STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
659 Degree project in economics
Spring 2016

The 2016 US Primary Elections and Twitter

A methodological study of econometrics, machine learning and their intersection

Adrian Ahmadi (23190) and Robert Hu (23102)

Abstract: In this thesis modern methods in using social media data from Twitter to predict the 2016 US primary elections are investigated. The collected data is processed via a simplified sentiment analysis and later used in modelling election results with less traditional linear models and methods within statistical learning. This investigation uses a limited methodology in sentiment analysis since a rigorous analysis would require methods within natural language processing, which is a thesis topic on its own.

The methods used in this thesis achieves significant results and there is potential for improvement using more exact analysis. Further, it is clear that the more refined methods yields more precise estimates.

In conclusion, the results suggests that the approach is highly plausible for future research and under less bold assumptions. This is concluded from significant results despite a relatively simplified approach.

Keywords: Big data, Large Data Sets, Statistical Learning, Sentiment Analysis

JEL: C550

Supervisor:	Maria Perrotta Berlin
Date submitted:	May 13th 2016
Date examined:	June 7th 2016
Discussant:	Mamud Miyan and Sara Davidsson
Examiner:	Kerem Coşar

Acknowledgements

We would like to thank Maria Perrotta Berlin for supervising this thesis and giving us feedback and guidance throughout the process. Further, we would also like to thank Örjan Sjöberg, for encouraging our work and providing inspiration and our fellow economics students, for providing feedback and suggestions for our thesis. Lastly, we would like to thank our classmates, friends and our family for bearing with our shenanigans for the last three years. Unfortunately, this will probably not be the end of them.

Adrian Ahmadi
Stockholm, 2016-05-07

Robert Hu
Stockholm, 2016-05-07

Table of Contents

List of Figures

Abbreviations

AIC	Akaike Information Criterion
DNC	Democratic National Committee
BLUES	Best Linear Unbiased EStimator
GOP	The Republican Party
MSS	Mean Sum of Squares
OLS	Ordinary Least Square
SE	Standard Error
SSE	Sum of Squared Errors
SSR	Sum of Squared Residuals

Symbols

X	Matrix of Covariates
β	Regression Coefficient Vector
e	The Error Vector
x_i	The i:th Column of X
A^{-1}	The Inverse of a Matrix A
$\|x\|$	The Norm of a Vector x
k	The Number of Covariates
n	The Number of Observations
α	The Significance Level
A^T	The Transpose of a Matrix A
Y	The Vector of Independent Variables

1. Introduction

Background to the problem is outlined. Further the academic contribution as well as the study aim and limitations are discussed.

1.1 Background

The 2016 US presidential election exhibits some of the most controversial personalities and opinions seen in the modern times. One candidate, if elected president, intends to build a wall on the southern American border and make the southern neighbour pay for it. Another candidate displays, what in the US is generally considered as, strong socialist tendencies. All of this in a digitalised world where public opinion frequents social media. This paves the way for interesting questions such as to what extent the opinions aired on social media reflects the real world political landscape.

1.2 Problem Discussion

One of the recent global mega-trends that have had a significant impact on society and businesses is the phenomena of digitalization.[1] A term which is closely related to this trend is *Big Data*. No doubt over 4 million hits on Google Scholar[2] show how this relatively recent field of research has intrigued academics and led to a tremendous amount of publications. With over 300 million so-called tweets posted each day[3] on the micro-blog Twitter it is a very promising candidate for conducting Big Data related research.

Can a 140 character compact tweet really convey a political message? If you ask the campaign manager of a serious presidential candidate the answer is, at least to some extent, yes as Twitter is one of their main channel of communication with potential voters. The online presence of the 2016 White House candidates is unsurprisingly stronger than ever before. With everything from important political topics to childish dramas unfolding in the candidates Twitter feeds, filled by both proponents and opponents, it is interesting to see whether any value can be derived from these feeds. For the campaign manager this value can come in form of more votes. For the eager econometrician perhaps the value can be derived from how well he or she can use the data to make inference and forecast the outcome of future elections.

[1] *B. El Darwiche et al., Digitization for economic growth and job creation Regional and industry perspectives Accessed April 21st 2016*

[2] *J. Manyika et al., Big data: The next frontier for innovation, competition, and productivity Accessed April 21st 2016*

[3] *J. Edwards, Leaked Twitter API data shows the number of tweets is in serious decline Accessed 2016-04-21*

1.3 Problem Formulation

This study aims to use methods which treat vast amounts of data from social media, several orders of magnitude more than the average econometric study at this academic level, in order to make successful predictions. To test the methods the paper will investigate the 2016 US presidential primary elections. More specifically the study aims to assess how well US presidential primary elections can be predicted using Twitter data. It intends to use sentiment analysis to do so. Thus the research question can be defined as:

- *How well can the outcome of US presidential primaries be predicted using Twitter data?*

1.4 Study Aim and Limitations

The aim of the study is to make a methodological contribution by investigating the usefulness of models and methods which uses considerable amounts of data. If the models and techniques used are proven successful it may possibly have implications for other fields, such as online marketing.

Specifically, for the field of Economics, we make an attempt to contribute to the study of prediction markets in the hope of providing a new simplified approach of extracting Twitter data and using it to perform predictions. Since the idea of prediction markets is a marketplace for trading the outcome of events, i.e. betting, we hope that this thesis might offer new methods to utilize digital data more efficiently in order to increase the accuracy of certain prediction models used in prediction markets to predict the probability of certain events occurring. Prediction markets are as mentioned primarily for betting, but they also have an interesting side effect: They serve as an indicator of some event occurring, depending on the odds they are offering for the outcomes of that event. As an example, most prediction markets have offered very high odds that Donald Trump would become the republican nominee, suggesting that Donald Trump is very unlikely to get the nomination.

In order to be able to conduct the study a series of limitations are imposed. First of all the paper is limited to Twitter data for the simple reason that it is the only major social media platform which allows for free collection of data. Only primaries from 2016 are considered, as the data can only be collected in real-time this is a reasonable, and necessary, limitation. Further, only candidates running as of April 1st are considered as that is the approximate starting period of this study. For obvious reasons not all tweets available are analysed, although the sample size is considered to be large enough not to give rise to any problems regarding sample size. Moreover, irony in the tweets will not be addressed as this would require methods far beyond the scope of this study. In addition to this, demographic data and other variables which cannot be deduced from Twitter will not be considered. A point brought forward in previous research is that the incumbent effect should be considered but as there is no clear incumbent in the 2016 US primaries this will be hard. Finally, assumptions are made on autocorrelation on the data. Two cases of autocorrelation are considered in this study.

2. Theoretical Background

Previous research in the area is discussed. Further the relevant theory associated with sentiment analysis and multiple linear regressions are reviewed.

2.1 Previous Research

One of the first attempts to use Twitter to predict election results were carried out by Tumasjan, Sprenger, Sandner, & Welpe (2010). The paper start by establishing, after analysis of over 100,000 tweets, that tweets indeed, and unsurprisingly, contain political messages. Further, Tumasjan et. al. claim the number of mentions on social media is positively correlated with actual election results for the 2009 German general election. Definitely worth mentioning is that, instead of using the popular econometric method of least square errors, the authors utilise a Mean Absolute Error, MAE, estimator. Further, instead of comparing the predictions to actual outcome they are compared to polls, which are widely regarded to contain many degrees of uncertainty.

Using Twitter sentiment to predict the outcome of elections has experienced different levels of success in past papers. Among the ones who claim, at least modest, success in the field are Choy, Cheon, Nang Laik, & Ping Shung (2011). They collected in total 16,616 tweets during the first eight days of the 2011 Singaporean presidential election. Further, the data is processed to get of rid of problems such as duplicate tweets. Choy et. al. makes two central assumptions in their framework, namely:

- The people who voted in the general elections are most likely to be voting along the party lines.
- The online sentiment is representative of the people who are expressing their views.

Following these assumptions the aggregate sentiment for each candidate is calculated. In order to predict the expected percentage of votes for each candidate the aggregated sentiment is used together with; demographic data, percentage of people using computers per age group, percentage of people using social media per age group and percentage of people for party and candidate in each age group. In simpler words Choy et. al. (2011) use a set of tweets to aggregate the Twitter sentiment for each candidate and consequently generalise this to the entire population by using demographic data. The predictions in the paper are proven fairly accurate for three out of the four candidates. However, the winner is not correctly calculated.

Gayo-Avello (2012) thoroughly walks the reader through both the successful and unsuccessful cases of where Twitter data is used for prediction. The paper finds that the predictive power of Twitter regarding elections has been greatly exaggerated. One of the arguments that are highlighted throughout the paper is that in previous research all tweets are assumed to be trustworthy, e.g. the sometimes very ironic and sarcastic climate of social media is ignored. Another argument is that many researchers apply sentiment analysis as a

black-box tool instead of actually going through the effort of understanding and correctly applying sentiment analysis. The main points of criticism can be summarised as:

1. Majority of papers focus on predicting result post-election. I.e. no real prediction is done.
2. Incumbent effect is disregarded.
3. No consensus on how to apply sentiment analysis.
4. No common basis for comparison between different predictive models.
5. Sentiment analysis is blindly applied.
6. Astro-turfing[4] is ignored.
7. No adjustment due to demographics.
8. Self-selection bias is ignored.

After incorporating the above feedback Choy et. al. applies a slightly modified method in (2012) where they successfully predict the outcome of the 2012 US presidential election to be Barack Obama. However they find it hard to address all the points brought forward by Gayo-Avello (2012). The criticism regarding astro-turfing and self-selection bias are pointed out as the most difficult and costly to address as it in many cases would require reading the tweets manually.

But not all attempts to predict election results Twitter are as positive as Choy et. al. (2012). In a critical paper (Metaxas, Mustafaraj, & Gayo-Avello, 2011), after analysing the results from multiple elections, claims that Twitter data predictions are only slightly better than the baseline, which is taken to be incumbency. The authors stress the importance of a clearly defined procedure for sentiment analysis in order to avoid a black-box approach.

Adam Bermingham and Alan F. Smeaton writes in their paper (2011), that there indeed exists predictive power in social analytics based on volume-based measures and sentiment analysis. In their conclusion they find that volume may be a stronger predictor than sentiment in the sense that sentiment is more reactive and thus more often reflect an immediate response to an event or piece of news rather than a consistent political standing. On a technical note, they also use the MAE.

To summarise the previous research, there is no clear consensus as to Twitter's predictive power. Neither does a unified approach for sentiment analysis exist.

[4] *Astro-turfing is when support of an agenda is portrayed as "grassroot" social movement.*

2.2 Modelling

As sentiment analysis and linear regression are the methods used throughout the paper the main theory relevant to the study are briefly presented and discussed in this section.

2.2.1 Sentiment Analysis

In this study, it is necessary to quantify the sentiment of a tweet. Intuitively, what we do is given an arbitrary tweet, we count the number of "positive" words as +1 and "negative" words as -1 in the tweet and sum everything in the end. To get a measure of whether the tweet was positive or negative.

When we say "positive" and "negative" words, we simply mean words attributed to describing something positive and negative respectively. In order to know which words are positive and negative we need predefined lists of positive and negative words. These predefined lists are acquired as web resources from professor Bing Liu´s (a prominent researcher within the field of sentiment analysis) personal webpage. The lists are accumulated words associated with positive and negative opinions that have been continuously mined from large quantities of customer reviews.

In the approach used in this paper each positive word in a tweet results in the sentiment score being increased by one (starting with 0), and each negative word results in the sentiment score being reduced by one. The sentiment score is then normalised using the length of the tweet. For example the tweet "Hillary is a good person." contains one positive word (good) and five word in total (Hillary, is, a, good, person). The normalised sentiment score would thus be 1/5.

In order to formalize this, measure theory is needed. In this thesis a signed measure is constructed in order to measure sentiment given a universal set Ω and a corresponding sigma algebra $\mathcal{F} = \sigma(\Omega)$. The definition of a signed measure is the following:

Given a measure space $(\Omega, \mathcal{F})$, a signed measure μ is a mapping $\mu: \mathcal{F} \to [-\infty, \infty]$ such that:

- $\mu(\emptyset) = 0$
- μ assumes at most one of the values $\infty, -\infty$
- If $\{E_i\}$ is a sequence of disjoint sets in $\mathcal{F}$ then:

$$\mu(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i) \ , (\sigma - additivity) \tag{1}$$

In our case, the universal set will be the union of three disjoint sets $\Omega := \mathcal{O} \cup \mathcal{G} \cup \mathcal{B}$, where $\mathcal{G}$ is the set of all words associated to positive sentiment, and $\mathcal{B}$ is the set of words associated with negative sentiment and $\mathcal{O}$ is the set of words without any sentiment. In our case $\mathcal{G}$ is the list of positive words and $\mathcal{B}$ the list of negative words.

Our measure is then defined as:

$$\mu(\tau) = |\mathcal{G} \cap \tau| - |\mathcal{B} \cap \tau| \quad (2)$$

Where $\tau \subset \sigma(\Omega)$ is the tweet set, i.e. the set of words that the tweet contains. For a proof that this indeed a signed measure please see Appendix 5.

2.2.2 Dummy Variables

A dummy variable, or indicator variable, is a variable which takes on the values 0 or 1. It is common to use dummy variables in econometrics in order to indicate the presence or absence of some categorical effect. For example if one seeks to assess the difference between the salary of men and women one should include a dummy variable for woman, or man, where a 1 indicates that the salary is associated with a woman and 0 if it is not. It is important to make sure not to include a dummy variable for man as it would render in perfect multicollinearity.

2.2.3 Interaction Effects

In order to capture the interaction effects in a regression model it is useful to introduce a new covariate which is the product of two or more other covariates. E.g. if one seeks to test if the return to education is the same for both sexes one can regress the logarithm of the wage on experience, education and the product between a female dummy and education. Here it is assumed that the return to experience is the same for men and women. Under the null hypothesis the slope coefficient for the new covariate, namely female dummy times education, is zero.

2.2.4 Ordinary Least Squares

In this study the popular Ordinary Least Square, OLS, is used as opposed to the Mean Absolute Error estimator used in for example Choy et. al. (2012). The difference is that OLS seeks to minimise the sum of squared residuals while the MAE seeks to minimise the sum of absolute errors.

The OLS estimator is given as:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (3)$$

The popularity of the OLS follows from it being the Best Linear Unbiased Estimator, BLUES, as proven by e.g. Lang (Elements of Regression Analysis, 2014, s. 7). Unbiased meaning that the estimated beta will tend to the true beta as the numbers of observation goes to infinity.

2.2.5 Underlying Assumptions

In order to get unbiased and correct results when using the OLS estimator the assumptions below should be fulfilled.

- Linear relationship between the dependent and independent variables. As mentioned before this is almost never limiting as one could easily transform the desired variables into the desired form.

- No close-to-perfect multicollinearity. This is never a problem when the dataset is sufficiently large.
- $Var(e_i) = \sigma^2$. I.e. no heteroscedasticity. More on this later in the theory section.
- No autocorrelation, i.e. the errors from different observations should not be correlated.
- The residuals follows a Gaussian distribution with mean zero. Please note that as long as the residuals are independent and identically distributed random variables most cases will render in meaningful results. I.e. non-normality will not always cause issues.

2.2.6 Hypothesis Testing

After estimating the regression coefficients using OLS one typically wants to test for the null hypothesis, H_0, of the coefficient being equal to zero. Thus H_0 used throughout this study can be formulated as:

$$H_0: \hat{\beta}_j = 0 \tag{4}$$

where $\hat{\beta}_j$ is the OLS estimate for the j:th covariate.

The null hypothesis is tested against the alternative hypothesis, H_1, formulated as:

$$H_1: \hat{\beta}_j \neq 0 \tag{5}$$

The hypothesises are tested at the significance level $\alpha = 0.05$ as this is the conventional level most commonly used in econometric papers. Thus the probability of rejecting a true null hypothesis is $\alpha = 0.05$.

To test the null hypothesis it one can employ the F-statistic defined as following for one covariate:

$$F = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \right)^2 \tag{6}$$

The null hypothesis is rejected if the following is true:

$$F > F_{\alpha,\ 1,\ n-k-1} \tag{7}$$

where $F_{\alpha,\ 1,\ n-k-1}$ is the tabulated F-value for k number of covariates, n the number of observations and significance level α. If the null hypothesis is not rejected it implies a dependence between the covariate and the independent variable.

The lowest level α for which (9) still holds is called the p-value. In this study the conventional codes for significance defined as follows are used.

p-value	*Significance code*
≤ 0.1	.
≤ 0.05	*
≤ 0.01	**
≤ 0.001	***

Table 1: *Significance levels associated with each p-value. p-values larger than 0.1 have no significance code.*

2.2.7 Model Efficiency

There are several techniques to determine the efficiency of a model. The ones used in this study are presented below.

2.2.7.1 Coefficient of Determination – R^2

The most commonly used measure for model efficiency is the coefficient of determination, R^2. It essentially measures how much of the total variance the model explains, and is defined as:

$$R^2 = \frac{|\hat{e}_*|^2 - |\hat{e}|^2}{|\hat{e}_*|^2} \tag{8}$$

where $\hat{e}_*$ is the residuals from the regression on only a constant term and $\hat{e}$ the residuals from the considered model. It is sometimes warranted to use the adjusted R^2, which penalises overfitting, but due to the vast amount of data used in this study it is not needed.

2.2.7.2 Akaike Information Criterion – AIC

The Akaike Information Criterion, AIC, is a relative measure of overall model efficiency. Thus it is a means of model selection. One seeks the model which minimises the AIC. The measure is defined as:

$$AIC = n\,ln(SSR) + 2k \tag{9}$$

where k is the number of covariates, n is the number of observations and SSR is the sum of squared residuals. An interpretation of the formula is that AIC rewards goodness of fit while penalising over-fitting. In practice, statistical software with built-in algorithms for finding the model with the lowest AIC is used.

2.3 Problems Associated with Regression Analysis

If the assumptions regarding linear regressions presented earlier are violated several problems may arise. The problems relevant to this study, and suggested remedies, are presented in this section.

2.3.1 Heteroscedasticity

Heteroscedasticity, or the absence of homoscedasticity, is when the underlying assumption of equal variance for all residuals is violated. If heteroscedasticity exists one will not get reliable results of an F-test. A common case where one would suspect heteroscedasticity is when employing GDP as a dependent variable, as this varies greatly across countries of different size, population and level of development. In this case one can make a smarter choice of variable, e.g. GDP per capita, or even better the logarithm of GDP per capita.

There are several techniques available for detecting heteroscedasticity. The one used in this study is called the Breusch-Pagan test. To carry out the test one first run the model one wants to investigate. Secondly, one regresses the squared residuals on the original covariates. More formally:

$$\hat{e}_i^2 = x_i\gamma + v_i \qquad i = 1, \dots, n \tag{10}$$

where γ is the regression coefficient and v_i the residuals. Subsequently an F-test is used to test for the null hypothesis of homoscedasticity, H_0: $\gamma = 0$.

2.3.2 Remedies for Heteroscedasticity

If one cannot find a better choice of variable as suggested in the GDP example above one can instead try to use a consistent variance estimator. One popular consistent variance estimator is White's estimator (A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, 1980). It is defined as:

$$Cov(\hat{\beta}) = \frac{n}{n-k-1}(X^TX)^{-1}X^T\left(\sum_{i=1}^{n}\hat{e}_i^2 x_i^T x_i\right)X(X^TX)^{-1} \tag{11}$$

According to Lang (Elements of Regression Analysis, 2014, s. 17) it is warranted to always employ White's estimator. Therefore White's estimator will be used throughout this study.

2.4 Statistical Learning

In this thesis, methods of statistical learning are briefly explored. A good choice of reference literature in the field is Friedman, Hastie, & Tibshirani (The Elements of Statistical Learning: Data Mining, Inference, and Prediction., 2009).

2.4.1 Basic Idea

Statistical learning is a subset of methods within machine learning that focuses on methodologies within mathematical statistics. Essentially, statistical learning deals with the problem of finding a predictive function based on a given dataset.

The primary reason for exploring this area is to investigate if the methods presented can be of use to improve predictive econometric models. Specifically, one can argue that the advantage of statistical learning is that the methods can adapt to data that changes frequently, for example opinions on Twitter, stock market data and public approval ratings.

In the case of traditional econometrics, models are constructed and regressed on large datasets in order to find patterns and significant inferences. In this case, a framework of methods that can adapt a model as new data arrives in order to maintain predictive power are used.

2.4.2 Mathematical formulation

Similarly to the ordinary regression, the problem of finding the best linear predictor using statistical learning can be formulated as an optimization problem:

$$\min_{w \in \mathbb{R}^d} f(w) \tag{12}$$

where $f(w)$ is a convex function defined as:

$$f(w) := \lambda R(w) + \frac{1}{n}\sum_{i=1}^{n} L(w^T x_i, y_i) \tag{13}$$

where w is the weights of the linear predictor for the vector $x_i \in \mathbb{R}^d, \lambda$ is a real valued constant and $R(w)$ and $L(w^T x_i, y_i)$ denotes the regularization function and loss function respectively. Further, n denotes the number of datapoints used and d denotes the number of covariates. In this context, w can be interpreted as the $\hat{\beta}$ obtained from regular regressions.

The loss function $L(w^T x_i, y_i)$ can be seen as a punishment of inaccuracy for the linear predictor. Further, one also have a regularization function $R(w)$ which is used to prevent overfitting of the predictor. This can be seen as a punishment on having large coefficients when fitting a model. When the coefficients, w, are too large overfitting occurs, as seen in the right box in the picture below:

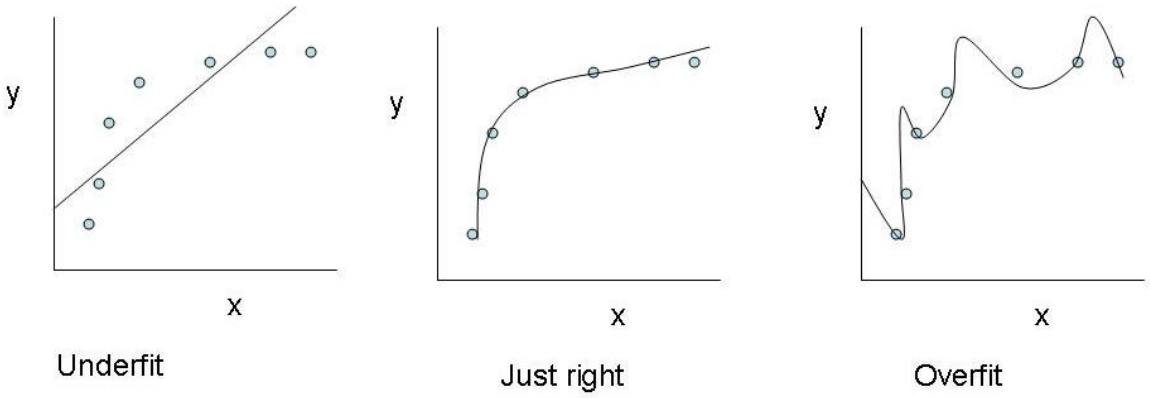

Figure 1: *Picture showing different fits achieved by using statistical learning.*

In order to avoid this behaviour a punishment on overfitting is imposed. Here λ is chosen as the impact of overfitting a model has, i.e. a larger λ leads to stronger bias against overfitting.

In general, a predictor is trained on a dataset known as the "training set" and tested against a dataset known as "test set". The results from testing is then measured with the help of a metric usually selected to be the prediction accuracy. Accuracy in this case, is defined as number of correctly predicted points in the test set divided by the total number of points in the test set.

2.4.3 The Logit Predictor

When investigating statistical learning, the logit estimation is used as a predictor. This implies that the loss function and regularization function becomes:

$$L(w^T x_i, y_i) \coloneqq \log\left(1 + e^{-y_i w^T x_i}\right), y \in \{-1,1\} \tag{14}$$

$$R(w) \coloneqq \frac{1}{2}|w|^2 \tag{15}$$

where $L(w^T x_i, y_i)$ essentially defines the maximum likelihood estimator. It turns out that the estimator in general lacks a closed form solution and thus numerical methods such as Newton's method, gradient descent or in our case the Broyden–Fletcher–Goldfarb–Shanno, BFGS, algorithm are necessary. The BFGS algorithm is essentially a variation of Newton's method, but less computational heavy.

3. Method

Methods and means of data collection are described. Further the models used are introduced.

The computer programs and libraries used in this study to analyse, collect, compile and sort data are MongoDB, Python, Scala, Apache Spark and R.

It is very important to note that every individual tweet is used as an observation in the regression as one do not know if there would exist a direct correlation between an aggregated data point compared to just one tweet. In order to obtain a model that essentially takes the mean given a regression on a large set of tweets, we regress on each individual tweet rather than aggregated data points. The prediction thus instead takes a mean over a sample of tweets rather than a lot of aggregated ones.

3.1 Setup

The general setup of software used is that MongoDB is used as a database, Python for pre-processing, R for regular regressions and Scala with Apache Spark for the statistical learning approach.

It is recommended to have previous programming experience in order to attempt recreating this setup. To properly set everything up, Python 2.7.11 is installed together with the PyMongo and Tweepy API, which can be readily installed through the command prompt calling “pip install”. A more comprehensive tutorial can be found at Sean Dolinar’s webpage, “Collecting Twitter data: Storing tweets in MongoDB”.

For the statistical learning setup, the programming language Scala must first be installed. After this, Apache Spark must be built together with Scala. It is favourable to adopt the practices of the guide “How to Build Apache Spark on Windows 8”.

The setup and scripts used have been uploaded and can be found at this GitHub page.

3.2 Data Sources

The sources of data are tweets and election results. The election results are collected from realclearpolitics.com as a table and then saved as a .csv file, which is later processed together with the tweets into a table format.

3.3 Data Collection

The data collection process is carried out using the Twitter API “Tweepy”. Tweepy allows us to specify a list of search words, or tags, and to collect the stream of tweets that match these. This stream of data is then downloaded and stored locally using the database tool MongoDB, which helps store the tweets in a compressed and readily accessible manner.

The data collection process is implemented through a script, which is looped over the search queries of candidates for 3 hours evenly distributed over 3 days before the primary. Then

while the stream is open, every tweet recorded is stored in a compressed manner using MongoDB. In order to simplify the handling of MongoDB, a .bat script is used to make activation of MongoDB easier.

It is important to use as objective keywords as possible, in order to obtain as unbiased data as possible. Therefore some popular keywords associated with the candidates such as "MakeAmericaGreatAgain" and "FeelTheBern", for Mr Trump and Senator Sanders respectively, have been excluded.

The following tags are used throughout the study:

Donald Trump	***Ted Cruz***	***John Kasich***	***Hillary Clinton***	***Bernie Sanders***
Trump	Cruz	Kasich	Clinton	Sanders
TRUMP	tedcruz	KASICH	hillaryclinton	SANDERS
Donald Trump	CRUZ	John Kasich	CLINTON	berniesanders
donaldjtrump	Ted Cruz	johnkasich	Hillary Clinton	Bernie Sanders
realDonaldTrump	TedCruz	JohnKasich	HillaryClinton	BernieSanders

Table 2: *Tags used for each candidate*

3.4 Pre-processing

As the tweets are recorded, they are stored in a .json format which requires pre-processing before further analysis. The fields recorded for each tweet are:

- Timestamp
- Number of followers of the account that posted the tweet
- Number of friends of the account that posted the tweet
- Number of retweets of the posted tweets
- Number of times the tweet has been marked as a favourite
- User location
- Sentiment score
- Dummy variable for the candidate associated with the tweet
- Dummy variable for the state the primary was held in
- Actual outcome of the primary
- Dummy variable for the winner of the primary

3.5 Sentiment Processing

In order to obtain this data, algorithms and natural language processing are in place. This section will briefly explain the method used.

The first part is to identify which candidate the tweet is about. This is done by using pre-specified lists of candidate tags, where each tag is associated with a candidate. Formally, these lists can be seen as sets $T_i = \{'\text{candidate name}', '@\text{candidate}', \text{candidate indetifier 2}', ...\}$ Then given a tweet that is decomposed into a set of words $\tau := \{'\text{Some}', '\text{tweet}', ...\}$, the candidate the tweet is about is indentified as finding $i{:}\,th$ candidate that satisfies:

$$\max_i |T_i \cap \tau|$$

Where $|(\cdot)|$ denotes the cardinality operator, which counts the number of elements in $(\cdot)$. In the case when there are more than one candidate that satisfies the maximum cardinality, one simply take the candidate that occurred first in the tweet. As an example, consider the tweet:

$$\tau_1 = \{@\text{DonaldTrump}, \text{You}, \text{are}, \text{horrible}, \text{Hillary}, \text{for}, \text{president}\}$$

In this case Donald Trump and Hillary Clinton occurs equal amount of times in the tweet, and thus our algorithm will determine that this tweet is about Donald Trump, since his name occurred first in the tweet. The motivation behind this is that Twitter uses a mechanic known as the "@" symbol, which means in the example above translates to "Addressed to: Donald Trump". In this, it is thus always assumed that the first candidate that appears in a tweet is the person the tweet is most likely addressed to.

The sentiment score of a tweet is in loose terms defined as the quantified sentiment the tweet expresses about the candidate it is about. It is defined as:

$$Sentiment\ score := \frac{\mu(\tau)}{|\tau|}$$

Where $\mu(\cdot)$ is the measure defined in section (2.2.1). The score is normalised with the length of the tweet, $|\tau|$, in order to ensure that the sentiment score is bounded.

3.6 Data Treatment

As mentioned, one needs to treat the raw data before transforming it into vectorised form. Below the process is described. For a more exact review, please see the code attached in Appendix 5.

First, after data is collected, some common tags are generated, i.e. #candidate_i @candidate_i etc to the search queries used to create a larger and more covering identification set for each candidate. Secondly, the results of a primary is processed by transforming the data on the form:

Connecticut	***Votes***	***Percent***
Trump	518,601	58
Kasich	214,755	12
Cruz	123,894	28

Connecticut	Votes	Percent
Clinton	1,037,344	52
Sanders	752,739	46
Delaware	Votes	Percent
Trump	518,601	61
Kasich	214,755	16
Cruz	123,894	20
Delaware	Votes	Percent
Clinton	1,037,344	60
Sanders	752,739	39
Maryland	Votes	Percent
Trump	518,601	54
Kasich	214,755	19

Table 3: *Example data in original format.*

To the form:

State	***Trump***	***Cruz***	***Kasich***	***Clinton***	***Sanders***
Connecticut	58	28	12	52	46
Connecticut winner	1	0	0	1	0
Delaware	61	20	16	60	39
Delaware winner	1	0	0	1	0

Table 4: *Example data in processed format*

This new table in then used to generate dependant variables, which in this case is chosen to be the percentage result.

Lastly, the saved tweet from MongoDB are loaded, and extract the following variables directly:

- Timestamp
- Number of followers of the account that posted the tweet
- Number of friends of the account that posted the tweet
- Number of retweets of the posted tweets
- Number of times the tweet has been marked as a favourite
- User location

And the following data points are extracted from processing the actual text of the tweet using the scoring system and identification system described in the theory section:

- Sentiment score
- Dummy variable for the candidate associated with the tweet
- Dummy variable for the state the primary was held in
- Actual outcome of the primary
- Dummy variable for the winner of the primary

The state tags and dependent variables are generated directly from Table 4: *Example data in processed format*, and using the generated information about which candidate the tweet is describing.

3.7 Model Setup

In this section the variables used in the models in this study are presented.

3.7.1 Dependent variable

The dependent variable, Y, is the actual outcome from each primary election. This is a natural choice of dependent variable given the research question.

In our statistical learning approach, Y is chosen as a dummy for winner of each primary. This yields a binary logistic formulation.

3.7.2 Independent variable

The goal is to remove all insignificant or otherwise unsuitable independent variables to arrive at the model which most accurately explains the outcome of a primary election. The candidates for the independent variables, X, are presented below:

- Timestamp
- Number of followers of the account that posted the tweet
- Number of friends of the account that posted the tweet
- Number of retweets of the posted tweets
- Number of times the tweet has been marked as a favourite
- User location
- Sentiment score
- Dummy variable for the candidate associated with the tweet
- Dummy variable for the state the primary was held in
- Actual outcome of the primary
- Dummy variable for the winner of the primary

As only some of the above variables are deemed suitable for building forecasting models not all of the above will be included in further analysis. For example the dummy variable for the state the primary was held in is for obvious reasons not suitable in a forecasting model.

3.8 Model Variation

As to answer the research question a set of different models are proposed in order to find the model which can best predict the actual outcome. The full regression is referring to the model using the following variables:

Variable	*Description*
result	Dependent variable, percentage vote for each candidate
$candidate_i$	Dummy variable for candidate i
followers	Number of followers of the account that posted the tweet
$followers:candidate_i$	Interaction term
score	Sentiment score for the tweet
$score:candidate_i$	Interaction term
interaction	Sentiment score * followers, interaction term
$interaction:candidate_i$	Interaction term
Friends	The number of friends on Twitter the user have

Table 5: *List of all variables used in the regressions, including brief descriptions.*

Please note that in the logit regression the dependent variable has to be transformed. This transformation takes the form:

$$\tilde{y} = \log\left(\frac{result\ in\ percent}{1 - result\ in\ percent}\right)$$

where $\tilde{y}$ is the new dependent variable used in the logit regression.

In some sense, the models are presented in an increasing level of sophistication.

3.8.1 Model 1 – Simple Linear Regression

The first model proposed is called model 1. In model 1 first the full regression, i.e. with all covariates, is carried out. Then it is reduced, one covariate at the time, starting with the least significant one, until all covariates are significant at the pre-specified significance level $\alpha = 0.05$.

3.8.2 Model 2 – Stepwise Simple Linear Regression Based on AIC

In model 2 a stepwise regression based on finding the lowest AIC is carried out, starting with the full regression. An advanced built-in statistical algorithm in R is used to find the model with the lowest AIC.

3.8.3 Model 3 – Logit Regression

In model 3 a stepwise logit regression based on the AIC is carried out, starting with the full regression. Again, an advanced built-in statistical algorithm in R is used to find the model with the lowest AIC. The regression is carried out by employing the OLS estimator.

3.8.4 Model 4 – Inflated Zero Logit Model

As over half of the tweets in the sample are neutral it is somewhat warranted to test a model where observations with sentiment score equal to zero has been excluded. Even though one ideally would want to include all tweets this approach might provide some extra robustness. Except the use of a sub-sample the procedures for Model 4 and Model 3 are identical.

3.8.5 Model 5 – Statistical Learning

Here logit regression is used, but on a dummy of the winner in each primary. Besides this, the parameter λ is chosen to be 0.01. An important thing to note is that an additional, very brave, assumption on the data is made. More specifically it is assumed that online opinions on twitter vary very slowly and that data used post-election results can still be used to train the model, since it still represents the online opinion before the election.

Further, the GOP nomination is separated from the DNC nomination and thus two predictors are obtained. The reason for doing so is because this model does not include interaction effects and therefore does not capture the difference in effect of sentiment on the prediction results. In both cases, the training data and test data are sampled from two separate instances.

4. Results and Analysis

The results are presented, commented and briefly analysed. Further analysis and discussion is left to the Discussion and Conclusion chapter.

To provide the reader with intuition for the sample some graphs are presented below.

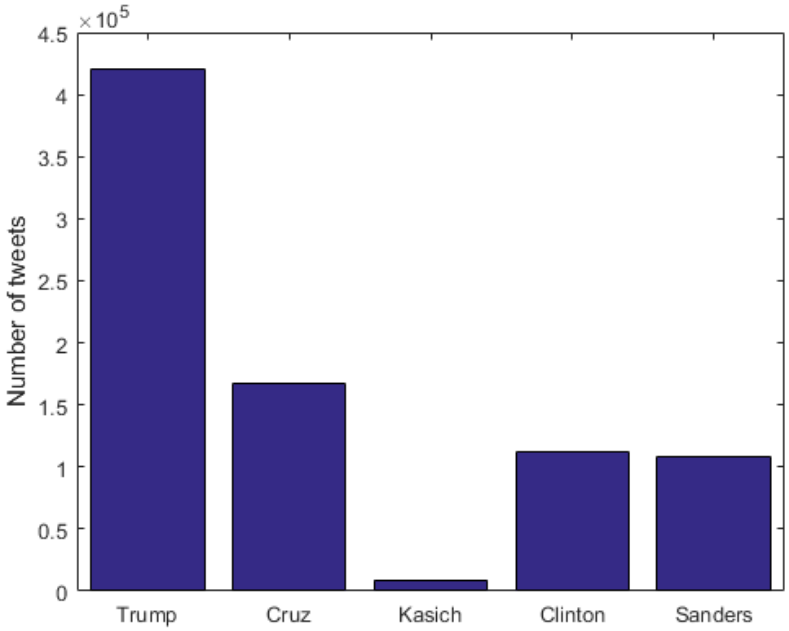

Figure 2: *Number of tweets collected per candidate. 816,265 tweets in total.*

Hardly surprising the controversial businessman Donald Trump is the candidate which gives rise to the highest amount of tweets. Further John Kasich is the one who, by far, gives rise to the least number of tweets.

As the most interesting variable included in this paper is the sentiment score its sample distribution is plotted below.

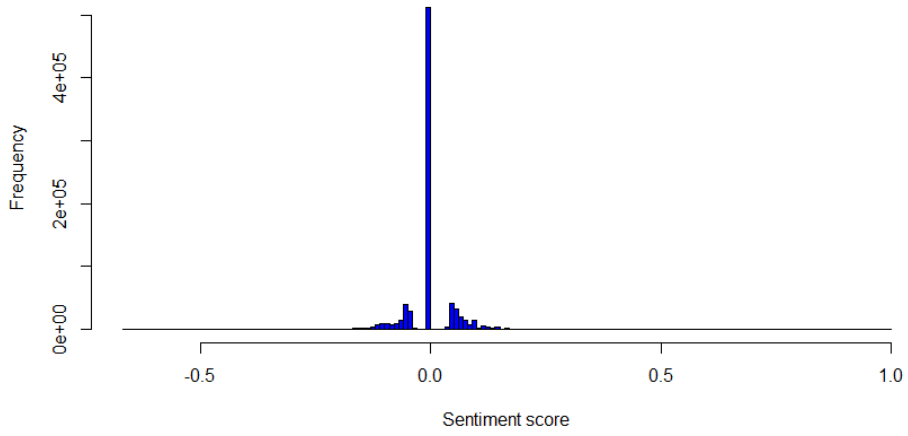

Figure 3: *Histogram for the sentiment score. As one see can from the plot the sample consists of a majority of neutral tweets.*

As over 500’000 of the approximately 800’000 tweets are regarded as neutral, i.e. sentiment score equal to zero, the plot of the distribution of non-zero sentiment scores might also be interesting and is thus presented below.

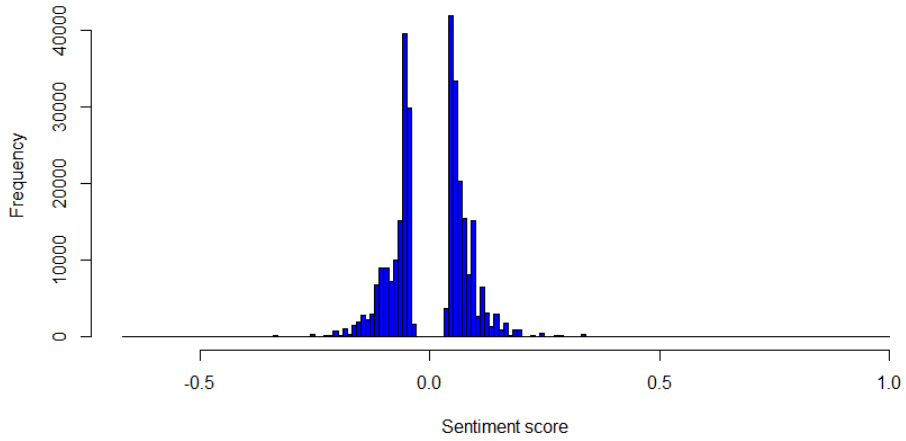

Figure 4: *Histogram for the sentiment score after excluding all neutral observations.*

One can observe a quite symmetric distribution with mean zero.

Employing a Breusch-Pagan tests for all three models render in p-values of less than 0.001. Thus White’s consistent estimator is used for the rest of this study. Further, to apply White’s estimator is always an advisable approach according to Lang (2014).

4.1 Model 1 – Simple Linear Regression

Model 1 is obtained by starting with the full regression and removing the least significant variable on-by-one until all variables are significant at the significance level $\alpha = 0.05$. The output from the final regression is presented below, for earlier steps please refer to Appendix 1.

	Estimate	*Std.Error*	*p-value*	
(Intercept)	48.995	0.053	< 2e-16	***
score	7.613	1.129	1.57e-11	***
trump	-13.132	0.060	< 2e-16	***
cruz	-8.841	0.069	< 2e-16	***
clinton	-3.691	0.076	< 2e-16	***
score:trump	-6.681	1.252	9.43e-08	***
score:cruz	-4.765	1.443	0.001	***
score:clinton	-4.490	1.562	0.004	**

Table 6: *Regression statistics for the final Model 1. Senator Bernie Sanders and Governor John Kasich are the benchmark for candidates. Please note that in the full regression only Senator Bernie Sanders is the benchmark for candidates. 816'265 observations,* R^2*=0. 0687.*

The first thing to note is that, somewhat surprisingly, the dummy for John Kasich is not significant. Neither are any of the variables containing followers, this may imply that the effect of a tweet is uncorrelated with the number of followers the person who posted tweet has. This is counterintuitive as one might expect that a tweet that reaches a wider audience, on average, has a larger impact. Moreover, as expected, a positive sentiment score implies a higher result, even though this effect varies for the different candidates.

It is also interesting to note that the candidate with the lowest *ceteris paribus* intercept, i.e. intercept plus dummy for the candidate, also has the lowest interaction for sentiment score and so on.

The coefficient of determination implies a quite poor fit since the model only explains around 7% of the variation in the data.

4.2 Model 2 – Stepwise Simple Linear Regression Based on AIC

Model 2 is obtained by starting with the full regression and employing a statistical algorithm to find the model with the lowest AIC. The output from the final regression is presented below, for earlier steps please refer to Appendix 2.

	Estimate	*Std.Error*	*p-value*	
(Intercept)	51.570	0.054	< 2e-16	***
followers	-4.53e-07	1.98e-07	0.022	*
trump	-15.700	0.061	< 2e-16	***
cruz	-11.410	0.070	< 2e-16	***
kasich	-32.960	0.195	< 2e-16	***
clinton	-6.259	0.076	< 2e-16	***
score	3.593	0.579	5.44e-10	***
followers:trump	4.68e-07	2.75e-07	0.089	.
trump:score	-2.661	0.785	0.001	***

Table 7: *Regression statistics for the final Model 2. Senator Bernie Sanders is the benchmark for candidates. 816'265 observations,* R^2*=0. 100.*

For Model 2 all variables except the interaction term between number of followers and Donald Trump is non-significant at the chosen significance level $\alpha = 0.05$. Further, the dummy for John Kasich is significant, as opposed to Model 1. And so is the variable for number of followers, although it comes with an unexpected sign. Moreover, the interaction variable for sentiment score is only present for Donald Trump, implying no difference with respect to sentiment score between e.g. Hillary Clinton and Ted Cruz.

The coefficient of determination, still, implies a quite poor fit since the model only explains around 10% of the variation in the data. However, the model exhibits a large improvement compared to Model 1.

4.3 Model 3 – Stepwise Logit Regression Based on AIC

Model 3 is obtained by starting with the full logit regression and employing a statistical algorithm to find the model with the lowest AIC. The output from the final regression is presented below, for earlier steps please refer to Appendix 3.

	Estimate	*Std.Error*	*p-value*	
(Intercept)	0.084	0.003	< 2e-16	***
followers	-2.12e-08	9.46e-09	0.024	*
trump	-0.791	0.003	< 2e-16	***
cruz	-0.573	0.003	< 2e-16	***
kasich	-1.601	0.009	< 2e-16	***
clinton	-0.315	0.004	< 2e-16	***
score	0.249	0.053	2.53e-06	***
followers:trump	2.28e-08	1.32e-08	0.082	.
trump:score	-0.207	0.059	4.33e-04	***
cruz:score	-0.110	0.068	0.105	
clinton:score	-0.117	0.073	0.111	

Table 8: *Regression statistics for the final Model 3. Senator Bernie Sanders is the benchmark for candidates. 816'265 observations,* R^2*=0. 109.*

In Model 3 the interaction term between number of followers and Donald Trump, the interaction term between sentiment score and Ted Cruz and Hillary Clinton are insignificant at the chosen significance level $\alpha = 0.05$. Again, the coefficient for followers comes with an unexpected sign. One can see that the pattern from Model 1, with lower interaction terms between sentiment score and candidate for lower stand-alone dummy for candidate, repeats itself.

The coefficient of determination, still, implies a quite poor fit since the model only explains around 11% of the variation in the data.

In Model 3 it is important to be very careful when interpreting the above estimated slope coefficients as the regression is a logit regression, i.e. it is non-linear. If the estimated slope coefficient, beta, is positive is implies a positive effect on the dependent variable, how much so depends on the values of the other covariates.

4.4 Model 4 – Inflated Zero Logit Model

Model 4 is obtained by starting with the full logit regression and employing a statistical algorithm to find the model with the lowest AIC. The output from the final regression is presented below, for earlier steps please refer to Appendix 4.

	Estimate	*Std.Error*	*p-value*	
(Intercept)	0.070	0.005	< 2e-16	***
followers	-1.82e-08	1.423e-08	0.200	
trump	-0.807	0.005	< 2e-16	***
cruz	-0.545	0.006	< 2e-16	***
kasich	-1.587	0.016	< 2e-16	***
clinton	-0.292	0.006	< 2e-16	***
score	0.269	0.054	5.28e-07	***
followers:trump	4.04e-08	1.97e-08	0.041	*
trump:score	-0.212	0.059	3.46e-04	***
cruz:score	-0.139	0.068	0.043	*
clinton:score	-0.136	0.074	0.067	.

Table 8: *Regression statistics for the final Model 4. Senator Bernie Sanders is the benchmark for candidates. 303'626 observations,* R^2*=0. 109.*

Yet again, the coefficient for followers comes with an unexpected sign. Further, the most interesting part in this model is that the estimated slope coefficient for score is positive, just as in the previous three models with "neutral" data included.

The coefficient of determination, still, implies a quite poor fit since the model only explains around 11% of the variation in the data.

In Model 4 it is important to be very careful when interpreting the above estimated slope coefficients as the regression is a logit regression, i.e. it is non-linear.

4.5 Model 5 – Statistical Learning

To present the results from the model based on statistical learning model would not make sense due to its dynamic nature. However, it is used to make prediction as showed in the last part of the results chapter.

4.6 Predictions

In order to answer the research question the models built are used to predict the outcome of the primary elections using more recent data. As the data is collected before the time point of multiple primaries across different states the results below should be interpreted as an overall result.

Candidate	*Model 1*	*Model 2*	*Model 3*	*Model 4*
Trump	35.9%	35.9%	33.0%	32.4%
Cruz	40.2%	40.2%	38.0%	38.3%
Kasich	49.0%	18.6%	18.0%	18.0%
Clinton	45.3%	45.3%	44.2%	44.4%
Sanders	49.0%	51.6%	52.1%	51.7%

Table 9: *Predictions made for each candidate and for the first four models.*

For Model 5 the overall predictive power for both models (GOP and DNC) are very plausible. For the republican model the accuracy is 100% on a test set and for the democratic model the accuracy is 80%. Accuracy in this case, is defined as number of correctly predicted points in the test set divided by the total number of points in the test set.

Interestingly the sum of predicted share of votes for the republican candidates in Model 1 sums up to more than 100%. This problem does not arise in the subsequent, more sophisticated, models.

It is important to note that in the first four models the results are predicted as percentage of votes each candidate will receive but in the statistical learning model one instead predict the winner and loser.

5. Discussion and Conclusion

In this chapter the results and analysis are discussed. The methods and data utilised throughout the study are critically evaluated. Further the conclusions are presented together with the contribution of the study and suggestions for future research.

Overall results

From the prediction results of the four first models, we see that the classical econometric approach is not suitable for prediction of primary results in the sense that the models predictions are wrong by a lot. On the other hand using a statistical learning approach and reducing the outcome to win or loss, we are much more successful and thus this suggests that the way forward might be this new area combining machine learning and statistics.

In our results we observe an R^2 of approximately 0.11 for the regression models. The direct interpretation of an R^2 of 0.11 is that 11% of the variance in data is explained by our models. This might not sound like a good result from an objective standpoint, but considering the large amount of data points used to yield this result, it is at the very least significant. In this aspect a significant explanation of 11% of the total variation indeed suggests that there is an explanatory power in twitter data if used in the correct context.

From the statistical learning approach, the results obtained are excellent, indeed suggesting that there is a strong predictive power in twitter data. Since we made assumptions on data with this method, it is debatable on how valid our results are.

From previous research it has been concluded that sentiment and volume of tweets related to a candidate are two main aspects that contribute to the outcome of election results. In the next section we analyse our models and their interpretation in the context and how sentiment and volume have played a part in yielding the results we have obtained.

Models

Model 1 – Simple Linear Regression
In the linear model we see that the best R^2 obtained is approximately 0.07. Since the final form of this model was obtained from gradual reduction with respect to significance of each covariate, we are at least assured that this model tells us which candidates are relevant on Twitter. In our case when candidate Bernie Sanders is taken as a reference, all other candidates seen to be at a disadvantage in terms of online popularity, except for candidate John Kasich, who turns out not even to be significantly different from Bernie Sanders on Twitter. This result is surprisingly in line with reality. When comparing each candidates total followers on Twitter,[5] John Kasich indeed have the lowest amount: 0.24 million, compared to second lowest Ted Cruz at 0.93 million. A possible conclusion from this observation is that a

[5] *Number of Twitter followers of 2016 U.S. presidential candidates, as of May 31, 2016 [Accessed 2016-05-02]*

candidate that is not heard or seen on Twitter, is neither talked about nor mentioned on Twitter.

In terms of online popularity, we observe that there indeed is a positive correlation between sentiment of the tweets and election results. Given that we made a very naïve estimate of sentiment, this suggest that sentiment indeed plays a large role in determining a candidate's favourability from Twitter. In our case, we simply counted the positive and negative words and assigned a favourability score based on numbers of words. Clearly, a more sophisticated method is needed to account for sarcastic remarks and such. But if we already can establish that there is a strong correlation with good significance with a naïve method, there are reasons to believe that a candidate's popularity indeed has something to do with how the candidate is perceived on Twitter.

Looking at the dummy variables for each candidate, we see that almost all, except Kasich, are significant, which is in line with previous research in the sense that this implicitly reflects popularity based on number of tweets related to the candidate. A very important remark on this aspect is to observe the age group bias in the context. Clearly, there are more young adults and teenagers on Twitter compared to senior citizens.

Model 2 – Stepwise Simple Linear Regression Based on AIC
Here, we observe that the sentiment is still significant, implying that is efficient in explaining variance further strengthening our suspicion that sentiment plays an important role in predicting election results.

Additionally, we see that the number of followers is an efficient but less significant explanatory variable. This is very interesting in the sense that a tweet with a lot of followers tends to have a larger impact on the Twitter since a larger population will see and react to the tweet.

Overall, we see that the R^2 has improved by a lot, compared to Model 1, which implies that there indeed are prediction capabilities using Twitter data with the correctly selected model.

Model 3 – Stepwise Logit Regression Based on AIC
Since we are regressing on election results based on fractions of total votes, it would indeed make sense to interpret this as a probability and use a logistic estimation. In this model we see that R^2 has increased again yielding our best estimate so far.

Here we see that the sentiment is still significant suggesting that there exists a correlation between primaries outcome and sentiment. Further, we observe that more covariates are included in logit estimation, which may be due to a better model specification.

Model 4 – Inflated Zero Logit Model
In this regression, we observe that sentiment is still significant and that it has increased in impact. Since the results from the logit model with sentiment scores equal to zero removed are alike the results from this model, it can be inferred that the results from Model 3 are robust.

Model 5 – Statistical Learning
Using the statistical learning approach, our model has a regularization constant of 0.01. The choice of this constant is discussable in the sense that perhaps a larger or smaller chosen constant would yield even better results, in terms of presenting a better regularization of weights. Unarguably, the model approach of using a logit regression is yet again proven to be a successful concept from the surprisingly good prediction results, whose validity we will discuss later in this section.

Covariates

Candidate dummies
In every model, we see that the candidate dummies are significant for every candidate in every model except for candidate John Kasich, who is non-significant in Model 1. Moreover, it is important to note that Bernie Sanders is the benchmark candidate.

The results regarding the candidates are surprising, in the sense that Donald Trump is expected to be less favoured then Ted Cruz with regards to the number of tweets related to each respective candidate. This may be a methodological issue, which we will discuss in a later section.

Sentiment score
This is the most interesting result, in the sense that it is significant in every regression and that it is measured through a rather naïve method.

The combined covariates – interaction terms
From the combined we observe an interesting result in the sense that only Hillary Clinton get lower scores when combined with the dummy variable and the "score" variable. One interpretation of this is that Hillary actually gets more popular based on how much she is disliked, which seems more like a sample bias.

Method

One reason why Donald Trump is expected to be less favoured compared to Ted Cruz in winning according to tweets about each candidate may come from that the results are non-proportional in distribution. In this context, we mean that when we assign each tweet a result of a pre-election, a large portion of tweets about Trump was assigned when he lost big in Utah to Ted Cruz, explaining why there is a bias towards Cruz.

Another aspect about the regression is that we have not separated the democratic election and the republican election. This implies that when using Bernie Sanders as a reference point in regressing, republican candidates tend to be biased towards lower scores since there are not equally many candidates running in each party, implying that each candidate tend to get a smaller share of votes in the republican race since there are three candidates competing rather than two. However, this should be accounted for by including the interaction terms.

The different primary mechanics are something we did not account for in this essay. By this, we mean that in open primaries for instance, everyone is allowed to vote, which essentially makes "vote swaying" a lot more important for both parties and candidates. In this sense we argue there might be unaccounted bias in data we have not measured nor mitigated. On the same side, there is also the issue of closed primaries, which only allows registered voters. In this scenario we argue that Twitter data might be unreliable in the sense that the voters already have decided who to vote for making Twitter unrepresentative.

Data

An important issue about the data is that there is a population bias in Twitter data as previously mentioned. The implications of this is that we are not guaranteed that the data we are using is representable for the entire population. In this sense, it would be suitable to estimate how much of the voting population actually are young voters and investigate if this group is actually representable for the general population. This is something that Choy et al. (2012) accounted for in their paper using a consensus corrected Twitter model. Although, it is indeed efficient, we use a slightly different approach in this paper argue that it suffices to make multiplicative correction on the prediction for either a downward bias or upward bias depending on voter decomposition.

Additionally, we indeed assume that there is an autocorrelation in data and thus adjust our estimates to account for this in the sense that we only use data collected before a primary election to conduct our estimates. This might not be a necessary procedure in the sense that even if the opinion of a population changes, we argue that Twitter represents only the entire online population which has lesser direct impact on results rather than if the opinion in a state changed.

Another issue about data is the unobservable stigma for certain candidates that might be present. We argue that candidates like Donald Trump, who has publically announced less "politically correct" assertions have a hidden voter base unwilling to publically support him, in fear of the social backlashes that might occur, i.e. being fired from a job, losing friends etc. This stigma is something that we can neither measure nor observe and thus does not properly account for.

On the same note, there has also been reports that Twitter Inc. itself is biased in censoring and banning its content.[6] Under these assumptions, there is even a stronger bias against certain candidates who are known to be "outsiders".

An interesting effect that has been discussed in previous papers are the "incumbent effect", which is when a president is running for a second term and is thus in a more favourable position with regards to popularity compared to opponents. In our case this might be an issue when it comes to the candidate Hillary Clinton, who is the secretary of the state in the current government. In this sense, this effect might be hard to find or quantify through twitter data.

---

[6] *Twitter's new 'Safety Council' makes a mockery of free speech [Accessed 2016-05-03, Author: Brendan O'Neill]*

In the statistical learning case, we choose a training set a lot larger than the test set. This might yield the test inaccurate, since the test set might have been biased due to its smaller size.

Assumptions

As mentioned before, our algorithm does not account for irony. We argue that for a large amount of tweets, a naïve approach is a sufficient approximation of sentiment in the sense that if a tweet is sarcastically negative, it is very likely that there is a tweet that is sarcastically positive. We argue that in large amounts of data these tweets take each other out and ultimately gives a representative sentiment regarding a candidate.

When we trained our statistical learning model, we made the assumption that twitter sentiment varies very slowly with time and thus data post up to a week post primary election is still valid when it comes to training a model. This assumption may have induced a forecast bias, which we have not accounted for.

Further contribution

The aim of this study is to investigate how well a rather unorthodox method of conducting econometric research works. From the results we see that there is still room for improvement in terms of data processing and filtering. Despite a naïve approach to sentiment analysis we see that the approach still has some potential with large data sets, since significance is established for a non-zero slope coefficient.

In contrast to traditional econometrics which is studied on historical data sets, we have applied econometric concepts to big data sets of more dynamic nature, which we believe will become a larger interest of study in conjunction with the digitalization of society and evolution of social media. The dynamic aspect is reflected in the sense that the opinions of people on Twitter might change with certain news being released about each candidate, meaning that the obtained data for each pre-election might not be representative for the next pre-election.

We argue that the Twitter sentiment obtained in this paper is of value as a component in a future model, as it seems that sentiment indeed is positively correlated with election outcomes. This conjecture makes it plausible to investigate further in more advanced sentiment analysis accounting for more advanced linguistics used on Twitter.

Further, we find that more refined methods of statistical learning are better suited for building models from huge sets of social media, compared to the more traditional and simple approached commonly used within econometrics. A secondary, implicit, aim of this study is to serve as a reference or go-to-guide for econometricians who are interested in starting to use statistical learning and sentiment analysis to analyse large data sets for popular topics such as

stock price movement and policy evaluation. At a first glance the methods used can seem very technical but much of the advanced coding needed is related to data extraction, i.e. if one already has a data set much of the procedures described can be skipped. Even though this paper might not serve as an exhaustive source it certainly provides the basic intuition needed for a first attempt at statistical learning and sentiment analysis.

Looking back at what this could imply for the study of prediction markets, we conclude that there is indeed a possibility to use Twitter data to predict at least binary outcomes of various political occurrences such as elections, given that the historical data are provided of outcomes and covariates. In our case, the prediction occurs by taking the mean of several tweets to obtain an average predictor.

Future research

From the above results, we discover that simplified methods of sentiment analysis still yields viable results in terms of prediction and inferences made. A topic for further research would thus be to use more advanced techniques within the field of natural language processing in order to obtain more accurate estimates of sentiment.

We discover that statistical learning indeed has a lot of potential in predictive modelling based on data from streaming sources. A further topic to research is to see which other potential areas within econometrics, such as macro economy, micro economy or behavioural economics also could benefit from these methods.

References

Bermingham, A., & Smeaton, A. F. (2011). On using Twitter to monitor political sentiment and predict election results. Sentiment Analysis where AI meets Psychology (SAAIP) Workshop at the International Joint Conference for Natural Language Processing (IJCNLP).

Burnham, K. P., & Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.). New York: Springer-Verlag.

Choy, M., Cheon, M., Nang Laik, M., & Ping Shung, K. (2011). A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. Retrieved from http://arxiv.org/abs/1108.5520

Choy, M., Cheon, M., Nang Laik, M., & Ping Shung, K. (2012). US Presidential Election 2012 Prediction using Census Corrected Twitter Model, Cornell University Library, version 3, 11 Nov 2012, Retrieved from http://arxiv.org/abs/1211.0938

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). New York: Lawrence Erlbaum Associates.

Edwards, J., Leaked Twitter API data shows the number of tweets is in serious decline, Business Insider, 2 February 2016, http://uk.businessinsider.com/tweets-on-twitter-is-in-serious-decline-2016-2, (accessed April 21st 2016)

El Darwiche, B et al., Digitization for economic growth and job creation: Regional and industry perspectives, Strategy&, 2013, http://www.strategyand.pwc.com/reports/digitization-economic-growth-job-creation, (accessed April 21st 2016)

Friedman, J., Hastie, T., & Tibshirani, R. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag.

Gayo-Avello, D. (2012). I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper, Cornell University Library, version 1, 28 April 2012, Retrieved from http://arxiv.org/abs/1204.6441

Hall, R. E., & Lillien, D. M. (1995). EViews User Guide. Irvine, CA: Quantitative Micro Software.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). Applied Linear Regression Models (4th ed.). Irwin, CA: McGraw-Hill.

Lang, H. (2014). Elements of Regression Analysis. Stockholm: KTH Department of Mathematics.

Liu, B. Professor Bing Liu's personal website, [website], 2016, https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon, (accessed 15 February 2016).

Manyika, J et al., Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, May 2011, http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation, (accessed April 21st 2016)

Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011). How (Not) to Predict Elections. 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 165-171.

Newbold, P., Carlson, W., & Thorne, B. (2010). Statistics for Business and Economics (7 ed.). New York: Pearson Education Inc.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 122-129.

Theil, H. (1961). Economic Forecasts and Policy. Amsterdam: North-Holland Pub. Co.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 178-185.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica, 48, 817-838.

Wooldridge, J. M. (2008). Introductory Econometrics: A Modern Approach 4th Edition. East Lansing, MI: South-Western College Pub.

Web Sources

http://www.realclearpolitics.com/ [Accessed May 1st 2016]

http://stats.seandolinar.com/collecting-twitter-data-storing-tweets-in-mongodb/ [Accessed May 2nd 2016, Author: Sean Dolinar, A blogg post about MongoDB and storing tweets, last updated January 29 2015]

https://weiwutao.wordpress.com/2015/01/22/how-to-build-apache-spark-on-windows-8/ [Accessed May 2nd 2016, Author: Julius, A blogg post about how to build Apache Spark on windows 8, last updated January 22 2015]

http://www.statista.com/statistics/509579/twitter-followers-of-2016-us-presidential-candidates/ [Accessed May 2nd 2016, last updated May 31 2016]

http://spark.apache.org/docs/latest/mllib-linear-methods.html [Accessed May 3rd 2016, last updated Mars 6 2016]

https://github.com/MrHuff/twitterCrawler [Accessed May 7th 2016, Author: Robert Hu, One of the authors of this essay's personal GitHub page with code used for the project]

Appendix 1

Below is the full regression for Model 1. Please note that this is also the full regression in Model 2.

	Estimate	Std.Error	p-value	
(Intercept)	51.560	0.054	< 2e-16	***
trump	-15.700	0.061	< 2e-16	***
cruz	-11.400	0.070	< 2e-16	***
kasich	-32.950	0.195	< 2e-16	***
clinton	-6.256	0.076	< 2e-16	***
followers	6.01e-7	3.57e-7	0.091	.
score	5.598	1.155	1.24e-6	***
interaction	3.95e-7	5.29e-6	0.940	
trump:followers	6.50e-7	4.11e-7	0.114	
cruz:followers	-1.38e-7	5.43e-7	0.799	
kasich:followers	6.57e-8	1.90e-6	0.972	
clinton:followers	4.91e-7	4.95e-7	0.321	
trump:score	-4.675	1.271	2.34e-6	***
cruz:score	-2.817	1.454	0.053	.
kasich:score	-4.152	4.232	0.327	
clinton:score	-2.490	1.569	0.112	
trump:interaction	1.07e-6	6.21e-6	0.863	
cruz:interaction	9.73e-6	9.86e-6	0.324	
kasich:interaction	1.49e-6	2.22e-5	0.946	
clinton:interaction	2.11e-6	8.17e-6	0.797	

Table 10: *Regression statistics for the first regression of Model 1. Senator Bernie Sanders is the benchmark for candidates. 816'265 observations,* R^2*=0. 0100.*

The variables are then reduced in the following order:

1. kasich:followers
2. kasich:interaction
3. interaction
4. cruz:followers
5. clinton:interaction
6. trump:interaction
7. kasich:score
8. cruz:interaction
9. clinton:followers
10. followers

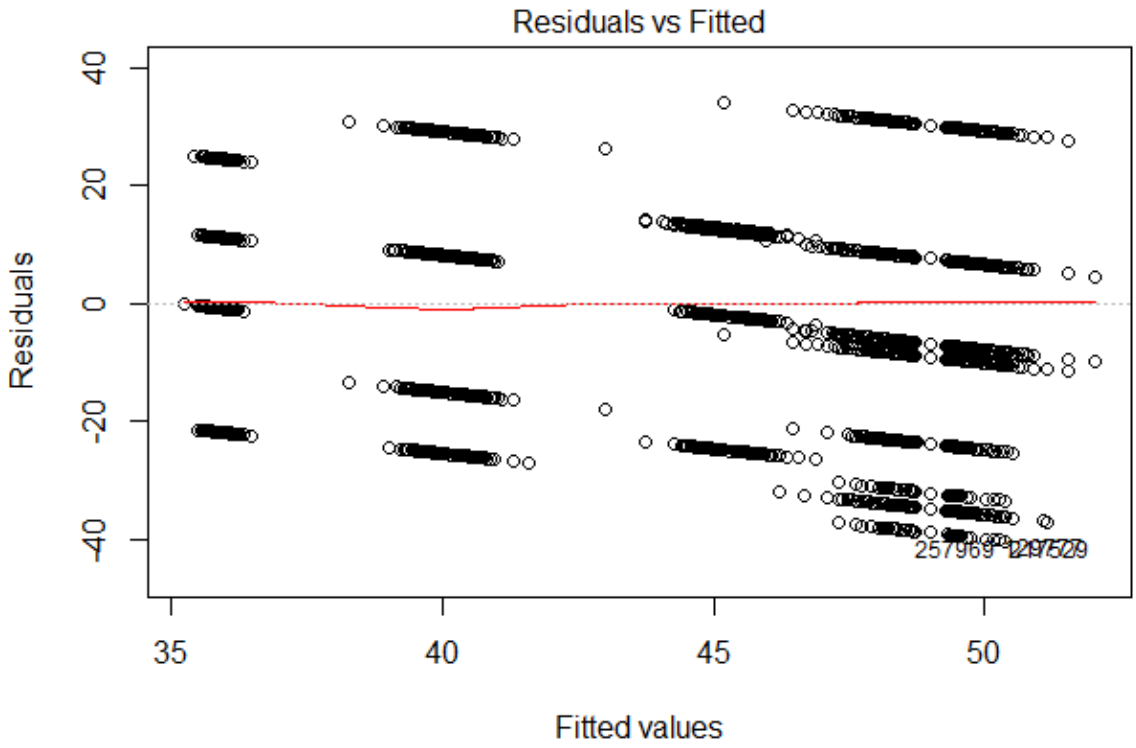

Figure 5: *Fitted values versus residuals for final Model 1.*

Appendix 2

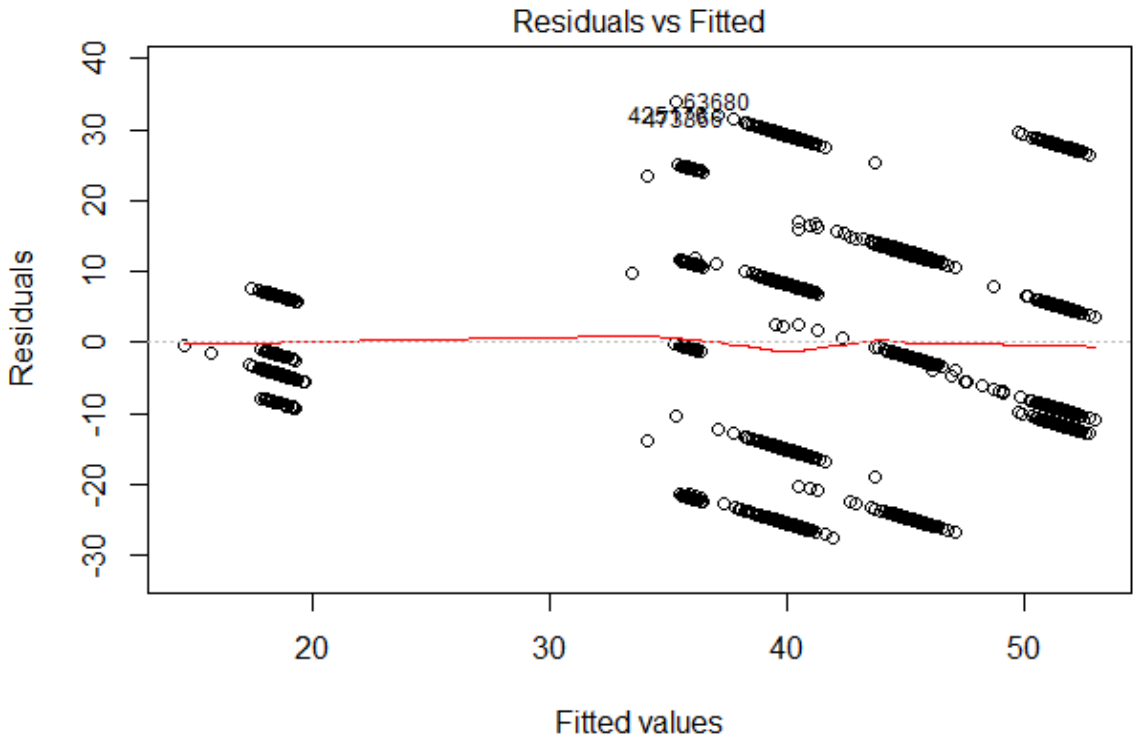

Figure 6: *Fitted values versus residuals for final Model 2.*

Below the steps taken to reach the final model are showed. Please note that in R the covariates satisfies the following property: $x_1 * x_2 = x_1 + x_2 + x_1 : x_2$ where : denotes scalar multiplication.

Starting point
result ~ followers * (trump + cruz + kasich + clinton) + score * (trump + cruz + kasich + clinton) + interaction * (trump + cruz + kasich + clinton)

Step 1: AIC=4706502
result ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:cruz + followers:clinton + trump:score + cruz:score + kasich:score + clinton:score + trump:interaction + cruz:interaction + kasich:interaction + clinton:interaction

Step 2: AIC=4706500
result ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:cruz + followers:clinton + trump:score + cruz:score + kasich:score + clinton:score + trump:interaction + cruz:interaction + clinton:interaction

Step 3: AIC=4706498
result ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:cruz + followers:clinton + trump:score + cruz:score + kasich:score + clinton:score + cruz:interaction + clinton:interaction

Step 4: AIC=4706496
result ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:cruz + followers:clinton + trump:score + cruz:score + kasich:score + clinton:score + cruz:interaction

Step 5: AIC=4706494
result ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:clinton + trump:score + cruz:score + kasich:score + clinton:score + cruz:interaction

Step 6: AIC=4706493
result ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:clinton + trump:score + cruz:score + clinton:score + cruz:interaction

Step 7: AIC=4706492
result ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:clinton + trump:score + cruz:score + clinton:score

Step 8: AIC=4706491
result ~ followers + trump + cruz + kasich + clinton + score + followers:trump + followers:clinton + trump:score + cruz:score + clinton:score

Step 9: AIC=4706490
result ~ followers + trump + cruz + kasich + clinton + score + followers:trump + trump:score + cruz:score + clinton:score

Step 10: AIC=4706490
result ~ followers + trump + cruz + kasich + clinton + score + followers:trump + trump:score + cruz:score

Step 11 (Final): AIC=4706489
result ~ followers + trump + cruz + kasich + clinton + score + followers:trump + trump:score

Appendix 3

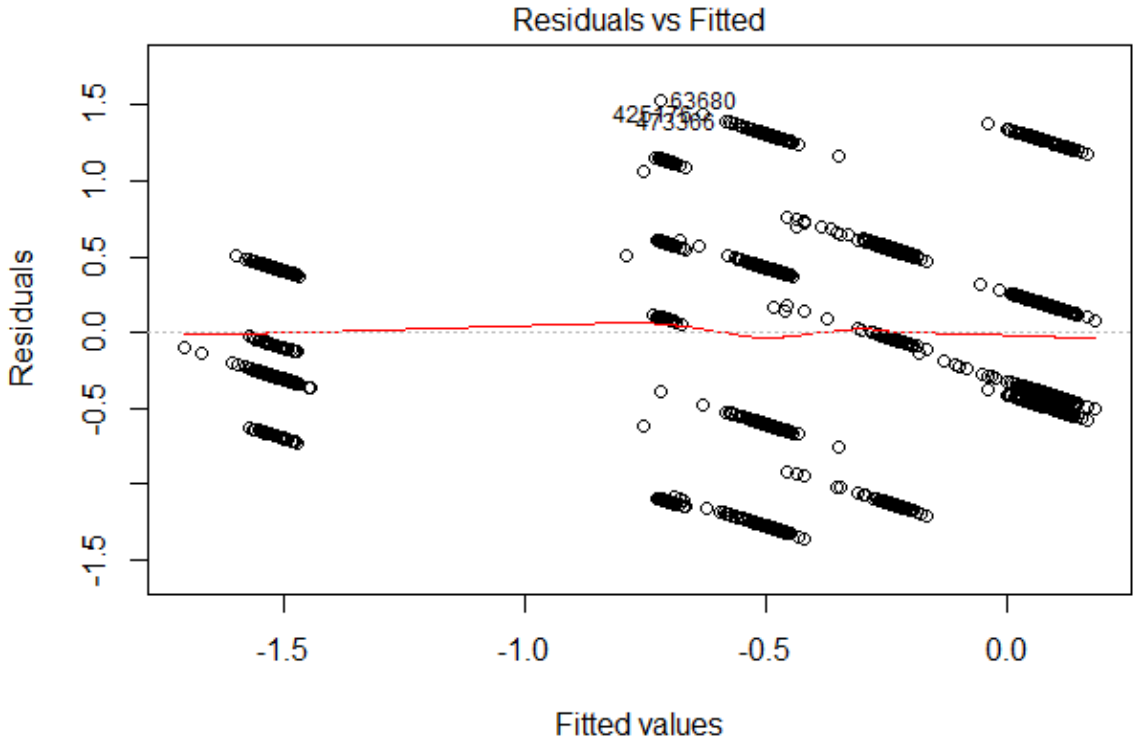

Figure 7: *Fitted values versus residuals for final Model 3.*

Below the steps taken to reach the final model are showed. Please note that in R the covariates satisfies the following property: $x_1 * x_2 = x_1 + x_2 + x_1 : x_2$ where : denotes scalar multiplication.

Starting point: AIC=-260472.3
logity ~ followers * (trump + cruz + kasich + clinton) + score * (trump + cruz + kasich + clinton) + interaction * (trump + cruz + kasich + clinton)

Step 1: AIC=-260474.3
logity ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:cruz + followers:clinton + trump:score + cruz:score + kasich:score + clinton:score + trump:interaction + cruz:interaction + kasich:interaction + clinton:interaction

Step 2: AIC=-260476.3
logity ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:cruz + followers:clinton + trump:score + cruz:score + kasich:score + clinton:score + cruz:interaction + kasich:interaction + clinton:interaction

Step 3: AIC=-260478.3
logity ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:cruz + followers:clinton + trump:score + cruz:score + kasich:score + clinton:score + cruz:interaction + clinton:interaction

Step 4: AIC=-260480.2
logity ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:cruz + followers:clinton + trump:score + cruz:score + kasich:score + clinton:score + cruz:interaction

Step 5: AIC=-260482.1
logity ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:clinton + trump:score + cruz:score + kasich:score + clinton:score + cruz:interaction

Step 6: AIC=-260483.4
logity ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:clinton + trump:score + cruz:score + clinton:score + cruz:interaction

Step 7: AIC=-260484.3
logity ~ followers + trump + cruz + kasich + clinton + score + interaction + followers:trump + followers:clinton + trump:score + cruz:score + clinton:score

Step 8: AIC=-260485.5
logity ~ followers + trump + cruz + kasich + clinton + score + followers:trump + followers:clinton + trump:score + cruz:score + clinton:score

Step 9 (Final): AIC=-260486.1 logity ~ followers + trump + cruz + kasich + clinton + score + followers:trump + trump:score + cruz:score + clinton:score

Appendix 4

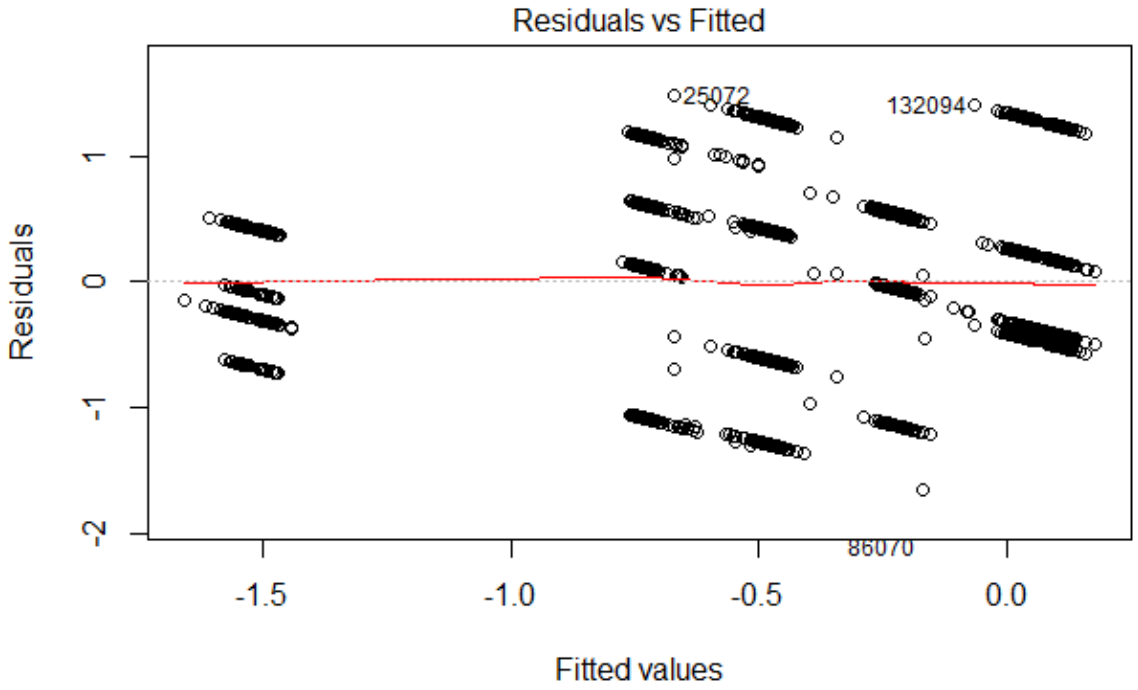

Figure 8: *Fitted values versus residuals for final Model 4.*

Below the steps taken to reach the final model are showed. Please note that in R the covariates satisfies the following property: $x_1 * x_2 = x_1 + x_2 + x_1 : x_2$ where : denotes scalar multiplication. Here

Starting point: AIC=-93038.72
logity ~ followers_z * (trump_z + cruz_z + kasich_z + clinton_z) + score_z * (trump_z + cruz_z + kasich_z + clinton_z) + interaction_z * (trump_z + cruz_z + kasich_z + clinton_z)

Step 1: AIC=-93040.71
logity ~ followers_z + trump_z + cruz_z + kasich_z + clinton_z + score_z + interaction_z + followers_z:trump_z + followers_z:kasich_z + followers_z:clinton_z + trump_z:score_z + cruz_z:score_z + kasich_z:score_z + clinton_z:score_z + trump_z:interaction_z + cruz_z:interaction_z + kasich_z:interaction_z + clinton_z:interaction_z

Step 2: AIC=-93042.59
logity ~ followers_z + trump_z + cruz_z + kasich_z + clinton_z + score_z + interaction_z + followers_z:trump_z + followers_z:kasich_z + followers_z:clinton_z + trump_z:score_z + cruz_z:score_z + kasich_z:score_z + clinton_z:score_z + trump_z:interaction_z + cruz_z:interaction_z + kasich_z:interaction_z

Step 3: AIC=-93044.38
logity ~ followers_z + trump_z + cruz_z + kasich_z + clinton_z + score_z + interaction_z + followers_z:trump_z + followers_z:kasich_z + followers_z:clinton_z + trump_z:score_z + cruz_z:score_z + kasich_z:score_z + clinton_z:score_z + trump_z:interaction_z + cruz_z:interaction_z

Step 4: AIC=-93046.34
logity ~ followers_z + trump_z + cruz_z + kasich_z + clinton_z + score_z + interaction_z + followers_z:trump_z + followers_z:clinton_z + trump_z:score_z + cruz_z:score_z + kasich_z:score_z + clinton_z:score_z + trump_z:interaction_z + cruz_z:interaction_z

Step 5: AIC=-93048.1
logity ~ followers_z + trump_z + cruz_z + kasich_z + clinton_z + score_z + interaction_z + followers_z:trump_z + followers_z:clinton_z + trump_z:score_z + cruz_z:score_z + kasich_z:score_z + clinton_z:score_z + cruz_z:interaction_z

Step 6: AIC=-93049.43
logity ~ followers_z + trump_z + cruz_z + kasich_z + clinton_z + score_z + interaction_z + followers_z:trump_z + followers_z:clinton_z + trump_z:score_z + cruz_z:score_z + kasich_z:score_z + clinton_z:score_z

Step 7: AIC=-93050.58
logity ~ followers_z + trump_z + cruz_z + kasich_z + clinton_z + score_z + interaction_z + followers_z:trump_z + followers_z:clinton_z + trump_z:score_z + cruz_z:score_z + clinton_z:score_z

Step 8: AIC=-93050.91
logity ~ followers_z + trump_z + cruz_z + kasich_z + clinton_z + score_z + interaction_z + followers_z:trump_z + trump_z:score_z + cruz_z:score_z + clinton_z:score_z

Step 9 (Final): AIC=-93051.33
logity ~ followers_z + trump_z + cruz_z + kasich_z + clinton_z + score_z + followers_z:trump_z + trump_z:score_z + cruz_z:score_z + clinton_z:score_z

Appendix 5

Proof of signed measure in (2.2.1).

Let $\tau \subset \sigma(\Omega)$. Then:

- If $\tau = \emptyset \rightarrow \mu(\emptyset) = |\mathcal{G} \cap \emptyset| - |\mathcal{B} \cap \emptyset| = 0$
- Since Ω is finite in our case, the second assumption holds.
- $\{E_i\}$ is a sequence of disjoint sets in $\sigma(\Omega)$ we have that:

$$\mu(\cup_{i=1}^{\infty} E_i) = |\mathcal{G} \cap \cup_{i=1}^{\infty} E_i| + |\mathcal{B} \cap \cup_{i=1}^{\infty} E_i| = \sum_{i=1}^{\infty} |\mathcal{G} \cap E_i| + |\mathcal{B} \cap E_i| = \sum_{i=1}^{\infty} \mu(E_i)$$

■