A Big Data Research Paper

On Donald J. Trump's Twitter Activity

Copenhagen Business School

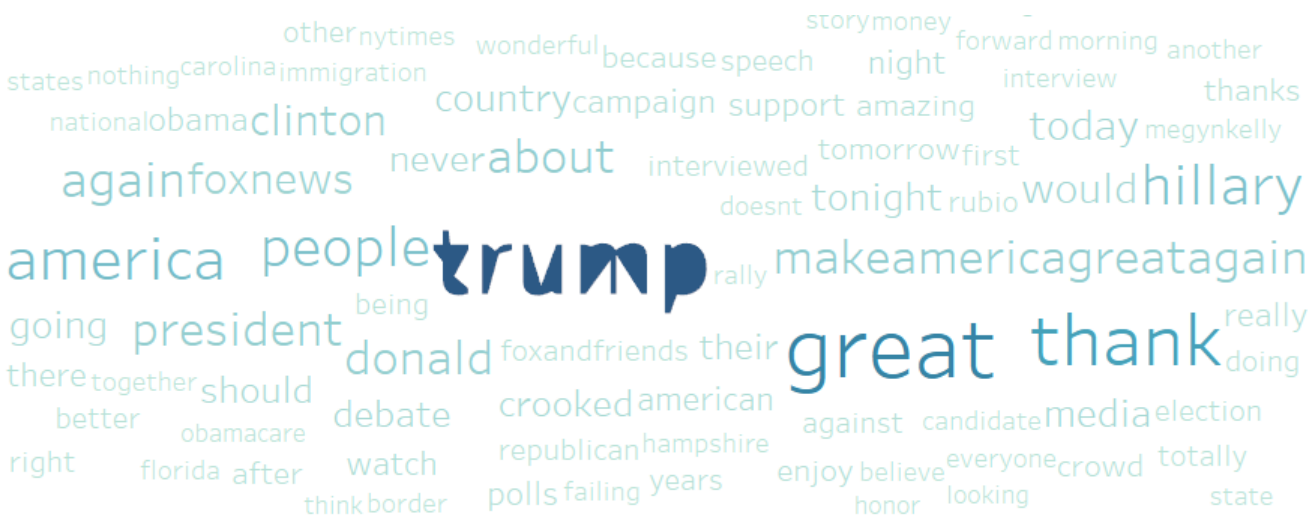

Figure 1: Trump Wordcloud

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. In addition to this, research what his impact was on the stock price of Lockheed Martin, which he tweeted negatively about. In addition to this, the ethical standpoint of data science was discussed in relation to the above.
This produced the following research questions in this research paper:

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 concepts used in this research paper were supervised datamining, where a linear regression model was made in order to analyze the overall impact of a negative tweet from Trump. In addition to this, text mining was used in order to generate a "bag of words", which could reveal patterns in the frequency of the words Trump used before and after his announcement of candidacy. The main analytics tools that were used to accomplish this, was Alteryx, which was used to filter, merge and preprare the datasets. In addition to this, a visual analytics tool called Tableau was used, in order to generate the models and illustrate the findings of datamining.

The most important results of this research paper, was that Trump shifted from tweeting about business and politics, to focusing more on politics after his announcement. Revealing a behavioral pattern, which suggested that his motivation of using social media is self-promotion. In addition to this, the linear regression revealed that a tweet from his Twitter account, can negatively impact the price of Lockheed Martin stock. Finally, from an ethical point of view, it was found that the possibility of analyzing behaviour on social media, can raise serious privacy concerns, which thankfully was found to be a field, which is getting increasingly regulated in especially the EU.

2 Introduction

Big Data has changed how data is viewed in society, businesses have invested heavily in infrastructure that can improve data collection. This new infrastructure makes it possible to use big data to analyze the behaviour of individuals on social media channels, such as Twitter, Facebook and Instagram. (Provost & Fawcett 2013) However, the vast amount of data available with modern technologies has proven difficult to analyze. According to Andrew McAfee and Erik Brynjolfsson, big data can be either structured or unstructured. The determining parameters are "Volume, Velocity, and Variety". (McAfee & Brynjolfsson 2012) In order to surmount this difficulty, the emerging field of data analytics is implementing what is called "Data Science" and "Datamining".

According to Foster Provost and Tom Fawcett , datascience and datamining are related, but not the same. Data science is the act of using fundamental principles, to guide extraction of knowledge from datasets and Datamining deals with extracting knowledge from data using technologies, which incorporates the principles of data science. (Provost & Fawcett 2013)

This research paper aims to use some of the fundamental principles of data science in conjunction with datamining technologies. This will be done in order to analyze the behavior of U.S president Donald J. Trump. and how this behaviour changed during his political campaign, as well as how his behaviour affected the U.S. financial markets. Therefore, the motivation behind this research, is to see how data can reveal a top politicians behaviour and thus also discover the actual benefits of data analysis.

2.1 Case introduction

Donald John Trump, born in 1946 in Queens. New York, is the president of the united states and the CEO of the Trump Organization. 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, BBC 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. Trump himself describes social media as a "key role" in winning the presidency on November 9th, it is therefore important to investigate how data science can help analyze his behaviour and impact on such social media channels.

3 Problem Formulation and Research Questions

From the case introduction a couple of research questions have been developed. Therefore, to summarize: The goal of this research paper is to analyze trumps behaviour on Twitter and how he used Twitter for political endeavors. A visual analysis on his behaviour between 2009 and 2017

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 after his announcement?.

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

How can a negative tweet from Donald Trump 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, or any other individuals behaviour, mean from an ethical standpoint?

3.1 Delimitations

Due to this not being a statistics course, in addition to the fact that statistics is not being taught on ha(it.), I will not dwell too much into the deeper statistical details of my datamodels. Furthermore, concepts such as R-squared, and other metrics are only barely mentioned in the course litterature. In spite of this, I will attempt to give light explanations of these metrics using other sources, because I believe it is one of the fundamental skills that a data scientist should posses.

4 Conceptual Framework

In this section I will discuss my conceptual framework which I used in order to investigate the research questions using data science.

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 there is a specific target (e.g. The predicted stock price of a company). 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 research 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 company is 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 lockheed-martin, when he tweets about them in a negative way. In these regression 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 linear functions will have the format : y=mx+b, where m is the slope and b is where the regression line intercepts the y-axis. This will be done by attempting to fit a linear relationship between the dependent (Y) and independent (X) variables. (Provost & Fawcett 2013)

There are other regression models that can be used with supervised datamining and in order to determine which method has the best fit for your data, it is possible to use various statistical metrics. Such as r-squared. In order to illustrate the importance of r-squared and other similar metrics, the terms overfitting and under-fitting 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. (Provost & Fawcett 2013)

According to Provost and Fawcett, overfitting is unavoidable to some extend. Therefore, there is not a specific data mining procedure that is "best" in terms of overfitting, nor does the authors argue that the answer is to produce a simple 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. In spite of this, I cannot deny that a more complex data model with more overfitting may have yielded better results. But such model has not been built and tested in this research paper.(Provost & Fawcett 2013)

Underfitting is the opposite of overfitting. if a model has severe underfitting of the data, it 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 under-fitting and the less useful the model will be in terms of representing the data (with exceptions, as previously explained).

According to the authors, an analyst should always know why any of the metrics are included and consider if there might be better metrics available for the model. R-squared values can help with determining the fit of a model, this value should be below the p-value in order to indicate a good fit. Other metrics that are included in the model is squared-error, which
This research paper will also use a residual-plot, in order to investigate the model's goodness of fit. Even if the model indicates a good fit according to our R-squared value and p-values.

4.3 Text mining

Text is according to Provost and Fawcett, often considered as "unstructured" data, and it therefore does not have the ordinary structure of data. Text can have a variety of lengths and words depend heavily on context. In addition to this, it can be very dirty e.g. grammatically incorrect. In order to conduct this analysis, I used the concept of "Bag of Words", where the data was treated as a collection of individual words. It ignored things such as grammar, order, sentence structure and punctuation. This makes it possible to measure the "frequency" of the words, which makes it possible to differentiate between word that were used a couple of times, with words that are frequently used in order to reveal a pattern. It is also important to filter the data, a term that is very rare, should not be presented if it is not important. The same applies to words that are too common and holds little value. (Provost & Fawcett 2013)

5 Methodology

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

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

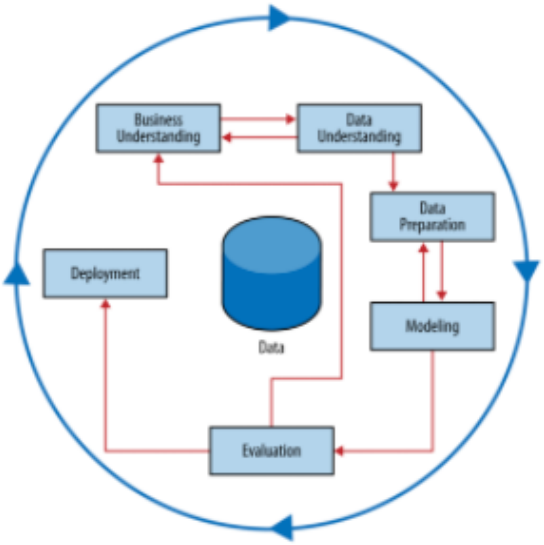

Figure 2: CRISP framework, (Provost & Fawcett 2013)

The model serves to illustrate how the data was processed throughout the project in order to yield a final result/answer to the research questions. The model is iterative and explores the Twitter data, so a better understanding of the data can be reached. "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 paper 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 in order to prepare our data, by merging and filtering, which I will elaborate on later in the paper.

"Modelling" was done with the help of the visual analytics tool "Tableau", 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 from "good" data. In this

case, usable data that can answer our research questions, by preparing it accordingly, 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 techniques are applied in order to create a model that is usable.

"Evaluation" is the stage where the data scientists tests whether the data model that has been built answers the overall research questions - and if the model has a good fit, so that it can be used without showing misleading correlations. 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 Process of Handling Data

I have developed the following model, to illustrate how I have used the analytics tools: 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.

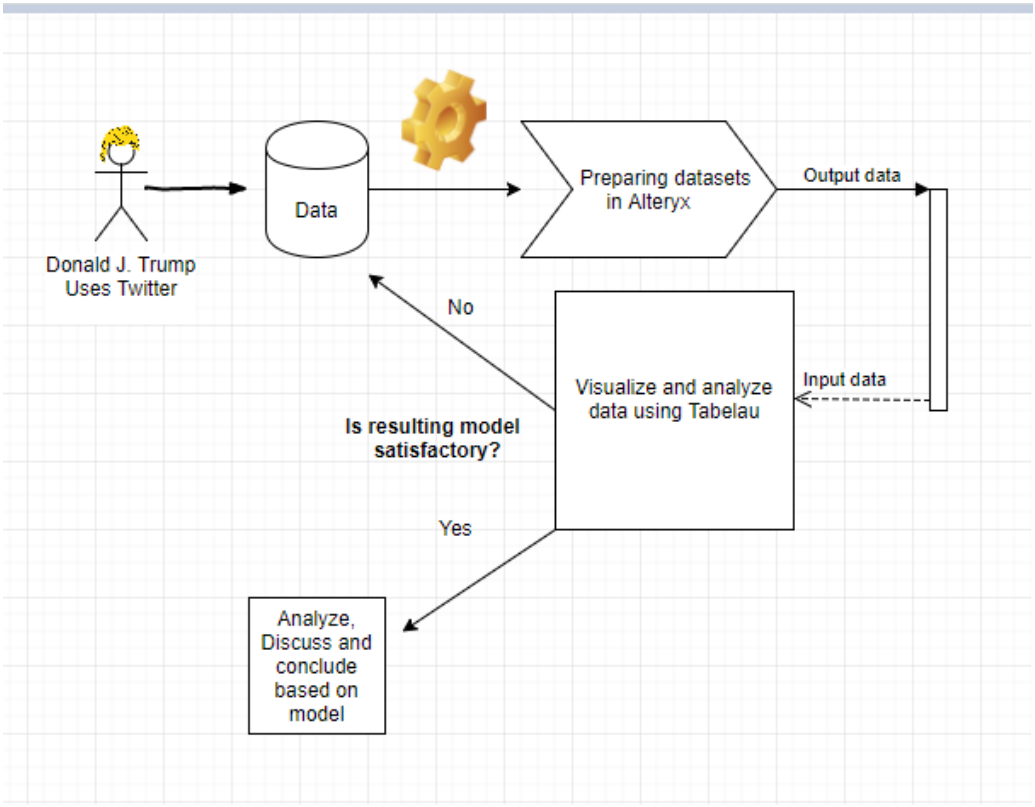

Figure 3: Model of my process of handling data

Throughout this project, I have attempted numerous models, most of which were discarded. Such as determining which news channels trump tweets most about, due to limiting the scope of the research paper.

5.2 Data Acquisition and Dataset Description

The first datasets consists of Twitter data on Trump's primary twitter channel. This dataset will mainly be used to analyze Trump and his behavior on Twitter. The strengths of this dataset comes from the fact that all the tweets made by Trump has been recorded, which can effectively be used to analyze his behaviour, as when he posted negative things about Lockheed-Martin. The dataset contains information about the date, what was in the tweet, how many retweeted and favorited the tweet, as well as the source of the tweet. (e.g. Android, Iphone). The one fundamental weakness of this dataset, is that it can be hard to determine if Trump was really the one, who made all of the Tweets, and not people from his political campaign team.

The second dataset was collected from Nasdaq and is used to collect information about the stock

price of lockheed-martin, which is a company that Trump tweeted negatively about. The strength of this dataset is that the dataset contains complete historical data, which makes it possible to make a correlation between this data and what Trump tweeted about. This also means that this particular dataset does not have any significant, if any, weaknesses.

In terms of reliability and validity, both datasets are collected from trusted, unbiased sources. The Trump Dataset was retrieved directly from his Twitter page and thus have a very high reliability and validity.

6 Results

In this section of the research paper, the results and insights of the datamining will be presented. The results will make it possible to start a meaningful discussion regarding Donald J. Trump's behaviour and impact on the stock prices of companies. Furthermore, the result section will be built around the CRISP framework, in order to better illustrate the process of the datamining.

6.1 Business understanding and Data Understanding

The understanding of the research questions and the data was an iterative process and is therefore hard to document in a research paper. As a result of that, the data understanding will be a part of the remaining sections ,where I will document my reflections and understanding of the research questions and data.

However, it is important to elaborate 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 Preparation

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

the research questions.

One use of Alteryx was to do text-mining, as explained in the conceptual framework, 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 4 shows an algorithm created in Alteryx that allows the creation of "word-clouds". This Algorithm filters the dataset, so it only presents the text values, generating our "bag of words". Furthermore, the algorithm removes empty data entries as well as other noise from the data. (e.g. ! : ;). Finally, the algorithm separates the sentences into new lines, so that every word can be counted by number of occurrence, resulting in the frequency of the words. Finally, the data can be put into the Tableau software in order to create a WordCloud.

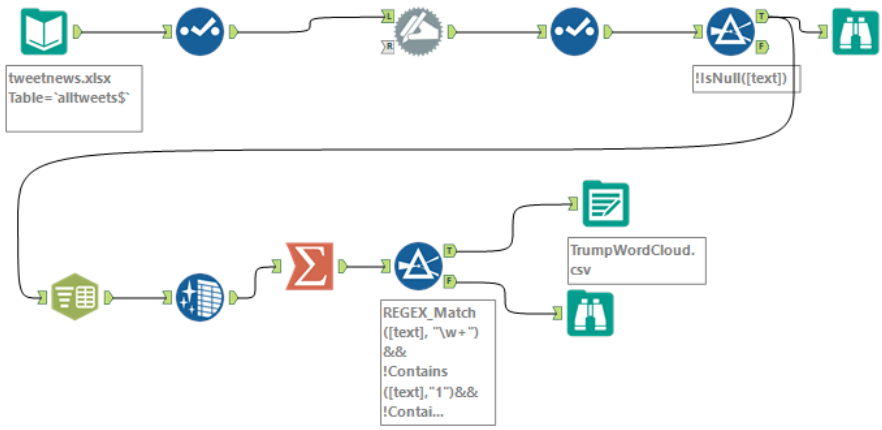

Figure 4: Alteryx algorithm used to produce a word-cloud

6.3 Modelling, Evaluation and Deployment

This section contains the models that I have produced with Tableau, the models will be used 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 text-mining. The result of the text-mining was a word-cloud, which showed which words he used frequently and thus revealed something about his behaviour. The first model illustrates Trump's most used words before his announcement of candidacy.

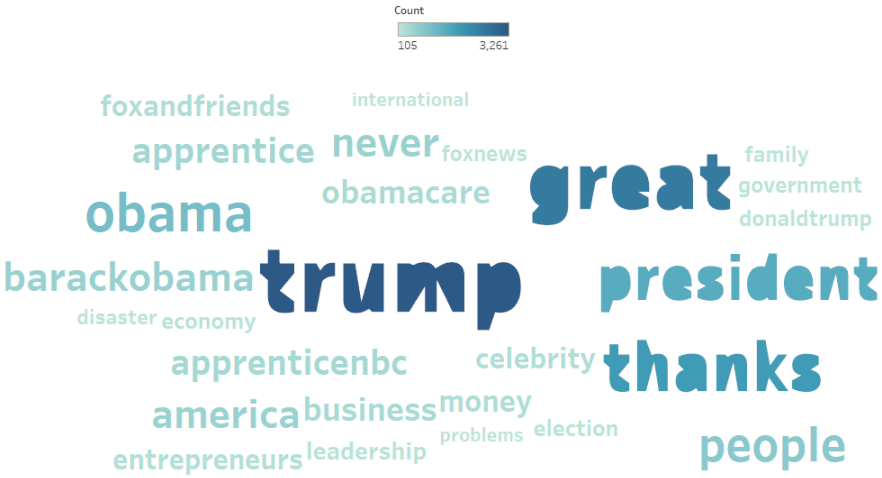

Figure 5: Donald Trump's frequently used words, before announcement of candidacy

As Figure 5 shows, Trump was invested in politics before his announcement as president, his frequently used words include "obamacare, president, government, barackobama". Furthermore, 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 word-cloud of his behaviour after he announced his candidacy.

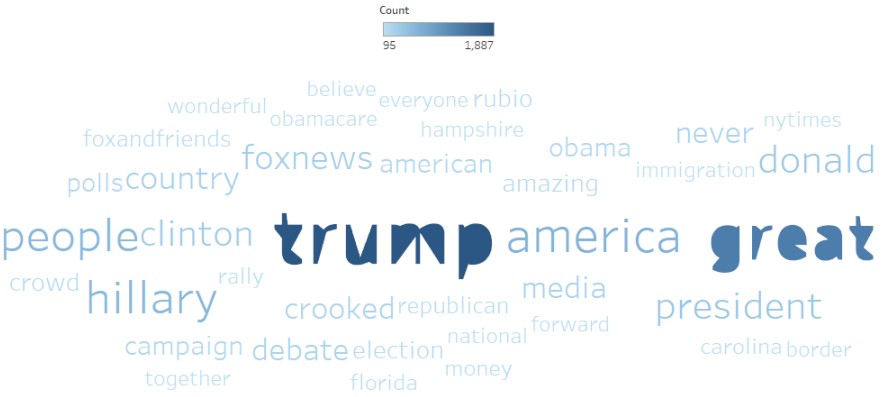

Figure 6: real, local caption for refrence

Here we see a similar, yet different story. Trump's tweets that were a mixed bag between business and politics no longer showed words relating to business and his tv-series, as the previous model did. Thus, we can see that Trump started focusing a lot more on politics and "fake news". Frequently using words such as "Hillary", "America", "media". We can also see, how Trump used twitter to further spread the nicknames of his political opponents, such as "crooked". Which in

this context was a nickname he gave Hillary Clinton during his political campaign.

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, affect the stock price of Lockheed Martin, which is a large aerospace company. The model on figure 7, has the days in december 2016 on the x-axis and the close-price of the lockheed martin stock on the y-axis.

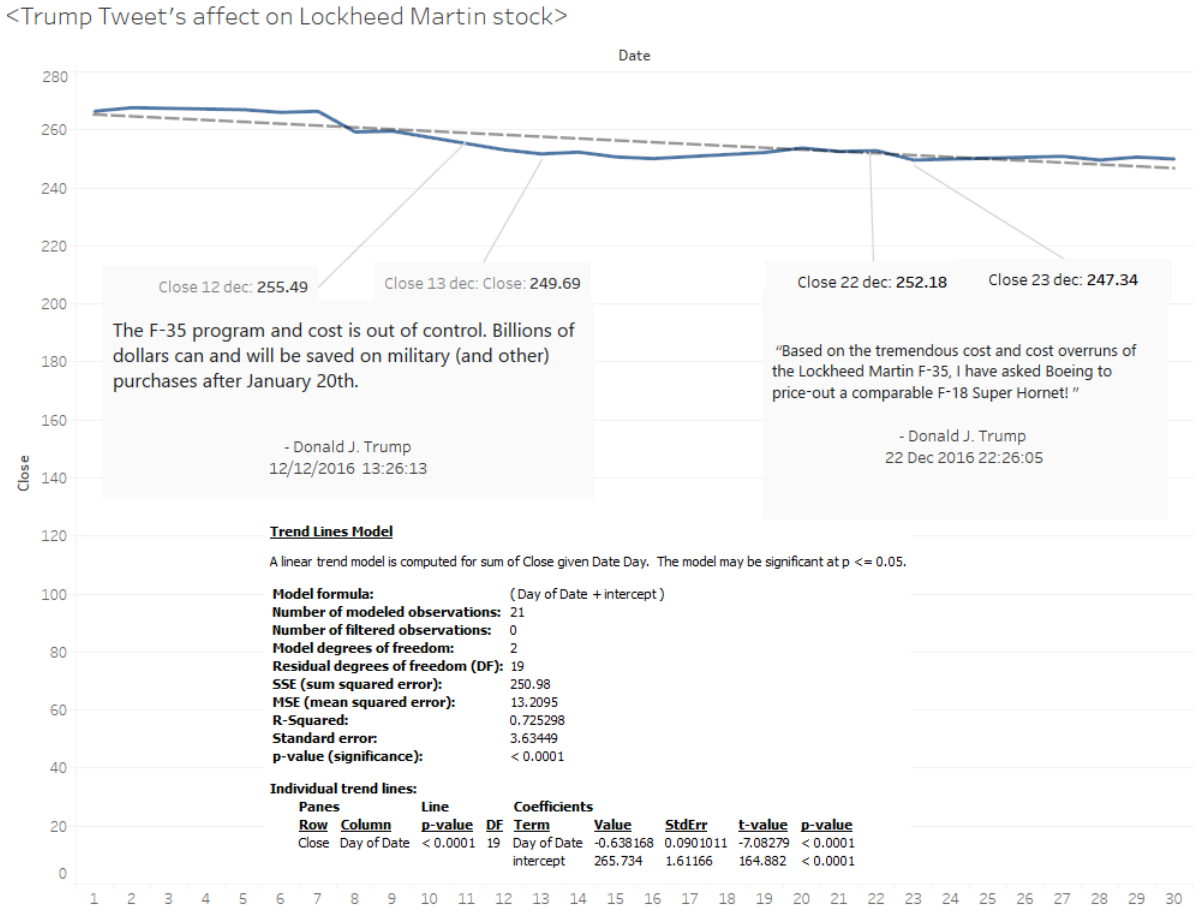

Figure 7: Trump tweet's affect on Lockheed Martin stock price

Figure 7 shows the stock price of the company, before Trump posted a negative tweet and the after effects of the tweet. The tweets were regarding his dissatisfaction with the F-35 program and how he would save money on military by cutting costs. This is the area in which Lockheed Martin services the U.S. army and thus likely has an effect on the company's stock price.

As the graph shows the closing price on december 12th, which is the day Trump made his first tweet, was 255.49 USD. This price quickly fell to 249.69 USD the following day. Furthermore, later in the same month, Trump tweeted that he would change supplier to Boeing, causing the stock price to fall from 252.18 to 247.34.

Going further from this standpoint, a linear regression was made of the stock price in the month of December. 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, the model has an R-squared value of 0.725, which indicates that the goodness-of-fit for the datamodel is more or less decent. However, it should always be kept in mind as mention in the section "Conceptual Framework" that even a good R-squared value, may not always mean that you have a good model. Therefore, I also factored in my own rational conclusions from seeing the overall trend in the datapoints.

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.

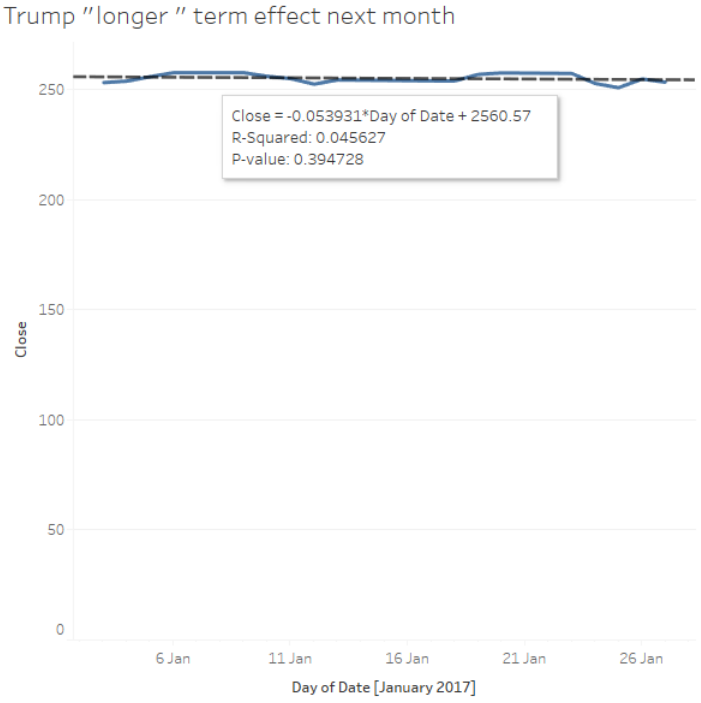

Figure 8: real, local caption for refrence

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. In spite of this, it is easy to see that the stock price in January is not as volatile, as it is in December.

7 Discussion

I will now discuss the results and reflect upon them.

7.1 Answer to 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.
The two word-cloud produced actionable insights into Donald Trump's behaviour on social media before and after his announcement of presidency. It is clear that he was more focused on his life as a businessman, before his announcement. He tweeted a lot about various tv-series, which he was involved in, as well as more business-oriented and economical topics. From this point of view it is very clear that Trump is a very career-driven person, who uses social media heavily to promote whatever endeavor he is involved, or interested in. After he announced his run for presidency, the results showed that he shifted to a political focus as expected. Looking at the results at the second word-cloud, it is very clear that he ran his political campaign in an unusual manner. Many of his most frequently used words, were found to be negative nicknames for his opponents. A prime example of this, as was seen in the results section, is how Trump tweeted "crooked" a lot, aimed towards his main political opponent, Hillary Clinton.

I believe that Trump found out about the strength of social media, in terms of using it as a base for distributing news to a large amount of people. The core of his behaviour before and after his announcement did not change, in both cases he forwarded his own self interest. What shifted was the target of his behaviour and it is thus clear that the use of data in the modern age, is very much capable of analyzing patterns in a persons behaviour. This raises numerous concerns, which I will touch upon later in this discussion.

How does Trump's behaviour affect the stock prices of companies ?
In addition to this, the second research question was answered. The resulting model of this research question, illustrated that something as small as a few megabytes of data in "the cloud", is enough to make an impact on the real world. This is an obvious fact to most, when the one tweeting is the president of the united states and that tweet contains information that make a large dent in a company's cash-flow. Nevertheless, it is an interesting aspect of data science, to be able to find such correlations between two different datasets.

What does using datascience to analyze Donald Trump's behaviour mean from an ethical standpoint?
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 and Provost, 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 data-mined, 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. http://ec.europa.eu/justice/data-protection/

7.2 Learning reflection

This research paper has helped me learn about the thought-process of a datascientist and how iterative the process is, as illustrated by the CRISP model. Unfortunately, I was unable to go as in-depth with the analysis as I wanted to, due to my limited knowledge of statistics and machine learning. In spite of all this, I do believe I arrived at some insights which allowed me to answer the research questions. In addition to this, the course and this research paper has taught me a lot about different analytics tools, which has given me a broad perspective of where I want to develop my skills in the future. Which for my case was in the realm of machine learning.

8 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 stock-prices 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 behavior 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.

References

McAfee, A. & Brynjolfsson, E. (2012), 'Big data:the management revolution'. [Last Accessed: December 1st 2017].
URL: *http://tarjomefa.com/wp-content/uploads/2017/04/6539-English-TarjomeFa-1.pdf*

Provost, F. & Fawcett, T. (2013), *Data Science for Business: What You Need To Know About Data Mining And Data-Analytic Thinking*, O'Reilly Media, Inc.

A

Stests