University of Copenhagen

Social Data Science Exam

Predicting Stock Prices from the Tweets of Donald Trump

Jeanette Køster
Elisabeth Bast Laurents
Joseph Massey
Marcus Orloff Søemod

01.09.2018

Jeanette (exam ID 73) contributed to the following parts: 1.1.1, 1.2.1, 1.3.2, 1.3.3, 1.3.4, 2.1, 2.2, 3.1, 3.2, 3.3, 4.1

Elisabeth (exam ID 121) contributed to the following parts: Introduction, 1.1.0, 1.2.1, 1.3.0, 1.3.1, 1.3.5, 3.7, 4.1

Joseph (exam ID 153) contributed to the following parts: Literature Review, Research Question, 1.2.1, 1.2.2, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 4.3

Marcus (exam ID 182) contributed to the following parts: 1.1.2, 1.2.3, 2.2, 3.6, 4.2

The conclusion has been written in collaboration.

Contents

Introduction

During his presidency and election campaign, American President Donald Trump has been a very diligent user of the social media platform Twitter. He has been criticised for his direct, impetuous, all-caps style of tweeting as well as his general behaviour that has been called narcissistic, delusional, and paranoid (Bulman, M., in the Independent, 2017). The mood of Donald Trump at any given time is a subject of debate (Coppins, M. in the Atlantic, 2018). But does it really matter? This project aims to answer whether Donald Trump's sentiment can be used to predict the changes at the American stock market and find out if the market is affected by Donald Trump's mood at a given time. Using sentiment analysis of the best publicly available source of Trump's mood - his tweets - and the reactions to it compared with stock market data from the Dow Jones Industrial Average (Dow Jones) and the National Association of Securities Dealers Automated Quotations (NASDAQ), we will implement different machine learning models that can predict changes on the stock market based on Donald Trump's sentiment. This report will look into whether there is a possibility that Trump's mood can be used to predict the stock market, in the same way that the general mood on Twitter has been used to predict the stock market (Bollen et al, 2011).

This report will begin by discussing relevant literature to form the foundations on which the report is built. Following this will be a comprehensive review of the methods used, including the cleaning, scraping and analysis of the data. This will be followed by descriptive statistics, a discussion of our results, and the report's limitations. Finally the project will conclude by discussing the significance of our report, and suggesting potential direction of future scholarship on the topic.

Literature Review

Financial markets are markets in which people trade financial securities and derivatives such as futures and options at low transaction costs. As such these markets are affected by the usual laws of supply and demand, however there is a extra focus on external events which can cause changes in the markets. This could be anything from a company's profits being lower than expected, to a new product release, to a fall in interest rates in the country where a company is based.

Stock markets were thought to have been based on historical stock prices, however recent studies suggest this is false, and markets actually react to various exogenous events. This is known as the Efficient Market Hypothesis, which states that 'financial market movements depend on news, current events and product releases and all these factors will have a significant impact on a company's stock value' (Pagolu et al., 2016). As these factors are somewhat random, predicting stock markets is hugely difficult, and cannot be predicted with more than 50 pct. accuracy (ibid).

The invention and expansion of the internet has augmented the volume and speed of transmission of information around the world. The presence of the internet allowed markets to attain information even faster, and as early as 2004, Antweiler and Frank were investigating how online message boards were affecting the markets (Serban et al., Undated). Social media, which has exploded in popularity in the last 10 years, has become an integral part of society - sites like Facebook, Instagram, and Twitter have millions of users everyday across the globe. The data created on these websites encapsulates the fabric of the lives which we live today - on these platforms you can understand how a group of people is thinking, gauge the mood of a nation, and understand more about the customers of your business.

This report will focus solely on Twitter. Twitter was created in 2006 and is a unique tool for researchers looking to capture the mood of a certain group of people. Twitter allows users to send 280 character tweets (formerly 140), which according to Java, A. et al. (2007) contain "daily chatter, conversations, URLs (sharing of links), and news". The short length, volume of tweets (around 500 million per day (Internet Live Stats, 2018)) and their content, described above, makes them perfect for analysing the public mood. Twitter has been used for multiple research projects on public opinion analysis (Pagolu et al., 2016), Asur and Huberman (2010) used public sentiment for movies on Twitter to predict box office collections for a movie prior to its release, and Eiji et al. (2012) forecasted flu outbreaks using twitter data.

Predicting stock prices, as mentioned above, is notoriously difficult and so researchers are constantly considering how new sources of information affect the markets. For example, Bollen, J. et al. (2010) used public mood on twitter to predict the daily up and down closing values of the

Dow Jones Industrial Average with 87.6 pct. accuracy. Mittal (2011) replicated the project and achieved 76 pct. accuracy. Further studies include those by: Rao and Sviristka (2012), Zhang et al. (2011), and Gilbert and Karahalios (2010). These studies point towards the importance of Twitter as an information source, and confirm the worth of Twitter as a gauge of public opinion.

Twitter also can help to predict other important parts of human life, with one example being politics. Jin et al. (2012) created a model which could provide up to date public feeling on presidential candidates through sentiment analysis on twitter. This project is interested in the intersection of stock markets, politics, and twitter sentiment, and our focus will fall on one account – @realDonaldTrump. There has been little study on Trump's twitter account before, with one example being Bae et al. (2012) considering the sentiments of social influencers. Trump's election campaign was characterised by his seemingly unconstrained twitter account, and he has continued to tweet freely since his election. This report will investigate the extent Trump's twitter account can influence the US financial markets by observing at the Dow Jones and NASDAQ index.

Research Question

To what extent are the sentiments of Donald Trump's tweets able to predict changes in US stock markets on any given day?

Our interest in writing this report was sparked by the plethora of pseudo-scholarship around the internet regarding Donald Trump's Twitter account. One example of this was an article detailing a drop of 1.4 pct. in the SP 500 as a result of a typical Trump rant on his Twitter account (Fortune, 2018). Whilst these articles have little statistically backing to them (they solely link a change in the markets to anything that has happened on Trump's account) we felt there was a gap in the social data scholarship which we could fill. Using these articles as inspiration, we sought to understand whether the sentiment of Donald Trump's tweets would have an effect on the markets. This base question led us to many other questions which this report will also seek to answer: would we then be able to predict future changes in the markets based on the sentiment of a single tweet? Can one single Twitter account truly have such a strong influence on the stock markets?

1. Method

1.1. Data collection

This section contains information on how we have collected the data for this project. We have exclusively used web based data, but our data is both quantitative (stock market prices) and qualitative (text based tweets). One of the main advantages of using web based data is that it is very easy to access, as it doesn't require any special permissions to use and doesn't require any sort of interviewing or other ways to usually access data. Therefore, it is ideal for at project written within a limited time frame, as the data can be accessed quickly.

Often, one has to be careful with using social data for data science. An important aspect in social data science is the privacy of the people included in the data used. However, in this project we only use data from publicly available sources such as stock market indexes and Trump's very public Twitter account with more than 54 million followers. Thus, none of the data we are using can be classified as private so our project has very limited issues concerning privacy.

1.1.1. Scraping Twitter Archive

To make an analyse of Donald Trumps tweets, we had to find a way to select all of his tweets. There are two ways to do this: use the Twitter API, or scrape the already existing twitter archive: `http://www.trumptwitterarchive.com/archive`

Since you only get 200 tweets per call while using the Twitter API, we decided to use the already existing archive. For each year a different URL request was used, which is why we defined a function `("tweets_by_year")`, the function contains both scraping but also structuring of dates.

1.1.2. Using the Alpha Vantage API for stock market data

In order to calculate the stock movements, for NASDAQ (IXIC), and Dow Jones (DJ), between the closing values from one day and the day before, we made an API call using Alpha Vantage (www.alphavantage.co). Through this API call we collected data on the volume, opening value, closing value, high, and low for each day.

One of the disadvantages of the Alpha Vantage, is that it does not include all historical data, but only data going back 18-25 years depending on the index. We could overcome this issue by using either Google Finance or Yahoo Finance APIs, however both recently closed access to their APIs. This indicates one of the disadvantages of using web based data, the lack of permanence. Web pages and APIs are notoriously unstable. They appear, change, move, and disappear regularly. This can be an issue if either you are scraping data using an API that no longer works or scraping

data from web-pages, which changes their page-codes and selectors. After concluding that the Google and Yahoo Finance APIs no longer worked, we found the Alpha Vantage API through which we were able to collect all stock data needed for this project.

The base function for obtaining data from the Alpha Vantage API call is:

`https://www.alphavantage.co/query?function={}&symbol={}&outputsize={}&apikey={}&datatype={}`

Where:

- Function: The observations / time series of you choice
- Symbol: The name of the index
- Outputsize: Sample or full data, with up to 20 years of back data.
- Apikey: An API key obtained from their website
- Datatype: .csv or. json format

From this URL, we defined a function `("alphavantage_call")` that both gathered, prepped and cleaned the necessary data needed for each stock price and its movement. This included calculation of the percentage change in the closing price movements, making a Boolean value for when the market goes up (1) or down (0), as well as cutting the data based on a pre-defined start date.

Both for the percentage change and the Boolean value we defined a for loop as a function, and included the function within the API call. For the start-date variable, we made an if formula within the API function, which by default just returns the full data, but if a predefined date in a `%y-%m-%d` format, the function will return only the values starting from the defined date.

1.2.Data preparation

1.2.1.Cleaning of data

Before analysing any data, it is important to clean through the data, looking for missing values, or unnecessary information. Before converting the tweets into a Data Frame, the dates were formatted correctly so they could be used at a later stage. We originally scraped all of Donald Trump's tweets from 2015, 2016, 2017 and 2018, this gave us 16,412 observations. It was decided that our study would begin from the date which Trump formally announced his presidential candidacy, the 16th of June 2015. By dropping those tweets from before 16th of June, we reduced our set of observations by 3,725 to 12,687 observations.

The scraping method collected every tweet on Trump's Twitter account. This includes his retweets, his replies and his posting of generic links. We concluded that these were unimportant to our research, as they did not fully reflect the President's sentiments and would affect the effectiveness of our model, so it was decided to remove them from our dataset.

By removing the other accounts Trump had retweeted, we removed 729 observations and were left with 11,958 observations.

During Trump's election candidacy he had many tweets in which he solely quoted a supporter, and then replied with words of thanks, an example being this tweet from the 31st of December 2015:

-""@AnonymousUser: @realDonaldTrump ALL OF AMERICA LOVES TRUMP! #TrumpOrNobody2016 #MakeAmericaGreatAgain" Thank you!"

We concluded that these tweets were not portraying Trump's personal sentiment, as much as legitimising others and so we decided to remove these quoted tweets. This was done by selecting those which started with "@. By dropping these tweets, the number of observations was further reduced by 2,488 to 9,470.

Lastly, during our analysis of Donald Trump's Sentiment, we noticed numerous tweets which were scoring 0.000 on Vader's Sentiment analysis (VADER will be explained later). We considered there to be two possible reasons for this, firstly that these statements were genuinely completely neutral, but after skimming over these tweets it became clear that this was not true. Thus, we developed a second hypothesis, that VADERs software was unable to analyse the sentiment of these tweets. After further inspection of the tweets it was clear that many of them were generic statements containing URLs to various websites, and so a decision was made to exclude these tweets from our sample. After dropping these 1,319 tweets, we were left with our final sample of 8,151 observations.

The final sample included tweets over 820 days in total, each day containing different amount of tweets and different sentiment scores. Since we only have one observation per day of the stock prices, we decided to give each day an average score of sentiment, and this gave us 820 observations to combine with the stock price data.

One of the main difficulties in joining the data from Donald Trump's tweets with the stock market data, is the fact that Donald Trump is tweeting when the stock market is closed. For instance, if Trump tweets on a Tuesday evening after the stock market has closed at 4 pm, the tweet cannot have an effect before the market opens at 9.30 am Wednesday. Also, the stock market is closed during weekends, so any tweets from Friday afternoon and during the weekend can first have a visible effect on Monday. Therefore, we have changed the date of any given tweets from after the

closing of the stock market (8 pm GMT) to the date of the day after. Similar we have changed changed the tweets from the weekends to the date of the following Monday. By doing this we are able to compare the tweets and the changes on the stock market directly without any delays or missing values. In our DataFrame our week is spanning from Monday to Friday with Monday starting at Friday 8 pm GMT and ending Monday 8 pm GMT where Tuesday starts and ends at 8 pm GMT on Tuesday.

1.2.2. VADER Sentiment Analyser

VADER (Valence Aware Dictionary and Sentiment Reasoner) is "a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains." (Hutto, C.J. Gilbert, E.E., 2014). VADER takes a sentence, or in our case a Tweet, as input and returns the polarity of the sentence in four categories – positive, negative, neutral, and compound – a combination of all three. VADER has an internal dictionary which rates each word of a sentence, and forms a score for the whole sentence. The creators of VADER used multiple different people to rate each word to creates a strong collective opinion, making it more reliable than those that use only one source.

This study will use the compound value, which scores between -1 for the tweet with the most negative sentiment, to 1 for the tweet with the most positive sentiment. VADER is particularly effective due to its ability to understand some punctuation, capitalisation, and acronyms. We believed the capitalisation and punctuation abilities would be particularly important for understanding some of Trump's more aggressive tweets.

To properly train the model, it is important to include all the most extreme sentiments, meaning those tweets with a compound score of over 0.8 or less than -0.8. Equally, as mentioned before the neutral tweets (scoring 0.000) will disrupt the training of our dataset, and dilute the quality of our analysis. It is not clear exactly why these tweets scored exactly 0.000, but we believe it could be due to an inability of the system to understand certain types of tweets (including generic sentences and URLs).

1.2.3. Joining the data

When merging the DataFrames, we encountered some observations were missing in both the Stock Data and the Twitter data. In total we were missing 17 observations from Trump's Twitter data. This could either be because Trump did not write any tweets in the given day or because the data was removed from Trump's Twitter profile. In order to complete the data, we assumed Trump had not tweeted on the dates where there was no tweets, meaning that the missing sentiment data would be equal to 0. We now had Twitter data available for all days in the giving period. However

data from the Dow Jones NASDAQ index was still absent for holidays when the market was closed. In order to complete this data, we approximated the missing values by using a concave function, since stock data usually follows a concave function. So, if the stock value on a day is x and the next value present is y with some missing in between. The missing value was approximated to be (y+x)/2 and the same method was followed to fill all the gaps.

1.3.Machine learning

1.3.1.How to split the data

For both of our prediction models it was necessary to split the data into training data (which we will use to make the prediction models) and test data (which we will use to test the accuracy of our prediction models). However, when dealing with time series data such as stock market data, we cannot use the normal random split function to split the data into test and training data sets because this model assumes that the observations are independent. In a time series, the observations are not independent, meaning we must respect the temporal order of the observations. We therefore have to split the data into groups using `TimeSeriesSplit`. This function splits the data in n groups (in our case n=5), where a fixed number of observations (the newest) are the test data and increasing number of observations are included in the training data. A visualisation of the split data sat for the Dow Jones index is provided below. Figure 1 shows the split of the boolean target value of the change in the index. One drawback of using `TimeSeriesSplit` is that the test data is static, and is unsuitable for a dynamic predictive model, as it cannot cope with new added observations.

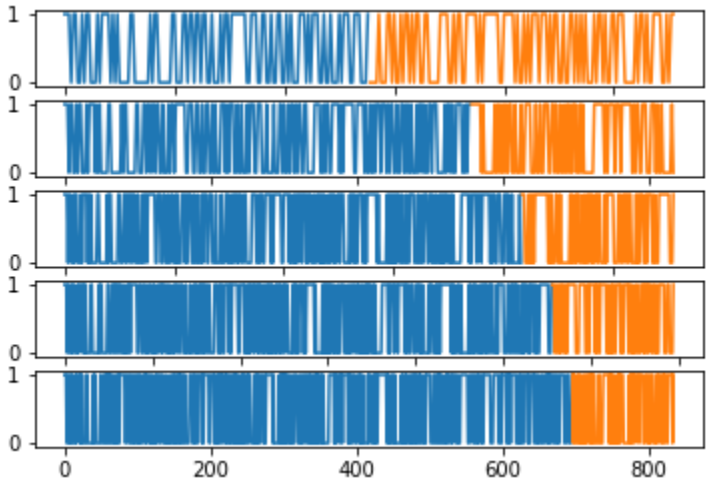


Figure 1: Dow Jones split

1.3.2.Ordinary Least Squares - OLS

Ordinary least squares (OLS) is a method in linear regression for estimation of the unknown parameters in the model. OLS minimizes the sum of the squared vertical distances (the difference between actual and predicted value), also known as the residuals or errors. (Raschka & Vahid, 2017)

OLS has a few areas where it is ineffective, including when data contains outliers or too many variables. Outliers has an effect on the constants, and too many variables can cause the model to include both significant and insignificant variables because it lacks the possibility of model selection.

Due to the fact that OLS has many instances of inefficiency, we decided to look to two regularisation techniques - Lasso (L1) and Ridge (L2). A Linear Regression which suffers from multicollinearity will have a very high variance but very low bias, resulting in overfitting. This means that our estimated values are very spread out from the mean and from one another, it therefore captures the noise and outliers in the data set along with the underlying patterns. This is an issue which Ridge seeks to combat. A model with low variance and high bias would result in underfitting of the data. This means that model is unable to find the underlying patterns within the dataset. This issue is one which we have employed Lasso to tackle.

1.3.3.Ridge

As mentioned above, the Ridge regression is a way to avoid overfitting when the data set contains multicollinarity - a correlation between the predictors. The Ridge model is basically the OLS model, including a squared sum of the weights (Raschka & Vahid, 2017). The L2 term, shown in (2) represents the regularisation term.

$$J(w)_{RIDGE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda||w||_2^2 \tag{1}$$

$$\text{Where: } L2 : \lambda||w||_2^2 = \lambda\sum_{j=1}^{m}(w_j)^2 \tag{2}$$

A ridge regression adds the "squared magnitude" of each coefficient as penalty term to the loss function. So in the case that lambda is zero then the function will be equivalent to an OLS model. However, if lambda is very large then it will add too much weight and causing the model to under-fit. This technique works very well to avoid over-fitting issue.

1.3.4.Least Absolute Shrinkage and Selection Operator - LASSO

LASSO is an extension of OLS and was introduced to improve the accuracy of the predictions. An important feature of the LASSO regression is the possibility of variable selection and regularisation, meaning iterating over different variables and setting some of the coefficients to 0. Which will cause LASSO to remove, some of the insignificant variables OLS keeps, which will improve the overall accuracy.

The notation for LASSO looks as (1), λ is the shrinkage parameter, and chooses the size of the coefficients, as λ goes towards 0, the model goes toward OLS (Raschka & Vahid, 2017). The L1 term, shown in (2) represents the regularisation term.

$$J(w)_{LASSO} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda||w||_1 \tag{3}$$

$$\text{Where: } L1 : \lambda||w||_1 = \lambda \sum_{j=1}^{m} |w_j| \tag{4}$$

The key difference between the Ridge and Lasso techniques is that Lasso shrinks the less important feature's coefficient to zero which can remove some features altogether. This is particularly effective for feature selection in the case of a large number of features.

1.3.5.Decision tree classifier

The decision tree classifier can be used to predict whether the stock market goes up (1) or not (0) on a given day based on Trump's sentiment score, number of retweets, number of favourites, and number of tweets. These variables are our features in the tree classifier. A decision tree classifier works by asking sequential questions for the features and by the answer to these questions choosing a particular route which eventually gives a specific result. Based on the answer to these questions asked by the model, the model can predict a given result. In our case, the model is used to predict whether the stock market change per day is positive or negative which is a binary variable. The decision tree can also be used as a regression tree, but here we are working with the model as a classifier, as our aim is to see whether Donald Trump's sentiment affects the stock market positively and not how large the effects is.

Random forest is a collection of decision trees whose results are aggregated into one final result. In this project we will use random forest as a prediction model. The advantage of using random forest is that random forest isn't as prone to over fitting as using a single decision tree as well as it reduces variance compared to using a single decision tree (Liberman, 2017). However, random forest works by choosing a random bootstrap sample of the training set. Since we are using time series data, we

are not interested in splitting the training data in a random way. Therefore, we will not be using random forest classifier, but instead do a type of k-fold cross validation with TimeSeriesSplit, so we ensure that the data always is split chronologically. The cross-validated decision tree is therefore made by using these five splits from the TimeSeriesSplit and then aggregating these results from the five trees created by using the split.

2. Descriptive statistics and visualisation

This section presents an overview of our data, by presenting descriptive statistics and figures showing the evolution of the American stock market since Donald Trump announced his presidential campaign on the 16th of June 2015.

2.1. The sentiment in Trump's tweets

In order to get the first impression of Donald Trump's sentiment during the last years, we made an average by year, as presented in Table 1. The average increases during the last two years, and presents Donald Trump as slightly positive since the announcement of his presidential.

Table 1: Donald Trump's average sentiment score on tweets during a year

Year	2015	2016	2017	2018
Mean	0.236	0.190	0.231	0.224

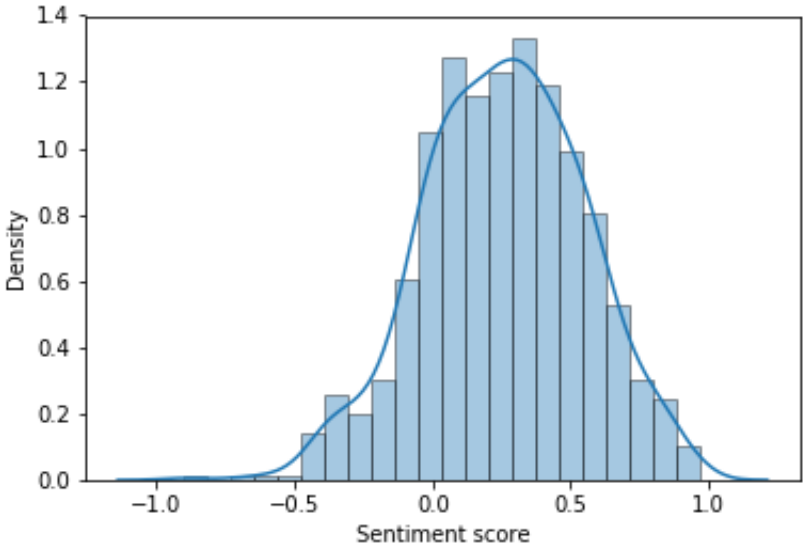
In general, during the period, he has an average score of 0.19, but besides the average positive sentiment of his tweets, the sample contains both very negative and positive tweets, due to the minimum and maximum score presented in Table 2.

Table 2: Descriptive stats

	Max	Min	Mean
Value	0.9899	-0.9864	0.216

Due to the average score during the period, Figure 2 presents a density plot of the distribution of scores. As the average is slightly positive, it is clear that the figure is skewed to the right, meaning Donald Trump has more tweets being slightly positive than negative.

Figure 2: Average sentiment per day



2.2.Stock market prices

While Donald Trump's average sentiment during the period was quite stable, both the Dow Jones and NASDAQ indexes have been increasing. Figure 3 and 4 present the Dow Jones and Nasdaq indices respectively, it is clear that the indices are quite volatile, but are increasing over time.

Figure 3: Dow Jones

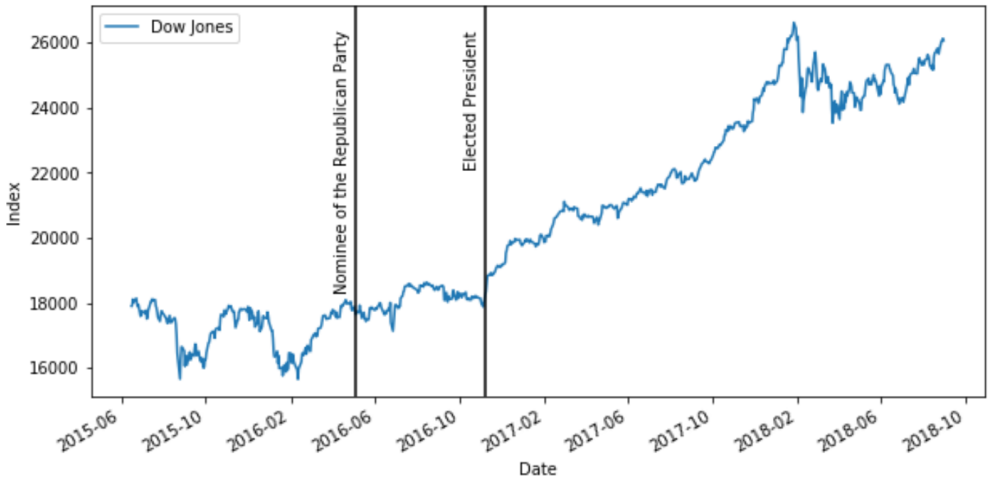
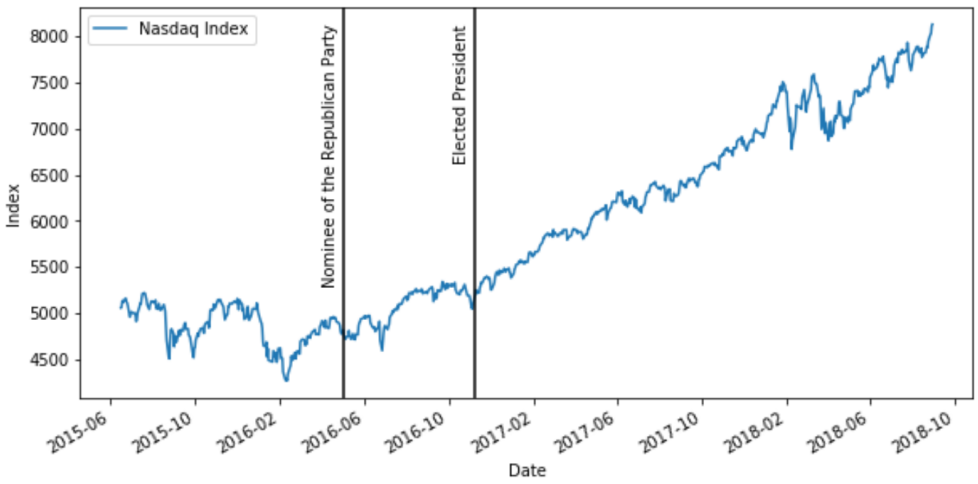


Figure 4: NASDAQ Composite



3. Predictive analysis

The aim with using machine learning is to create different models that can predict the daily change in the American stock market based on the sentiment of Donald Trump's tweets. In this project we will implement different machine learning models that will do the aforementioned prediction. As mentioned, we have chosen to use a LASSO model and Ridge models to improve the usual OLS model when predicting the changes in the stock market per day based on Donald Trump's sentiment score. Furthermore, we have chosen a decision tree as a classification model. This model will be used to predict whether the stock market will go up or down on a given day based on the sentiment of Trump's tweets.

To see whether the sentiment of Donald Trump's tweets have any sort of effect on the stock markets, we thought it would be interesting to see whether we would be able to predict what effect the sentiment of a given tweet would have on the stock markets. We will use Trump's average daily sentiment as our independent variable, and the daily percentage change of the Dow Jones and NASDAQ as our dependent variables. As other predictors we will also included the time passed (which counts the 15th of June 2015 as 1, and adds 1 for each additional day so we are able to model time passed as a quantitative variable), total number of re-tweets per day, total number of favourites per day, total number of tweets per day, and the volume of trading in each stock market.

The accuracy of our model will be measured in Root Mean Squared Error (RSME) which measures the difference between predicted and real population values. The RMSE values enable us to compare different models' ability to predict, and by selecting the model with the lowest RMSE we will find the model with the best ability to predict.

RMSE is calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{5}$$

This section will focus on the results from analysis on the Dow Jones index, whilst a smaller discussion of the NASDAQ values will follow at the end.

3.1.OLS

The Ordinary Least Squares model, a standard linear regression, is the simplest approach that this report will use. Our OLS model, using the `TimeSeriesSplit` validation technique returned an RMSE of 0.00955, which is small. Normally a small RMSE value means that the model is a good predictor, however in the case of our model this is not necessarily the case. Upon further inspection of the fit of our model, we find that the R-Square value is 0.011, which indicates a poorly fit model. A small RMSE combined with a small R-Squared indicates that the best regression model would just be a horizontal line. Furthermore, the F-test is not at an acceptable level to suggest that our model is statistically significant (P value = 0.164).

3.2.Ridge

We employed a Ridge model to attempt to reduce the RMSE of our model, which would enable us to have a better chance of predicting future events. The Ridge model is particularly effective in dealing with multicolinearity, which we felt may have been an issue in our report.

After splitting our data, the Ridge model returns an RMSE of 0.00955, which is only very slightly better than the OLS model. Ridge focus' on data which is over fitted may be a reason for such small change. This led us to believe that our data could be underfitted, and as such we employed a LASSO model.

3.3.LASSO

The LASSO model focus on underfitting. LASSO shrinks the less important feature's of your datasets coefficients to zero thus, removing some feature altogether. The reduction of unimportant variables enables the model to fit even better.

The LASSO model returns an improved RMSE of 0.00836. This could likely be because LASSO is able to reduce the effect that certain variable coefficients have on the model. This will be covered section 3.5.

Comparing across the three models, whilst they all perform similarly, it is clear that the LASSO model is superior to the other two models. However, it is important to note that our dependent variable is measured by a percentage, and so despite the values seemingly being quite small, actually they are almost equivalent to 1 pct. This means that the model and the real data are close to each other, and thus could potentially mean that our model is relatively successful in predicting changes with up to 1 pct. accuracy. But, as mentioned in 3.1., we still have a very small R-Squared, indicating that our model isn't very accurate.

Table 3: RMSE for Dow Jones (OLS, Ridge LASSO)

	RMSE
OLS	0.00955
Ridge	0.00955
LASSO	0.00836

3.4.Learning Curve

A Learning curve shows the relationship between the training set size, and an evaluation metric - in this case the mean squared error. The curves show that as the sample size increases, the errors of the training data are dramatically reduced. This however, means that at larger sample sizes our model is suffering from high bias - the model is under-fitting the data. One potential solution to this would be to increase the number of features, however as we had already added several features to improve the model we didn't believe this was a realistic possibility.

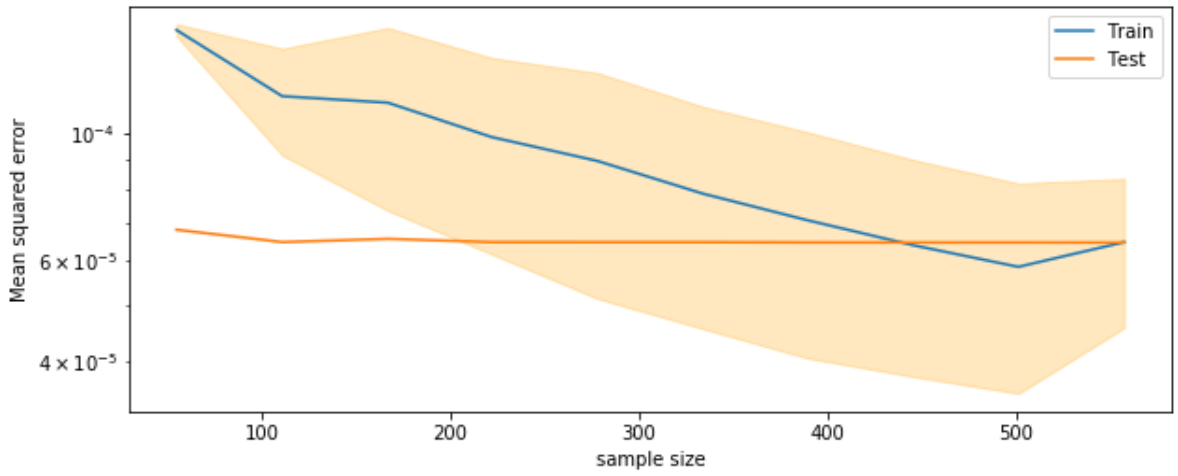


Figure 5: Learning Curve

3.5.Coefficients from the models

Table 4 below shows the coefficients from each of the regression models. In the results from both the Dow Jones and the Nasdaq, we can see that Lasso has removed the sentiment variable from the analysis. This has huge implications for our report, which will be covered in the discussion. Interestingly, the favourites from each day had a positive effect on an increase in the market indices, whereas the retweets had a negative effect. This could have some very interesting implications for further study on this matter, comparing the reach of a tweet (through retweets), to the acceptance of a tweet (through the favourites).

Table 4: Coefficients from the models

	Const	***Sentiment***	***Time passed***	***Retweets per day***	***Faverites per day***	***Tweets per day***	***Volume of trading***
				Dow Jones			
OLS	0.0005	-7.88E-05	4.71E-06	-1.01E-08	2.05E-09	2.68E-05	-8.84E-12
Ridge	0	-4.71E-04	6.46E-06	-7.63E-09	1.36E-09	3.59E-05	-8.49E-12
LASSO	0	0.00E+00	6.28E-06	-6.80E-09	1.27E-09	3.28E-05	-8.48E-12
				NASDAQ			
OLS	0.0088	0.0002	3.84E-06	-1.15E-08	1.39E-09	6.51E-05	-5.03E-12
Ridge	0	-3.23E-04	8.01E-06	-1.22E-08	1.79E-09	5.78E-05	-9.07E-12
LASSO	0	0	7.87E-06	-1.15E-08	1.71E-09	5.46E-05	-9.06E-12

3.6.NASDAQ Results

In Table 5 below, the results for the NASDAQ are presented - in this case the Ridge model is actually has the best ability to predict the markets using Trump's tweets as it has an RMSE of 0.0097555232 (very slightly smaller than Lasso). The results are again all very similar, and small. The fact that the Ridge model is the best suggests that the model could be suffering from multicollinearity. This means that least squares estimates are unbiased, but their variances are large so they may be far from the true value. Ridge adds a degree of bias to the regression estimates, which then reduces the standard errors.

Table 5: RMSE for NASDAQ (OLS, Ridge LASSO)

	RMSE
OLS	0.01042
Ridge	0.00969
LASSO	0.00969

3.7.Decision tree classifier

Table 6: Model accuracy for decision tree classifier

	Model accuracy
Dow Jones	0.496
NASDAQ	0.532

The result from the cross-validated decision tree classifier is a classifier model with an accuracy of 0.496 for Dow Jones data and 0.532 for the NASDAQ index, meaning that our decision tree can predict whether the stock market goes up or not on a given day with a accuracy of 49.6 pct. or 53.2 pct respectively. The L1 term, shown in (2) represents the regularisation term. This might sound like a good model, but considering that we have a binary variable to predict, as the variable for change in the stock market only can take the value 0 or 1, we would actually get a more accurate classification, if we were to predict that the stock market always go up, as both the Dow Jones index and the NASDAQ index rose in over half of our observations. With that in consideration, our models performance is less than impressive. This is a further indication that Donald Trump's Twitter isn't very useful as a prediction tool for the American stock market.

4. Discussion

4.1.General Discussion

Even though our models have small RMSE, it doesn't mean that the prediction models are good. Combining a very small or even negative R-Squared and very small coefficients, there is no indication that Donald Trump's sentiment or any of the other features we use in our models have any economic significance. Our classification model produces a less accurate prediction than a prediction of ever increasing stock prices would have been. The fact that the LASSO model drops the sentiment variable from its analysis indicates that it may not be a good predictor of stock markets at all. Whilst the LASSO model gives us the best RMSE, the fact that it has dropped the sentiment coefficient indicates the insignificance of Trump's sentiment. This allows us to conclude that the model is ineffective at predicting the stock market, using Trump's Twitter sentiment.

Our analysis has limitations, both in the models and in the prediction of the stock prices as a product of which sentiment Donald Trump has while tweeting.

First of all, our models do not contain variables that could cause a change in both the stock prices and the sentiment of Donald Trump, such as policy, events, or attacks. This will cause omitted variable bias.

Secondly, as much as the model should find the effect of the sentiment of Donald Trump's tweets, it doesn't show whether the stock prices at some point have had an effect on Donald Trump's sentiment, meaning that there is an issue with causality in our models. There are likely to be many cases where Trump is tweeting negatively about a fall in the markets, which would cause problems in our results.

Besides this, our model is built upon an average sentiment score of the day, which means that if Trump tweets 18 times one day, the one Tweet that might have had a strong effect on market prices gets neutralised by the rest of the more neutral tweets.

Though our models find that there is not any correlation between Donald Trump's tweets and the American stock market in general, it doesn't mean that Donald Trump's tweets cannot affect the stock market. For instance, when Trump in August 2017 tweeted "Toyota Motor said will build a new plant in Baja, Mexico, to build Corolla cars for U.S. NO WAY! Build plant in U.S. or pay big border tax.", the Toyota stock fell immediately afterwards (FXCM Insights, 2018). This is an indication that when Donald Trump's tweets are about policy or related to certain companies, his tweets can have an effect.

4.2.Ethics Discussion

One of the ethical issues arising from using social media data could be the issue of informed consent. Informed consent is in its classical form derived from research in biomedicine, where it requires informed consent to be given by participants at the point of data collection. (Dingwall et. al, 2003). An informed consent can be said to have been given when there is a clear understanding of the implications, and consequences of an action.

Though Twitter's API the Trump Twitter Archive, it is possible to collect tweets without users being aware at all that their tweets are being collected. Users are typically not approached directly to give their informed consent and take part in research. Instead, in the case of research involving Twitter data, informed consent and the acceptance by the user is typically assumed to have been given by the user acceptance of Twitter's Terms of Service that licenses Twitter to make ones content available for others (Twitter.com, 2018).

Oxford produced an updated guidance (Webb et al., 2017) on internet-based research. This advises that in the case of Twitter (as a public platform) researchers do not need to solicit consent to collect data but should seek consent to publish individual posts. Alternatively, they can create composite data for the purpose of publication.

In our case, it is quite difficult to anonymise the data, since the data is only taken from one

specific individual Twitter account whom the whole research questions is based upon. In general data accessed from open and public online locations, such as Twitter, present less ethical issues than data which are found in closed or private online spaces. Similarly, data posted by public figures such as politicians or celebrities on their public social media pages is less likely to be problematic, because this data is intended to reach as wide an audience as possible. Since Twitter is open by nature, where users already have given consent for the use of the data and since Trump is a public figure, the data is less likely to be problematic and not entitled to the scrutiny of an ethics panel (Socialdatalab.net, 2018). Also, since we are not using any sensitive information we would argue that we are not using harmful information. We have only used quotes to visualise "thank you" tweets to his users who has been anonymised.

4.3.Further Study

The focus of this project has been how Donald Trump's general behaviour on Twitter affects the American stock market and not how specific, policy related tweets can have an effect. Yet, looking at Donald Trump's tweets regarding policy and/or specific companies would be an interesting subject for further study. A model which selected specific tweets using the number of retweets to indicate their importance, would perhaps have been more effective in predicting the markets.

A further topic of interest could have been to compare Donald Trump to other key influencers around the globe, to really understand whether key users on Twitter can truly affect the worlds financial markets. This would again raise further questions surrounding Trump's integrity, and whether there is a conflict of interest in the President of the United States having a personal account which is able to influence so many parts of society.

5. Conclusion

We started this project by looking into the stock market prices and finding out if Donald Trump's mood can be used for predicting the stock market. After scraping the Twitter archive, we collected around 16,000 observations and narrowed it down to around 800 days after combining it with the stock market data. Our descriptive analysis shows how Donald Trump's tweets are both positive and negative, but also how the stock market prices have been increasing since 2015.

So, to answer our research question, to what extent are the sentiments of Donald Trump's tweets able to predict changes in US stock markets on any given day? In the model we have created, unfortunately not at all. Trump's Twitter continues to be a never ending source of new information, and whilst we suspect that specific tweets can still have huge effects on the markets, overall our model is unable to predict any effects. Through our predictive models, we were able to conclude that actually a constant line would have been just as good a predictor as our model, due to the R squared value combined with tiny RMSEs.

Even though Donald Trump's behaviour on Twitter is a subject of debate and specific tweets from him regarding policy or certain companies might cause shocks to the market, our models do not indicate that his general behaviour on the social media platform has any effect on the American stock market in general and therefore cannot be used as a prediction tool for how the markets might change on a given day.

Bilbiography

- Asur, S. and Huberman, B.A. (2010). Predicting the Future with Social Media. In: Proceedings of the ACM International Conference on Web Intelligence, pp. 492-499.
- Bae, Y. and Lee, H. (2012). Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. Journal of the American Society for Information Science and Technology, 63(12), pp.2521-2535.
- Bollen, J., Mao, H. and Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), pp.1-8.
- Bulman, M. (2017). 35 psychiatrists just met at Yale to warn Donald Trump has a 'dangerous mental illness'. [online] The Independent. Available at: `https://www.independent.co.uk/news/world-0/donald-trump-dangerous-mental-illness-yale-psychiatrist-conference-us-president-unfit-james-gartner-a7694316.html` [Accessed 27 Aug. 2018].
- Coppins, M. (2018). The American Obsession with Donald Trump's Mood. [online] The Atlantic. Available at: `https://www.theatlantic.com/politics/archive/2018/08/our-obsession-with-donald-Trump's-mood/568414/` [Accessed 27 Aug. 2018].
- Dingwall, R. and Murphy, E. 2003. Qualitative Methods and Health
- Eiji, A. (2011). Twitter Catches the Flu: Detecting Influenza Epidemics using Twitter. 1568-1576.
- Fortune (2018). http://fortune.com. [online] Fortune. Available at: http://fortune.com/2018/04/07/donald-trump-tweets-stock-market/ [Accessed 30 Aug. 2018].
- FXCM Insights (2018) https://www.fxcm.com [online]. Availiable at: https://www.fxcm.com/insights/president-Trump's-twitter-impact-forex-markets-stocks/ [Accessed 30 Aug. 2018].
- Gilbert, E. and Karahalios, K. (2010). Widespread Worry and the Stock Market. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp.58-65.
- Hutto, C.J. Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
- Internet Live Stats (2018). Twitter Usage Statistics - Internet Live Stats. [online] Internetlivestats.com. Available at: `http://www.internetlivestats.com/twitter-statistics/` [Accessed 27 Aug. 2018].
- Java, A., Song, X., Finin, T. and Tseng, B. (2007). Why we twitter. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA-KDD '07.
- Jin, F., Wang, W., Chakraborty, P., Self, N., Chen, F. and Ramakrishnan, N. (2017). Tracking Multiple Social Media for Stock Market Event Prediction. Advances in Data Mining. Applications and Theoretical Aspects, pp.16-30
- Liberman, N. (2017). Decision Trees and Random Forests. Available at: `https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991`

- Mittal, A. (2011). Stock Prediction Using Twitter Sentiment Analysis.
- Pagolu, V., Reddy, K., Panda, G. and Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES).
- Rao, T. and Srivastava, S. (2012). Analyzing Stock Market Movements Using Twitter Sentiment Analysis. /ACM International Conference on Advances in Social Networks Analysis and Mining.
- Raschka, Sebastian and Mirjalili, Vahid (2017) Python Machine Learning.
- Serban, I., Gonzalez, D. and Wu, X. (n.d.). Prediction of changes in the stock market using twitter and sentiment analysis. University College London.
- Socialdatalab.net. (2018). Ethics Resources – Social Data Science Lab. [online] Available at: `http://socialdatalab.net/ethics-resources` [Accessed 30 Aug. 2018].
- Twitter.com. (2018). Twitter Terms of Service. [online] Available at: `https://twitter.com/tos?lang=en#us` [Accessed 30 Aug. 2018].
- Webb, H., Jirotka, M., Stahl, B., Housley, W., Edwards, A., Williams, M., Procter, R., Rana, O. and Burnap, P. (2017). The Ethical Challenges of Publishing Twitter Data for Research Dissemination. Proceedings of the 2017 ACM on Web Science Conference - WebSci '17.
- Zhang, X., Fuehres, H. and Gloor, P. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". Procedia - Social and Behavioral Sciences, 26, pp.55-62. Policy Research. New York: Aldine de Gruyter