A Big Data Research Paper

On Donald J. Trump's Twitter Activity

Copenhagen Business School

Figure 1: Trump Wordcloud

Søren Kolbye Jensen
Bsc. Ha(IT)
#43124133123
December 1, 2017

Contents

1 Abstract

The topic in this research paper is an analysis of Donald J. Trump's Twitter account, using datascience and datamining tools to reveal his behaviour. Problemformulation The research questions were: What was trump's behaviour on social media, prior to his announcement of his presidency. And how did it change, during and after his political campaign. How does Trump's behaviour affect the stock prices of companies ? What does using datascience to analyze Donald Trump?s behaviour mean from an ethical standpoint? The research questions were answered using datasets from Trump's twitter profile, as well as nasdaq to get stock prices of companies he tweeted about.

2 Introduction

Big Data has changed how data is viewed in society, businesses have invested heavily in infrastructure that can imrpove data collection. It is now possible, to use Big Data to analyze the behaviour of Social Media channels such as Twitter, Facebook and Instagram. However, the vast amount of data available with modern technologies has proven difficult to analyze. According to ??, Big Data can be either structured or unstructured, where the determining parameters are "Velocity, Verocity and Validity". In order to surmount this difficulty, new datamining technologies has been invented, in order to aid the field of Data Science.

These principles are called "Datascience and Datamining". According to (INSERT BOOK), datascience and datamining can be defined as follows:

Data science is the act of using fundamental principles, to guide extraction of knowledge from datasets.

Datamining on the other hand, deals with extracting knowledge from data using technolgoies, which incorporates the principles of data science.

This research paper, aims to use the fundamental principles of Data science, in conjunction with datamining technologies. To analyze the behvaiour of U.S president Donald J. Trump. and how this behaviour affected his political campaign, as well as how his behaviour affected the American financial markets. The motivation behind this research, is to see how data can reveal a top politians behaviour and thus also discover the actual benefits from Data analysis.

2.1 Case introduction

Donald John Trump, born in 1946 in Queens New York, is the CEO of the Trump Organization and is well known as a real-estate broker in the United States. However, on june 16th, 2015, Trump officially announced his candidacy as the 20th US President". This political campaign was heavily involved with Social Media, as Trump was actively fighting big news channels, such as CNN, NBC etc. labelling them as "fake news". Thus, Trump figured out a way to distribute highly controversial statements and gain political support by his use of Social Media Channels such as Twitter.

I therefore believe that it is extremely relevant to utilize datascience and datamining in order to discover what impact Trump has been able to make on his external environment, by essentially using "distributed data" on the Social Media Network: Twitter.

3 Problem Formulation and Reserach questions

The goal of this research paper is to analyze trumps behaviour on Twitter and how he used Twitter for political endeavours. A visual analysis on his behaviour between DATE and DATE will be conducted to reveal the patterns in his behaviour in this timeframe. This leads to the research question:

What was trump's behaviour on social media, prior to his announcement of his presidency. And how did it change, during and after his political campaign.

In addition to this, due to Trump's political position, it would be interesting to investigate the impact that his tweets has on the stockprice of companies, during his campaign he has posted both good and bad things about nurmerous companies. It is therefore relevant to investigate, if his Tweets have an impact and if that impact is short-term or long-term. This leads to the research question:

How does Trump's behaviour affect the stock prices of companies ?

Finally, it is relevant to evaluate and discuss what the above actually means to the context of Data Science, both from an ethical point of view, but also from a practical point of view. This leads to the final research question:

What does using datascience to analyze Donald Trump's behaviour mean from an ethical standpoint?

4 Theoretical Framework

In this section I will discuss my theoritcal framework which I used in order to investigate the research questions using data science. Some aspects of this chapter, may "transition" into the realm of methodology, but are none the less relevant to mention in both contexts.

4.1 Supervised datamining

Supervised and unsupervised datamining comes from machine learning, in a supervised method the data scientist will "supervise" the data and provide target information. In this research paper I have chosen to use supervised datamining, because I had a specific target. There are two "main" subclasses when it comes to using supervised data mining. One is classification and the other is regression - these two differ in terms of what the target is. In this research paper, the target is stock prices and thus it can be defined as a numeric target. The other main subclass of supervised datamining is classification which deals with binary targets, however this has not been used in the reserach paper.
I will use supervised datamining methods, because my target is specified and the data on said target exists. Furthermore, the historical data of the stock value of the examined companies are complete.

4.2 Linear regression and fitting

In this research paper, models containing linear regressions will be used, in order to discover the relationship among variables in the data model. This will help create a simple predictive model of the datasets

showcasing Trump's short-term impact on certain companies, when he tweets about them, in a negative or positive way. In these regressional models, the x-variable will be the date the tweet was posted and the y-variable will be the closing price of a given stock. Thus, resulting Lienar Functions with the format : y=mx+b, where m is the slope and b is where y intercepts. This will be done by attempting to fit a linear relationship between the dependent (Y) and independent (X) variables.

There are other regressional models that can be used with supervised datamining. The use of statistical methods can be used, in order to discover which model has the best fit to your dataset. Such as R-squared values, but in this research paper only linear regression will be used due to time constraints and my own lack of practical knowledge to complete a more statistically advanced model. However, the R-squared values will be discussed in the report, in order to examine the fit.

In order to illustrate the importance of the R-squared value, the term overfitting and underfitting datasets has to be defined.

Overfitting is when a datamining procedure is completely tailored to the training data, in a worst case it is because a model is "memorized". This has a cost in terms of achieving a model that can generalize in terms of unseen data points and it may result in poor model performance and harmful consequences when determining correlations in a dataset.
According to the authors of the books, overfitting is unavoidable to some extend. Therefore, there is not a specific data mining precdure that is "best" in terms of overfitting, nor does the authors argue that the answer is to produce a more simply model in order to produce less overfitting. There is a trade-off when making more complex models and overfitting, it depends on the situation and such a decision must be considered throughly by the data scientist, if a model is too simple it may not convey the actual complexities and thus will be less accurate than an advanced model with more overfitting. In terms of this report, I have decided that a simple linear regression is enough in order to get the bigger picture of Trump's short term impact on a companies stock, although I cannot deny that a more complex data model with more overfitting may have yielded better results.

* Underfitting The opposite of overfitting is if we have a model that is underfitting. This means that the model that was produced is not good enough to represent the fitted data. In terms of the R-squared value, the lower it is the more the model will be underfitting and the less useful the model will be in terms of representing the data.
https://www.pugetsystems.com/labs/hpc/Machine-Learning-and-Data-Science-Linear-Regression-Part-6-978/

5 Methodology

To answer the research questions two different types of analysis is presented

Furthermore, the CRISP framework has been used in a modified fashion, in order to better illustrate the context of this research paper.

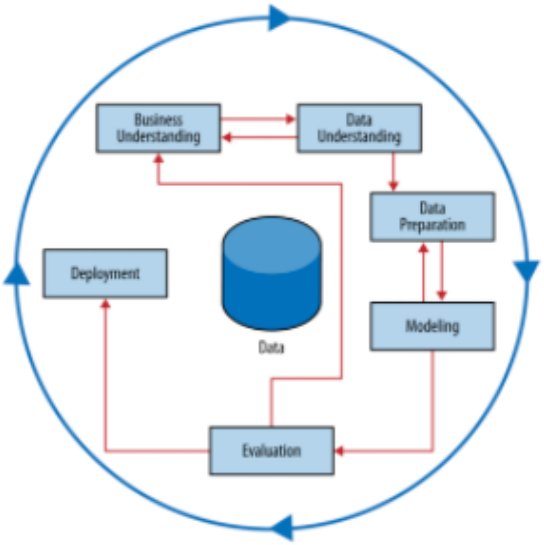

Figure 2: real, local caption for refrence

The model serves to illustrate how the data was processed throughout the proejct in order to yield a final result/answer to the research questions. The model is iterative and serves to explore the Twitter data, so a better understanding of the data can be reached. This process has been conciously iterated throughout the project.
”Business understanding” is in this context the overall problem formulation and research questions that we wish answered.
For every iteration a deeper understanding of the data is achieved, which affects the understanding of the overall research framework.

Data Understanding is how we collect the data from raw material and build a solution, it is therefore essential to critique your data and see what strengths and weaknesses that the data has. It can also be relevant to estimate the cost of the data, especially in a business context. However, for the purposes of this specific research paper, the economical aspect of data collection such as Cost/benefit won’t be explored. Due to the iterative nature of the CRISP framework, different solutions may change the direction of the analysis. The essential part to take away from this, is that we use data understanding in order to get a deeper understanding of our data and thus discover what the structure of the data is and how it can be used to ”solve” what in this research question is our business problem/research questions.
”Data preparation” is where we merge, cleanse and integrate multiple data sources. In this research paper, Alteryx has been used as shown in Table 1 later in the report. ”Modelling has used Tableau, which is a visual analytics tool. This makes it easier to explore and understand the patterns presented in the data. It has been a very closely fit process together with data understanding, in order to achieve the best results of ”good” data (In this case, usable data that can answer our research questions) and how to prepare it (So that frontend programs can be used in order to visualize findings)

Modelling is what we end up with, it is here our overall ”model” is made which captures patterns and regularities within the data. It is the stage where most data mining tehcniques are applied in order to create a model that is usable.
”Evaluation” Here we evaluate, whether the data model that has been built answers the overall research questions. After the model has been evaluated and validated, the model will go through ”deployment”, which in this sense will be the final conclusion to the research questions.

5.1 My overall process of handling data

I have developed the following model, to ilustrate how I have used Alteryx and Tableau in order to build a model. I have also attempted to illustrate how I used the CRISP model here, by determining if a given model is satisfactory. If not, I go back to the drawing board and reconsider what data I have and how I can prepare it in Alteryx, so that I can build a new model through the next iteration.

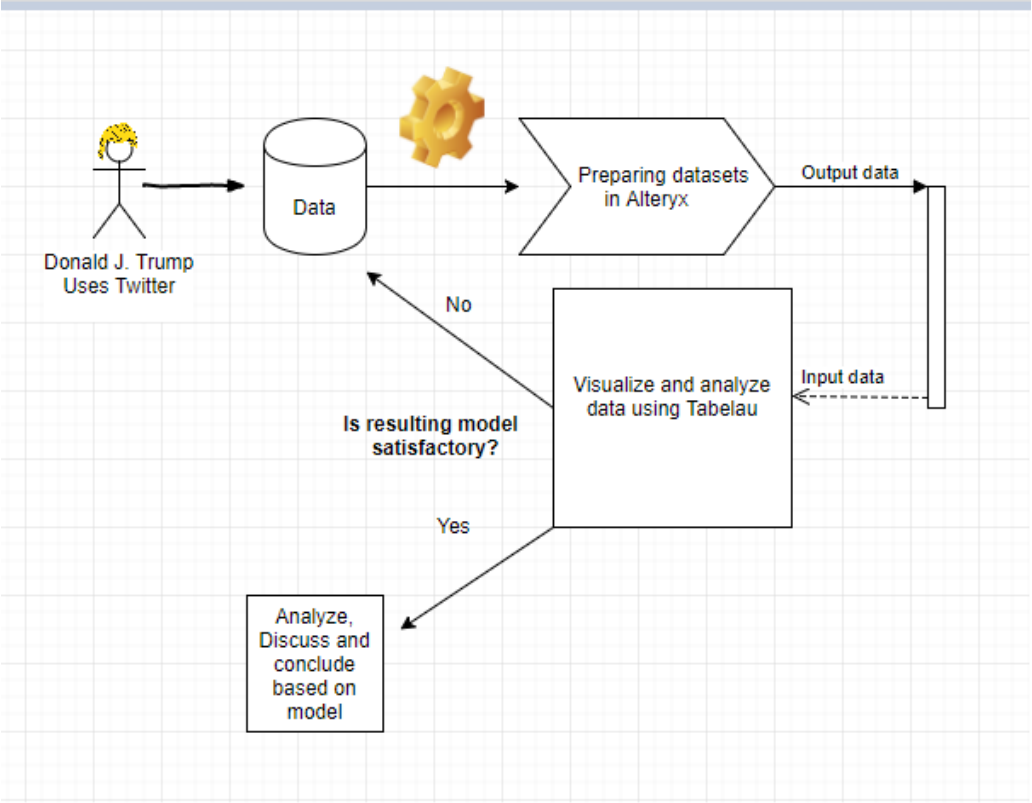

Figure 3: real, local caption for refrence

Throughout this project, I have attempted numerous models, most of which were discarded. Such as determining which fake news trump tweets most about, because I did not feel it answered the overall research question further.

To summarize, Table 1 shows the purpose and use of each datamining tool used in this research paper:

Table 1: My caption

Tool:	Purpose:	Use:
Alteryx	Data Analytics tool that provide data mining solutions by helping with data	Lorem ipsumLorem ipsumLorem ipsumLorem ipsum Lorem ipsumLorem ipsumLorem ipsum
Tableau	This tool helps with visual data representation.	Lorem ipsumLorem ipsumLore m iLorem ipsumpsumdasdLo remLorem ipsum ipsum

5.2 Data Acquisition and Dataset Description

The datasets in this research paper has been collected from Trump's Twitter page, ranging from 2009 to 2017.
The first datasets consists of Twitter data on Trump's primary twitter channel. This dataset will mainly be used to analyze Trump and his Campaign team's behaivour.
The second dataset consists of Historical data on the American and Mexican currency between Trump's announcement to 14/07/2017
The third dataset is collected through Sentione and will be used to analyze the sentiment towards Trump's campaign. Ranging from before his inauguration, until 02/11/2017.

6 Results

In this section of the research paper, I will go through the results of my datamining and discuss it in relation to the research question, and thus explain which insights I came across during my datamining, in order to create a meaningful discussion regarding Donald J. Trump's behaviour and impact on the stock prices of companies. The result section will be built around the CRISP framework, in order to better illustrate the process of my datamining.

6.1 Business understanding and Data Understanding

The understanding of both the overall research questions / Problem statement and the data was a very iterative process and thus is hard to document in a research paper, therefore my understanding will mostly be reflected in the following sections, where my ideas and reflections will be discussed. And hopefully create a reflection of how I arrived at the understanding of both my research questions and the data.

However, I would like to elobrate on the Data Understanding part, as this section includes evaluating the strengths and weaknesses of the data. In this research paper, Mostly data collected from Trump's twitter account has been collected, together with data on various stock prices on companies. As mentioned by Provost and Fawcett, there is rarely an exact match between the overall problem and the data, thus this has to be kept in mind when reading the next sections. Other factors that may have contributed to the

changes in stock prices from the analyzed companies is not within the scope of this research paper, nor is it in the datasets collected. Thus, this can be considered the weakness of the data. The strengths of the data however, is that there is enough variables within the dataset in order to make an analysis on Trump's behaviour, in addition, to be able to compare it to stock prices.

6.2 Data Preperation

In order to prepare the Datasets from Twitter and Sentione, I used a datamining tool called "Alteryx", which deals with the backend part of the data analysis. This tool was used to filter my data and join the relevant tables with each other, in order to make a comprehensive analysis. Data that was filtered out was things such as "Twitter ID" and "URL", as these were not needed for the direction I was going for.

One use of Alteryx was to do text-mining, which by nature is a very dirty and unstructured form of Big Data. While text-mining, it is important to keep in mind the context of the text (For example, what does Trump mean when he tweet "Crooked" a lot). Alteryx was used to find the frequency of the text, which shows how often a word showed up in the dataset.

Figure X shows an algorithm created in Alteryx that allows the creation of wordclouds. This Algorithim filters the dataset, so it only presents the text values. Furthermore, the algorithm removes empty data entries as well as other noise from the data. (e.g. ! : ;). Finally, the aglorithm seperates the sentences with new lines, so that every word can be counted by number of occurence, this data can then be put into the Tableau software in order to create a WordCloud.

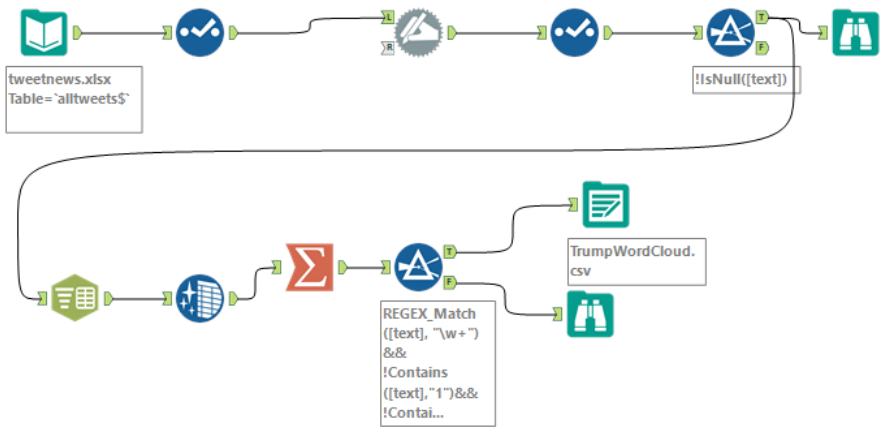

Figure 4: real, local caption for refrence

7 Modelling

This section contains the models that I have produced with Tableau, which is the product i used to make my visual data analytics. In this section I will use the model to answer the research questions.

In order to answer the first research question, regarding Trump's behaviour before and after he announced his presidency, I used textmining in order to build wordcloud, which shows which words he used frequently and thus can reveal something about his behaviour. The first model I want to show, is one that illustrates Trump's most used words before his announcement of candidacy.

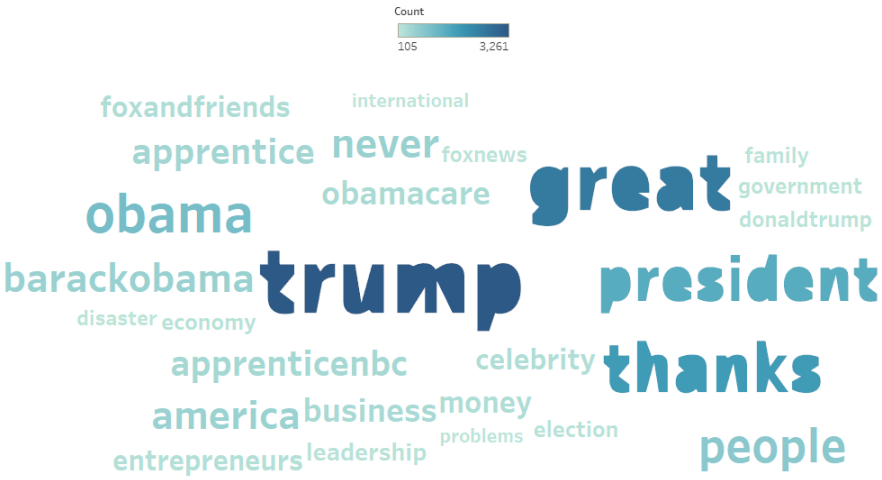

Figure 5: real, local caption for refrence

As Figure X shows, Trump was invested in the politics of USA before his announcement as president, having frequently used words such as "obamacare, president, government, barackobama". Furthermore, it can be seen that Trump's tweet weren't all about politics, but also about "business" and his TV-shows such as the apprentice (apprenticenbc). Therefore, it would be interesting to see a wordcloud of his behaviour after he announced his candidacy.

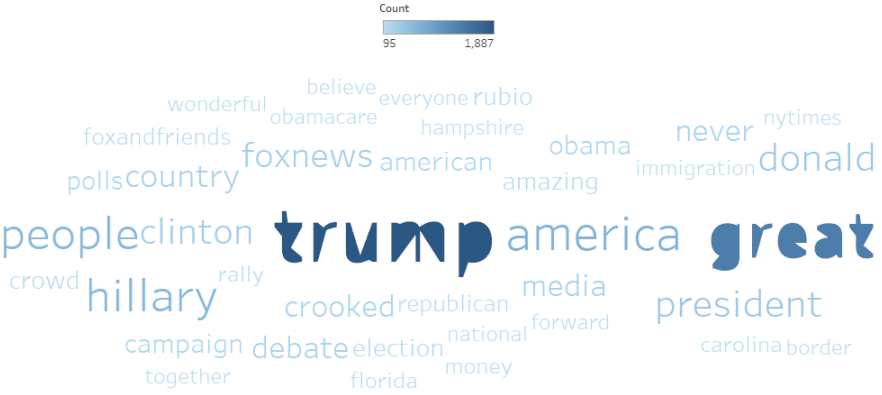

Figure 6: real, local caption for refrence

Here we see a similar, yet different story. Trump's tweets that were mostly unrelated to politics no longer showed up after filtering the word count, as they did in the previous model. Thus, we can see that Trump started focusing a lot more on politics and "fake news". Frequently using words such as "Hillay", "America", "media". We can also see, how Trump used twitter to further spread the nicknames of his political opponents, such as "crooked".

An interesting factor in this regard is to investigate the second research questions, so we can analyze how his behaviour on Twitter affected the "external environment", such as the stock prices of companies. I created a model which illustrates how a negative tweet from Donald J. Trump, may affect the stock price of LockHeed Martin aircraft company.

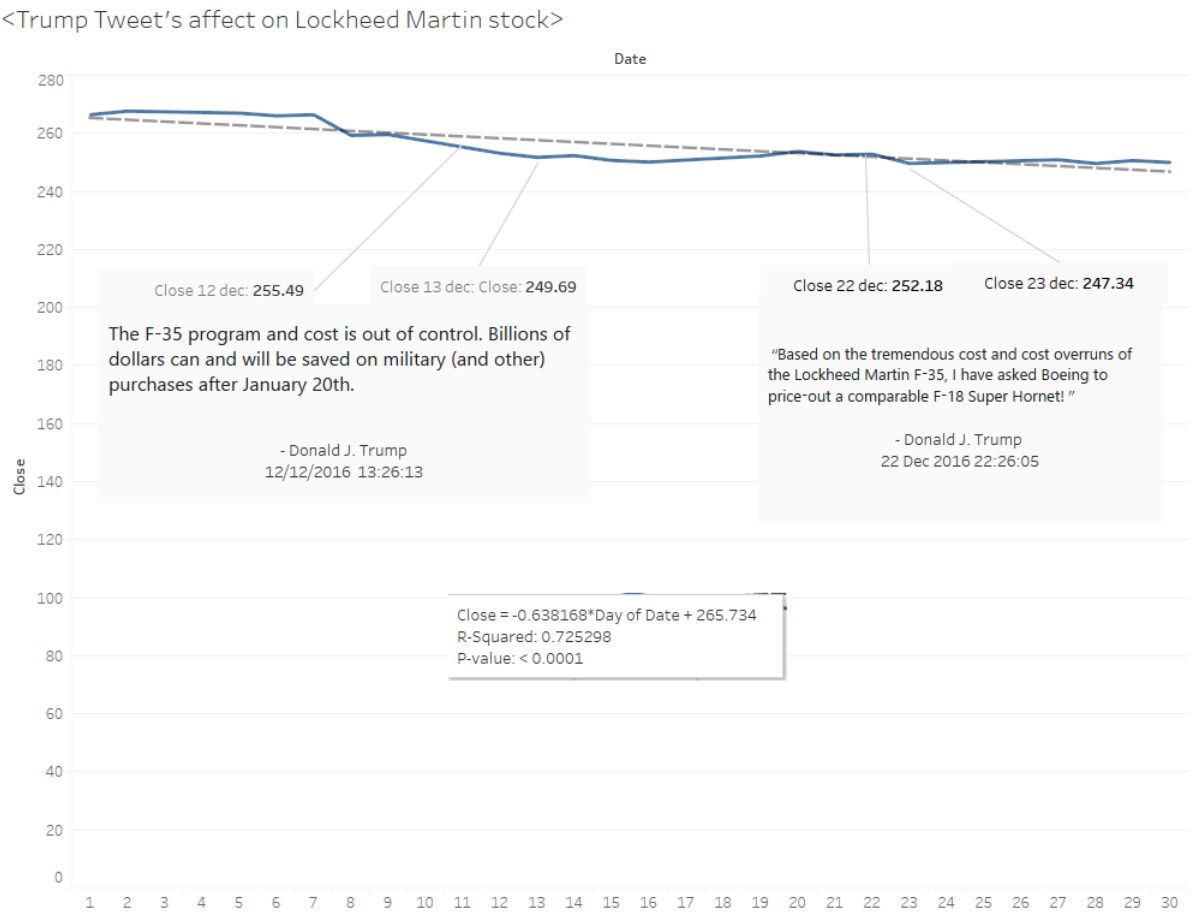

Figure 7: real, local caption for refrence

Figure X shows that the stock price before Trump posted a negative tweet, regarding his disatifaction with the F-35 program and how he would save money on military by cutting costs in the area that Lockheed Martin serviced the US.Army. As the graph shows, the closing price the day Trump made his tweet was 255.49 USD, but quickly fell to 249.69 USD the following day. Furthermore, later in the same month,

Trump tweeted again that he would change supplier to Boeing, causing the stock price to fall from 252.18 to 247.34.

If we go further from this standpoint, we can make a linear regression of the month of December, where Trump made these tweets. The linear trend line shows a negative slope of -0.6 USD in close price every day, which shows that Trump's tweets did indeed have a negative impact on the stock prices of Lockheed Martin. Furthermore, it can be said that the model has an R-squared value of 0.725, which indicates that the goodness-of-fit for the datamodel is decent. However, it should always be kept in mind as mention in the section "Theoretical Framework" that even a good R-squared value, may not always mean that you have a good model.

The above findings makes it interesting and relevant to expand our scope and look at the following month, to see if Trump's tweets' also affected the month of January 2017.

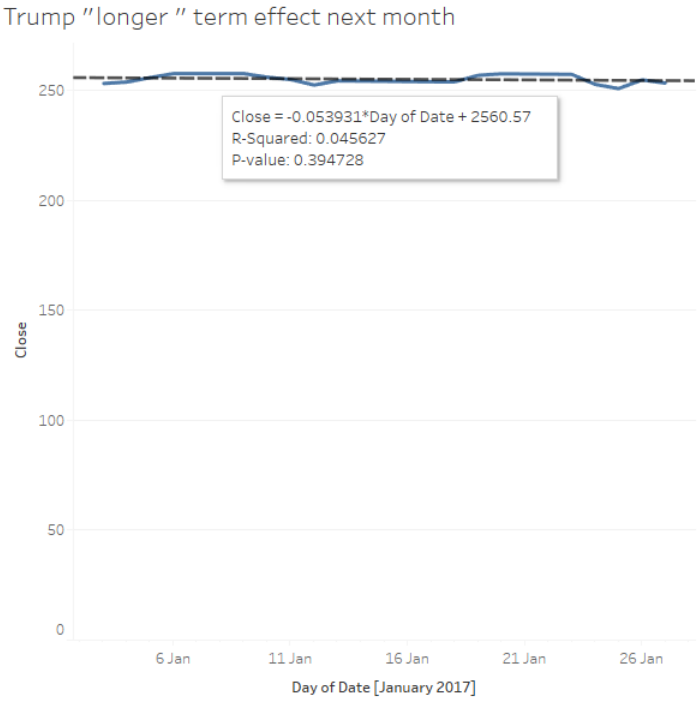

Figure 8: real, local caption for refrence

It becomes clear that Trump's tweets' only have an initial effect on the market, as the stock quickly stabilized itself in the following month. However, one most keep in mind that this model has a very bad fit for our data, with an R-squared value of > 0.1, which means that it doesn't fit our observations in the data very well.

8 To be added later

Furthermore, by the use of text-mining and using an algorithm, that sorts comments into differerent classifications ”positive”, ”negative”, ”neutral”. It is possible to determine, whether the reaction of Trump's presidency was overall positive or negative.

The results show that Trump's inital support at the announcement of his presidency was overall positive, with very few negative comments. The negative comments grew exponentially, as Trump got closed to winning his presidency, peaking around the time of his inagruation at xx/xx/xxxx

9 Discussion

In order to discuss the findings in this research paper in relation to an important topic of Data Science, I will discuss what ethical issues that datamining and datascience have on society. According to Fawcett, there is a correlation between good business decisions and the use of data to base them on. As an example, a lot of information about Trump's behaviour was datamined, revealing what topics he was interested in, what his impact is on his external environment etc. In Trump's case, he was mostly marketing his own ”brand”, but it is a concern, if such data was used on consumers without their consent and thus, raises questions about privacy. Companies will want to dig more and more into data about their consumers in order to reveal their behaviour and use it to increase the value in their business. This concern of privacy is a big issue, but is getting continuously regulated. As an example, a new personal data regulation in EU will help secure the privacy of consumers.

10 Conclusion

This research paper aimed to answer three different research questions:

What was trump's behaviour on social media, prior to his announcement of his presidency. And how did it change, during and after his political campaign.

For this research question, text-mining was used in order to reveal the frequency and analyze the context, in which Trump's tweets contained. It was found, that Trump was invested in politics before and after he announced his candidacy. Although, Trump was more invested in his TV-shows, such as the Apprentice before his announcement. After his announcement, it was found that Trump had a much bigger focus on political statements, naming political opponents and negative nicknames for them.

How does Trump's behaviour affect the stock prices of companies ?

This research question was answered using linear regression and data about a company in he aviation industry, called Lockheed Martin. It was here shown that Trump had a short-term effect on the stock price of the company, when he tweeted negatively about it. The findings showed that the following month,

the stockprices stabilized compared to the month of the tweets'.

What does using datascience to analyze Donald Trump's behaviour mean from an ethical standpoint?

From an ethical standpoint, it was concluded that the possibility of using datamining and datascience to reveal the behvaiour and impact that an individual can have on social media, is a threat to the privacy in society. It was concluded that such areas must be regulated in order to protect the rights of privacy. Using regulations such as the personal data regulation in EU.