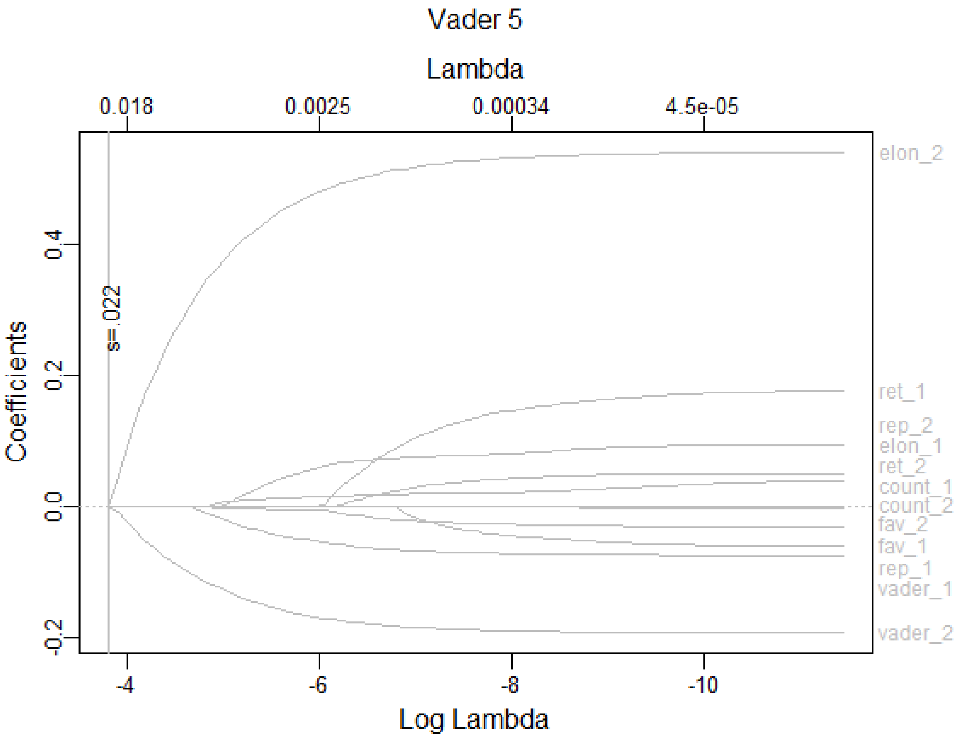

Figure A.5: Value of the coefficients depending on the value of log(10^{λ}) for the full data set on the five-minute interval.

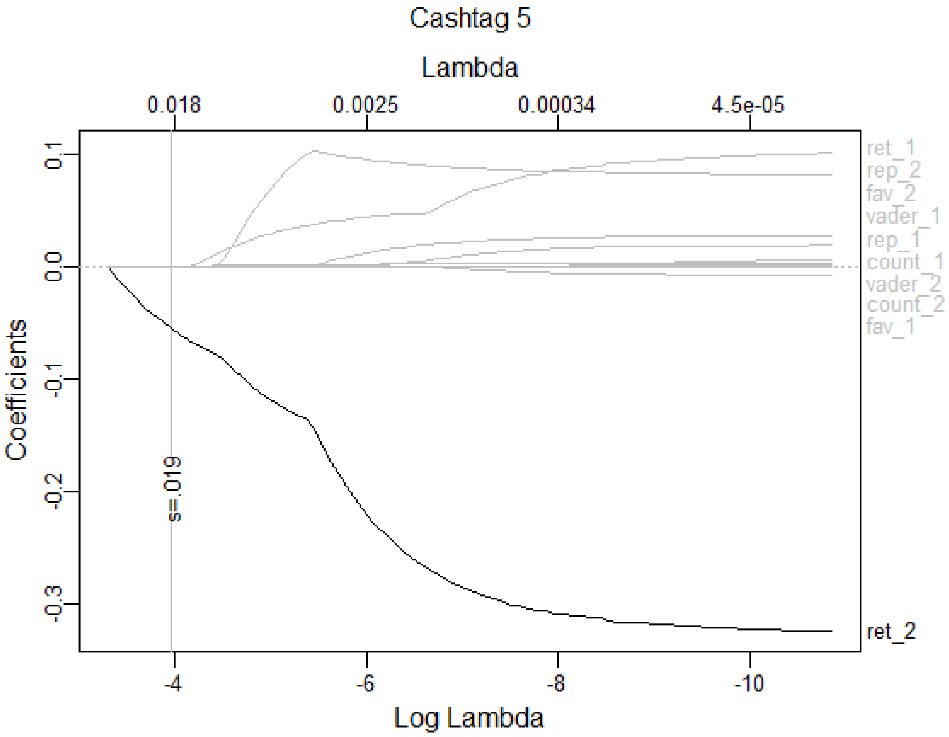

Figure A.6: Value of the coefficients depending on the value of log(10^{λ}) for the cashtag sample on the five-minute interval.

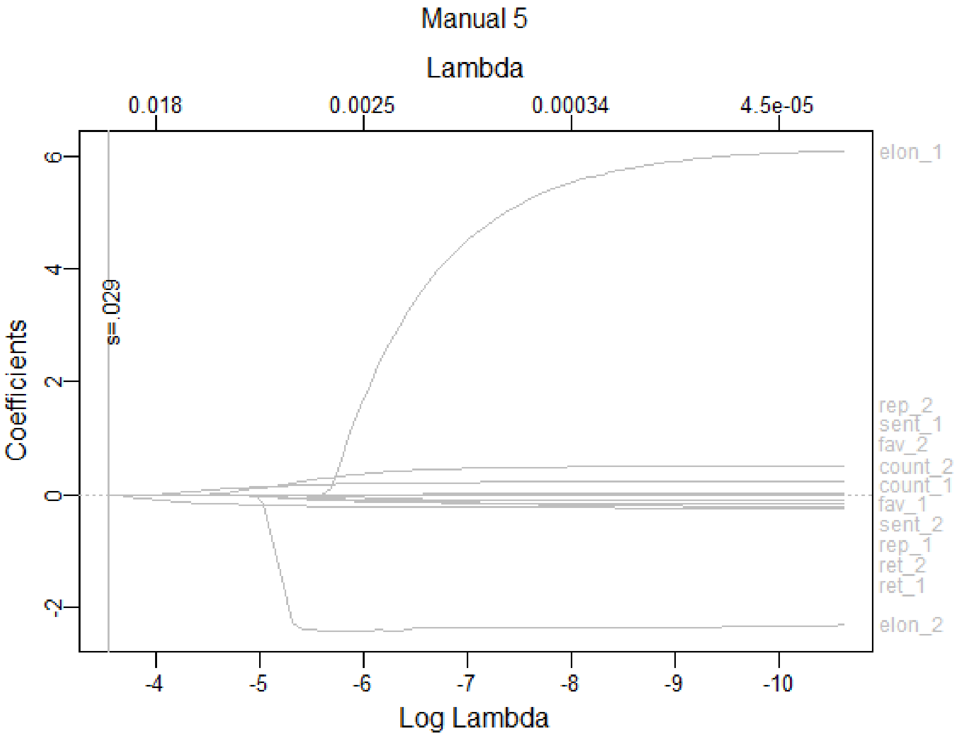

Figure A.7: Value of the coefficients depending on the value of log(10^{λ}) for the manually classified sample on the five-minute interval.

Appendix B Additional Tables

Table B.1: Results for the Lasso regression on the one-minute interval for the data sets, with the full search query (VADER), the manually classified tweets (Manual) and the data set with only the cashtag (Cashtag).

Variable	**VADER** **(1)**	**Manual** **(2)**	**Cashtag** **(3)**
Intercept	0.003427	0.003427	0.003427
sent_1	0	0	0
sent_2	0	0	0
sent_3	0	0	0
sent_4	0	0	0
sent_5	0	0	0
sent_6	0	0	0
sent_7	0	0	0
sent_8	0	0	0
sent_9	0	0	0
sent_10	0	0	0
fav_1	0	0	0
fav_2	0	0	0
fav_3	0	0	0
fav_4	0	0	0
fav_5	0	0	0
fav_6	0	0	0
fav_7	0	0	0
fav_8	0	0	0
fav_9	0	0	0
fav_10	0	0	0
rep_1	0	0	0
rep_2	0	0	0
rep_3	0	0	0
rep_4	0	0	0
rep_5	0	0	0
rep_6	0	0	0
rep_7	0	0	0
rep_8	0	0	0
rep_9	0	0	0
rep_10	0	0	0
ret_1	0	0	0

Table B.1: Results for the Lasso regression on the one-minute interval for the data sets, with the full search query (VADER), the manually classified tweets (Manual) and the data set with only the cashtag (Cashtag).

Variable	**VADER** **(1)**	**Manual** **(2)**	**Cashtag** **(3)**
ret_2	0	0	0
ret_3	0	0	0
ret_4	0	0	0
ret_5	0	0	0
ret_6	0	0	0
ret_7	0	0	0
ret_8	0	0	0
ret_9	0	0	0
ret_10	0	0	0
elon_1	0	0	-
elon_2	0	0	-
elon_3	0	0	-
elon_4	0	0	-
elon_5	0	0	-
elon_6	0	0	-
elon_7	0	0	-
elon_8	0	0	-
elon_9	0	0	-
elon_10	0	0	-
count_1	0	0	0
count_2	0	0	0.001367
count_3	0	0	0
count_4	0	0	0
count_5	0	0	0
count_6	0	0	0
count_7	0	0	0
count_8	0	0	0
count_9	0	0	0
count_10	0	0	0

Table B.2: Results for the Lasso regression on the five-minute interval for the data sets, with the full search query (VADER), the manually classified tweets (Manual) and the dataset with only the cashtag (Cashtag).

Variable	**VADER** **(1)**	**Manual** **(2)**	**Cashtag** **(3)**
Intercept	0.025289	0.025289	0.026778
sent_1	0	0	0
sent_2	0	0	0
fav_1	0	0	0
fav_2	0	0	0
rep_1	0	0	0
rep_2	0	0	0
ret_1	0	0	0
ret_2	0	0	-0.055386
elon_1	0	0	-
elon_2	0	0	-
count_1	0	0	0
count_2	0	0	0