UNIVERSITY OF HAWAII AT MANOA

Examining the Impact of the Internet News Environment on the Spread of Articles on Twitter

by

Michael Rodriguez

A thesis submitted to the graduate division of Univserity of Hawaii at Manoa in partial fulfillment for the degree of

Masters in Science

in
Electrical Engineering

December 2019

Thesis Committee:

June Zhang, Chairperson
Narayana P. Santhanam
Lee Altenberg

Declaration of Authorship

I, Michael Rodriguez, declare that this thesis titled, 'Examining the Impact of the Internet News Environment on the Spread of Articles on Twitter' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Der Unterschied zwischen Vergangenheit, Gegenwart und Zukunft ist für uns Wissenschaftler eine Illusion, wenn auch eine hartnäckige"

Albert Einstein

UNIVERSITY OF HAWAII AT MANOA

Abstract

Dr. June Zhang, Chairperson
Electrical Engineering

Masters of Science

by Michael Rodriguez

Time-Series prediction problems have proven to be a challenge in nearly every domain, especially domains that do not necessarily exhibit periodic trends, or may be influenced by outside actions. One such domain is the activity of users on a social network. The prediction task has been studied in the past, where methods included; individual article features and short term activity [1], social media activity alone [2] and probabilistic methods based on early social media activity [3][4]. In this work we will present a modification to the previous methods, by including previous periods frequency count of specific feature as a feature for prediction of spread of an article on Twitter. It is based on the biological theory of frequency-dependent selection. An empirical study of the impact of the features was conducted, and there appears to be no noticeable improvement when we consider the frequency of article features noted on the Internet to the prediction of the spread of those articles on Twitter. This suggests a weak connection between the two networks. The modification may work better if the frequency of articles on Twitter itself were considered.

Acknowledgements

I'd like to thank my Advisor, Dr.June Zhang, and my thesis committee, Dr. Narayana Santhnam and Dr. Lee Altenberg for their guidance through the process. Id like to thank Rudy for developing the base program for collecting Twitter data. Id like to thanks Chris for working with the data handling process. Id like to thank Honggen for working with me during the last few months of the project. Id also like to thank everyone in Pearl City, San Diego and San Antonio who helped make the pursuit of this degree possible. Id also like to thank everyone who was a sounding board during this process, there's too many of you to name personally, but know that I appreciate the time you gave.

Contents

List of Figures

List of Tables

Abbreviations

MLP	**M**ulti **L**ayer **P**erceptron
BNN	**B**ayesian **N**eural **N**etwork
RBFNN	**R**adial **B**ased **F**unction **N**eural **N**etwork
GRNN	**G**eneral **R**esgression **N**eural **N**etwork
K-NN	**K** **N**earest **N**eighbors
SVR	**S**upport **V**ector **R**egression
GP	**G**aussian **P**rocess
ARIMA	**A**uto **R**egressive **I**ntegrated **M**oving **A**verage
CSTS	**C**ross **S**ectional **T**ime **S**eries
MARS	**M**ultivariate **A**daptive **R**egression **S**plines
RF	**R**andom **F**orest
SVM	**S**upport **V**ector **M**achine
VARMA	**V**ector **A**uto **R**egressive **M**oving **A**verage
RNN	**R**ecurrent **N**eural **N**etwork
ML	**M**achine **L**earning
LR	**L**ogistic **R**egression
ANN	**A**rtificial **N**eural **N**etwork
MTS-GCNN	**M**ultivariate **T**ime **S**eries **G**roupwise **C**onvolutional **N**eural **N**etwork
NLP	**N**atural **L**anguage **P**rocessing
CBOW	**C**ontinuous **B**ag **O**f **W**ords
F-K	**F**lesch **K**incaid
D-C	**D**ale **C**hall
SMOG	**S**imple **M**easure **O**f **G**obbledygook
API	**A**pplication **P**rogram **I**nterface
URL	**U**niform **R**esource **L**ocator

MMSE	**M**inimum **M**ean **S**quare **E**rror
BRR	**B**ayesian **R**idge **R**egression
LMS	**L**east **M**ean **S**quares
MSE	**M**ean **S**quare **E**rror
MAE	**M**ean **A**verage **E**rror
R^2	**R** **S**quared
OLS	**O**rdinary **L**east **S**quares
JSON	**J**ava **S**cript **O**bject **N**otation
GI	**G**eneral **I**nquirer
JS	**J**ussie **S**mollett
NZ	**N**ew **Z**ealand
MR	**M**ueller **R**eport
NH	**N**ipsey **H**ussle
RO	**R**epresentative **O**mar

This work is dedicated to the memory of mi abuela, Cecilia Rodriguez.

Chapter 1

Introduction

1.1 Introduction

A 2018 Pew Research Center study reported that traditional mass media sources, such as print and television, have seen a decline as a source where people often get their news. News websites and social media, on the other hand, have seen a growth in this area [5]. Social media has a unique influence on the traditional mass media communication model. Traditional models of mass communication are built on general models of communication, where there is a transmitter, message, transmission channel, receiver and feedback. Limitations to the mass communication models include a low or time-delayed feedback. The Internet solved the feedback time delay issue, allowing mass communication broadcasters the ability to receive feedback from the audience in real-time. This also allowed the broadcasters to quickly respond to feedback and incorporate changes to their messages. Social media added in an extra layer of interaction to the traditional broadcaster and the audience. The audience now can take a more active role in broadcasting of messages, from acting in place of a news editor, selectively choosing what articles to share on their transmission channel, to becoming content creators, distributing their own messages along the social media transmission network.

Technological advances have increased the ability to measure the rate at which user engage with news compared to those metrics for traditional news media. One challenge in measuring the rates from online sources is availability of data to the general public. Tracking website propagation is limited with respect to free, computationally simple approaches for non-domain owners. Social media has emerged as a great resource to not only count the amount of instances of a news article on the network, but to also track individual users interactions with websites, including sharing, liking or commenting on

posts. In particular Twitter is a popular platform for researchers to monitor user activity. It boasts over 326 million active users per month, 500 million tweets per day, and 80% mobile users [6].

Figure (1.1) is an example of a Tweet and figure (1.2) is a Retweet both containing a link to an article. Both were taken from the CNN Breaking News account. The annotations mark particular features of interest for this study. The user is indicated by @cnnbrk on the left and @CNNPolitics on the right. In figure (1.2) you can see CNN Breaking News Retweeted indicating that the Tweet was retweeted by @cnnbrk from @CNN politics. The links to the articles were the highlighted portion of cnn.it/2nslZoC, and cnn.it/2mWwhgz, respectively. On the bottom of each of the Tweets is a highlighted number, 84, and 80, respectively, which indicate how many times that microblog has been Retweeted. The other highlighted number 289 and 237, respectively annotate the times the Tweet was favorited.

FIGURE 1.1: Tweet from CNN

1.2 Statement of Problem

This thesis will propose a modification to current methods for predicting the spread of news on Twitter. It will consider the frequency of article features on the Internet as a

FIGURE 1.2: Retweet from CNN

feature to predict the spread of those article on Twitter.

We define an instance of data as a post by a user on a Twitter about a specific topic. The article originates on the Internet, as a result of some outside actor. That article then can be transferred to Twitter, via a decision by Twitter User. Alternatively another Twitter User can create Tweet from a previous Tweet. That is known as a Retweet. This requires an original Tweet from another Twitter User.

We can consider three different feature sets, $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{Z}$, where $\mathbf{X}$ is a set of features drawn from the article, $\mathbf{W}$ is a set of features drawn from activity on Twitter and $\mathbf{Z}$ is a set of features based on activity on Internet. The goal then is to predict the total count of the article of data on Twitter, as a function of the three feature sets. This work will include a uniquely curated dataset, and will empirically examine the impact that each combination of the three sets of features sets have on predicting the output.

Chapter 2

Literature Review

The inspiration for the modification to previous methods is drawn from biological models used in population genetics, specifically frequency-dependent selection. The general idea of frequency-dependent selection is that the fitness of an allele can be modeled if the sexual selection method and the frequency of that allele in past generations is known. If we consider an allele as genetic information that is being propagated within a population network, we can draw parallels to an article, which is information, being spread in a social network. We will look at general methods of information propagation on networks, how social media changes those models, and basic interpretations of text, including sentiment and readability of text.

2.1 Information Propagation on Networks

Propagation of information in a network has been studied from a variety of perspectives. Study of activity spikes in network have been done, modeling the spikes with power law decay and periodic occurrences and as temporal point processes, where previous events are not considered independent of future events [7][8][9]. Individual user credibility has also been studied, using probabilistic models of user activity based on their engagement with a known database of news stories [10]. An extension of the user-based approach considers the interaction between two users in information exchange [11]. They considered a forceful user aggressively spreading their message to nearby users, which would lead to a domination of a network with forceful ideas. These forceful users can be media outlets, influencers, or other social agents [11]. Other approaches use cross-disciplinary techniques to study the propagation of information within a technological network. Kumar and Geethakumari presented a method to measure dissemination using evolutionary game and graph theory [12]. Approaching the spread of information from a psychological

perspective focuses on analytical methods of the sources characteristics as well as the characteristics of the message [13]. Social media provides a unique capacity for characterizing credibility of source and message through the network of users 8[13].

2.2 Sharing on Social Media

Social media does not fit traditional mass communication models, so we should be look at how and why information is shared on social media. Those differences impact content distribution on social media networks. Carlson suggests that the action of sharing news by users on social networks and engaging with the content through comments and polls has the ability to shape the way the news is shared with the users from the mass media sources [14]. He also suggests that within the social media network, the users can shape the meaning of the news stories through the sharing of the story [14]. Lee and Ma worked from traditional media sharing motivations with use gratification theory and social cognitive theory, focusing on motivations that include information seeking and status attainment from online media engagement into the social media news sharing sphere [15] [16] [17]. They found that status seeking was still a significant indicator of intent to share news on social media, as well as socialization and prior social media experience. Lee and Ma argued that information seeking was viewed more as an archival process, as opposed to the other characteristics that were found to be statically significant for intent to share [15]. Lee, Ma and Goh looked at key factors of news sharing through the lens of diffusion of innovations, and found that perception of opinion leadership, intensity of social relationship and a general preference for online news were indicators of news sharing on social media [18].

2.3 Social Media and Information Propagation

Research methods into the determining the spread of news on social media takes two general approaches, a k-class classification of popularity of information on a network, or a regression approach where the output is the prediction of the the number of times that information is published on the network.

Keneshloo et. al. considered short term popularity, 24 hours, of an article based on 30 minutes of activity on social media as a regression, with the goal of predicting page

views of an article as they relate to how the article is being spread on Twitter. They used a combination of metadata from the article, as well as context based measures from that data, temporal measures of the article and Twitter temporal features. They compared linear regression modes with tree regression models and different combinations of their features to find that overall multivariate linear regression models worked best, and the combination of all features generally performed better than any combination of the features alone [1]. Arapakis et. al. looked at the feasibility of cold start prediction of news based on five metrics, shares, likes, comments, Tweets and page views. They compared Yahoo articles based on Facebook and Twitter interaction. They found that classification methods did not result in meaningful performance, nor did regression models provide a meaningful representation of popularity, though they did find that there are certain features that exhibited weak correlation with popularity. They believed that early stage measurements were necessary to more accurately predict popularity of news [2]. Tatar et. al. reviewed a variety of methods that were developed to predict popularity of news on social media across different platforms, during different time periods and with different feature sets. They noted that the field was very diverse and there were multiple design parameters that could be considered when approaching the problem. One conclusion they drew was the need for richer models that may be able to look at long term popularity [19].

Ko et. al. suggested that the total number of Tweets on the network could be found as the sum of a product of probability a user will Tweet or Retweet, external stimuli and total Tweets at different times. They considered the volume of articles entirely, and the volume of articles from two representative sources, one conservative and one liberal. Their empirical study found that using these measures resulted in a better prediction than the null model they considered, which was based entirely on the observations from the internet alone [20]. Zaman et. al. considered the popularity of a Tweet based Retweets, early in the lifetime of the Tweet, using a Bayesian model. Posterior distributions were found through Markov Chain Monte Carlo calculations. This method considered a dynamic time graph built from Tweets with less than 1800 retweets over a period of a week [3]. Rizoiu et. al. explored a multivariate self-exciting temporal point processes, or Hawkes Process, with some modifications for prediction of views of a video based on Twitter and YouTube activity. They modified the point process to consider it as an continuous intensity process, which can be discrete given the nature of data observations. They showed that this method outperformed general multivariate linear regression models when predicting popularity of the videos [4].

Chapter 3

Methodolgy

3.1 Empirical Analysis

The proposed model will tested over a uniquely curated dataset. Articles written about five different events were collected from the Internet over a two month period. Four of the five centered on a political topic, and the last one was centered on a death in entertainment. Those articles activity on Twitter were then tracked, by looking for link to the article in the Tweet. The articles were chosen based of noted social media interest in the event, or perceived interested in the event, based on the possible social media discussions it may include. Details of the events are in the appendix.

To complete this task a data collection method was designed with existing API's and available programming packages. Cleaning and filtering of the dataset was performed to eliminate data that would skew results. Features were extracted from the remaining data and used to train the machine learning models and the adaptive filter for prediction of Tweets.

3.1.1 Data Collection

Two commercially available API's for web scraping and crawling were used to collect articles from the internet. The services used to crawl the web were NewsAPI and Event Registry. The service provided a URL list as an output. The query was limited to a predefined keyword, or keyword list, that roughly described the news event and a date parameter. The date was limited to one day, to preserve a time resolution of 24 hours.

Twitter provides a JSON format of data about their network, which includes a variety of data about the activity surrounding a Tweet.If a Tweet had a link to an article is was collected. The data collected is shown on table (3.1).

timestamp	tweet text	username	followers
favorites	retweets	tweet id	hashtags

Table 3.1: Twitter Data Collected

A filter was designed to remove stories that were not directly related to the subject or topic. A set of words that were relevant to the event was built from general knowledge about the event, subject or subjects that naturally would relate to the original event. The filter was used on the headline, the keywords and the first 50 percent of the text collected. The filter word bank was increased over time as needed. The initial word bank was used to filter the data. The articles that were filtered out were then reviewed to see if any articles were incorrectly filtered out. The filter work bank was updated over time until the articles removed were found to be acceptable.

The Twitter activity data had to be filtered due to the constraints in the collection process. Free collection was only available for seven days in the past. In order not to miss any Tweets, the collection process was ran every six days. Tweet ID was used to remove the duplicates.

3.1.2 Article Features

The Python package Newspaper3k was used to scrape individual features from the URL list generated by the API crawling services. The features that were extracted for this study are in in table (3.2) below.

Headline	Text	Authors	Publication Date	Keywords

Table 3.2: Article Features Collected

Headline, text, authors, and publication date were taken directly from the source code in the web page. The keywords are found with a NLP function from Newspaper3k. The package used a simple measure to find the keywords of an article. It first removed any stopwords from the text, using a predefined dictionary of stopwords included with the package. It then took a raw count of how many times a word occurred throughout

the text. It normalized the counts of the words by the length of the text without the stopwords and reported back the top ten keywords in descending order of frequency.

From that general data, headline sentiment, source bias and headline complexity were generated. The final features for articles consisted of those three measures.

Complexity was determined by readability measures. The elements of the measures were extracted using simple open-source tools and packages. Both the D-C and the F-K Reading ease measures were chosen for their straightforward application. The words and sentences were parsed and counted, with the natural language toolkit package, and difficult words were checked against a dictionary. A simple method for counting syllables was adopted from a generic code snippet found on Stack Overflow [21]. There was the assumption that each headline was only a single sentence. A binary classifier was built based on threshold values of high school graduate. If a Tweet exceeded a high school graduate score it was assigned a one, and a zero otherwise. The F-K method method is given below. The score can be translated to grade level with the by the bounds noted in table (3.3).

$$S = 206.835 - 1.1015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

The D-C Readability formula is given by the equation below. The score can be trans-

Score	School Level	Notes
100.00-90.00	5th Grade	Very easy to read. Easily understood by an average 11-year-old student.
90.00-80.00	6th Grade	Easy to read. Conversational English for consumers.
80.00-70.00	7th Grade	Fairly easy to read.
70.00-60.00	8th & 9th Grade	Plain English. Easily understood by 13- to 15-year-old students.
60.00-50.00	10th to 12th Grade	Fairly difficult to read.
50.00-30.00	College	Difficult to read.
30.00-0.0	College Graduate	Very difficult to read. Best understood by university graduates.

TABLE 3.3: Flesch-Kincaid Readability Score

lated to grade level with the table (3.4).

$$S = 20.1579 \left(\frac{\text{difficult words}}{\text{words}} * 100 \right) + 0.0496 \left(\frac{\text{words}}{\text{sentences}} \right)$$

Score	Notes
4.9 or lower	easily understood by an average 4th-grade student or lower
5.0-5.9	easily understood by an average 5th or 6th-grade student.
6.0-6.9	easily understood by an average 7th or 8th-grade student
7.0-7.9	easily understood by an average 9th or 10th-grade student
8.0-8.9	easily understood by an average 11th or 12th-grade student
9.0-9.9	easily understood by an average 13th to 15th-grade (college) student

TABLE 3.4: Dale-Chall Readability Score

Source bias was determined by considering the bias classification from two outside sites, allsides.com and mediabiasfactcheck.com. Both sites used a multivariate approach to classify bias. Priority was given to the ratings from Media Bias Fact Check, and All Sides was used as a supplementary source. Two classifiers were created. The first mapped a value of 0 to center, or least-biased sources, and mapped a value of 1 to right, lean-right, lean-left and left sources. The second classifier mapped a value of 0 to least-biased, lean-left and lean-right, and mapped a 1 to left or right sources. In both cases a missing source was mapped to -1.

Media Bias Fact Check utilizes a scoring system, mapping a value of [0,10] to different features. Zero was the least bias, and ten was extremely biased. The scoring is dependent on four categories, biased wording/headlines – such that loaded emotive words are used, or if the headline and the article match, factual/souring – if the article is factually based and well sourced, article choice – if the source gives equal time to both sides of a article and strength of political affiliation, which they cite as being rather subjective. Each category is assigned a value from [0,10] and the score is normalized for the four categories [22]. Media Bias Fact Check provided a larger more robust database of sources, as well as a better defined methodology.

Allsides.com is an 5-class classifier, with classes for Left, Lean-Left, Center, Lean-Right and Right. They assign a classification based on "a patented process to identify and display the average judgement of Americans. [They] update Media Bias Ratings and confidence levels as more data is gathered and assessed, or as outlets change their bias over time" [23]. Their method can be based on five different characteristics, blind bias survey, third party data, community feedback, editorial review of independent research [23].

Media Bias examined 1761 sources. All Sides examined 374 sources, some of which were not labeled. There was a number of local stations that were affiliates of larger nation-wide of global networks. In that case the bias of the larger source bias was assigned to the local station. There was also certain sources that did not have affiliates, but carried certain syndicated programs that were rated with a certain bias. In that case the bias of the syndicated program was transferred to the unlabeled source. The final adjustment was made for topical unlabeled sources, such as entertainment or sports sources. Entertainment sources were given a rating of 2 and sports sources were given a rating of 3. Questionable sources from media bias fact check that were not rated for a particular bias were omitted.

Headline sentiment analysis was performed with the Valance Aware Dictionary for sEntiment Reasoning, VADER, sentiment analysis. This method is a rule-based model for general sentiment analysis, based on qualitative and quantitative methods that is tuned for microblog content[24]. It builds from classical sentiment analysis methods using sentiment lexicons, like LIWC, GI and ANEW, to incorporate a more robust method for determining sentiment by looking at the following rules; punctuation, capitalization, degree modifiers, contrastive conjunctions and tri-gram preceding sentiment-laden lexical features. In this case capitalization may not necessarily lend much in terms of information about sentiment, given an expected level of professionalism in writing a headline, though that may not necessarily hold over some less traditional news sources. The output for the classifier is positive, neutral and negative. The positive and negative sentiment are combined into a single class, non-neutral, and the neutral class is considered as itself. Non-neutral was mapped to one and neutral was mapped to zero.

3.1.3 Twitter Activity Features

The raw Twitter activity data was binned to generate the features for a fifteen minute time interval. A datetime list was created and the Tweets were sorted into the bins by comparing the timestamp of the Tweet to the time interval of the bin. To standardize the size of the feature set the date time bins were started at 00:00 of the day that the article was published. This is not necessarily when the article was first published on Twitter, and there was no guarantee that the article received any Tweets that day, or any day in the future. When there were Tweets in a given time period, information was extracted. To count Retweets the text was extracted and if the first few characters were 'bRT@', it was counted as a Retweet. All other features were taken directly from the provided JSON file. Those features included, the number of times a Tweet was favorited, the number of followers a user who Tweeted had and the number of times that that Tweet had been Retweeted. For a set of values the maximum, minimum, mean and standard deviation of that list was taken and recorded.

Table (3.5) shows the features that were used.

Total Tweets	Original Tweets	Retweets	Time on Twitter
Max Favorites	Min Favorites	Mean Favorites	STD Favorites
Max Retweets	Min Retweets	Mean Retweets	STD Retweets
Max Followers	Min Followers	Mean Followers	STD Followers

TABLE 3.5: Twitter Features

3.1.4 Frequency Features

Frequency features were used to incorporate past information from the Internet. Fifteen features were considered for the frequency features set. They included, sentiment, as a single measure, bias, as two measures, readability, as two measures and keywords, as a measure for each of the top ten keywords. Keyword frequency by day was found by extracting the individual word from the keyword list and adding it to a set. The number of instances of a given words were counted over the set and then normalized by the size of the set to find the frequency of the specific word. All the frequency should sum to one, within rounding precision of eight digits. The lists were generated on a per day basis, filtering by the day of the publication. The other five frequencies, sentiment, bias and readability were found by summing instances of each class and dividing by the total

instances for that period. Those frequencies should also sum to one within rounding precision of the computer used to calculate the values. These frequencies were tracked for four previous days.

3.1.5 Model Fitting and Testing

We considered events individually. We used sixty percent of the data for training, twenty percent for validation and twenty percent for testing. The sets will be randomly selected by randomizing the order of the file list that builds the sets. We used two different methods predicting y(t+1), the number of Tweets at time t+1, based only on time t and predicting y(t+1) based on combinations of the features described above. Four different models will be fit to the data, General Regression, LMS adaptive filtering, BRR and RF regression models will be used to model and predict the change and the growth of total Tweets on the network. The Tweets will only be predicted for a period of 14 days. The log based transformation will be of the growth plus one to account for the points in time that the data is zero.

The general fit regressors will be fit with packages available in python, numpy and scipy. Numpy will be used to find a linear, a second degree polynomial and a third degree polynomial fit, while scipy provides curve optimization which worked best for the logarithmic fit and exponential fit. A parameter for exponential fit that required tuning was initial conditions. Because the focus of this study is not to provide a specific model for a specific prediction of a specific event, the model will not be tuned and a general initialization value of A=1 and B=-0.1 for the following equation; $y = A\exp(Bx)$. This method will be fit to the individual article and not generalized over all articles within the training set. The models and their input are given as

$$y(t+1) = at + b$$

This is the linear case, where y(t+1) is the prediction of Tweets in the next time period, a and b are learned constants, t is time.

$$y(t+1) = at^2 + bt + c$$

This is the linear case, where y(t+1) is the prediction of Tweets in the next time period, a, b and c are learned constants, t is time.

$$y(t+1) = at^3 + bt^2 + ct + d$$

This is the linear case, where y(t+1) is the prediction of Tweets in the next time period, a, b, c and d are learned constants, t is time.

$$y(t+1) = a + ln(bt)$$

This is the linear case, where y(t+1) is the prediction of Tweets in the next time period, a and b are learned constants, t is time.

$$y(t+1) = a \exp bt$$

This is the linear case, where y(t+1) is the prediction of Tweets in the next time period, a and b are learned constants, t is time.

To examine the impact of the frequency on prediction, different combinations of the feature sets were used to predict the outcome. The individual models and their input and output will be described below.

LMS adaptive filtering will be fit to the individual model and not generalized over all articles within the training set. Padasip, an adaptive filtering package developed for python was used for this task [25]. The general form of the solution is

$$y(t+1) = \beta x^T$$

where y is the predicted output, β is a 1xn vector of weights, and x is a 1xn vector of features, where n is the number of features for the specific case.

There are seven combinations of features, in general the input vector varies, and the order of the features is arbitrary, as the method solves for the weights based on the given order of features. The order of features just needs to be maintained through the individual case. A review of the method in which weights are determined is included in the index. To test the article features we can assign the input x as the combination of three different features sets, article features, **X**, Twitter Activity features, **W**, and the frequency of the previous four days of articles, **Z**.

For case one, we consider only the article features, the input vector x is given as a vector of the following features; headline sentiment, center source bias, extreme source bias, F-K readability, D-C readbility, total number of tweets at time t, and time, let us define

this set as **X**.

For the second case we consider Twitter activity as the set of vectors that include the following variables for the 15 minute period; total Tweets, original Tweets, Retweets, Maximum, minimum, mean and standard deviation from the list of times the Tweet has been favorited, Maximum, minimum, mean and standard deviation from the list of the number of followers of the user who Tweeted, Maximum, minimum, mean and standard deviation of the list of times Tweet has been Retweeted. The additional features are not based on the 15 mintue period, time on the network, total tweets and time from the beginning of the day that the article was published. We call this feature set **W**.

The third case we consider the frequency features where we take the center source bias frequency from the previous day, from two days prior, from three days prior and from four days prior, the same for the extreme source bias, the same for headline sentiment, the same for the F-K readability of the headline, the same for the D-C readability of the headline, and the same for each of the top ten keywords for the article. It also include the total number of Tweets at time t, and time t. This feature set is known as **Z**.

The next case considers the combination of **X** and **W**, not duplicating total number of Tweets at time t, or time t, but still including a single instance of both. The next case takes the combination of **X** and **Z**. The next case is the combination of **W** and **Z**, and the final case is the combination of all three feature sets, **X**, **W** and **Z**. Each case not duplicating the total Tweets at time t, or time t itself, but still including a features for both of those inputs.

BRR will be generalized for the entire test dataset. A pre-built function from the sklearn library will be used to solve the general equation below, based on methods described in the appendix.

$$y(t+1) = \beta x^T$$

where y is the predicted output, w is a 1xn vector of weights, and x is a 1xn vector of features, where n is the number of features for the specific case. As above, there are seven combinations of features, with the same description of the features, and general characteristics as before. No parameters will be tuned for this model.

RF regression build multiple decision trees to determine the outcome. Each tree make splits based on features that maximizes the information gain by variance reduction with MSE and MAE. For each feature the following values will be determined

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \mu|$$

where the output is the MSE, or MAE, respectively, y_i is the label for the instance, N is the number of instances, μ is the empirical mean of the output label. The features described above, and the combinations given above are the basis for the different cases and features that were used for each case. The final label is taken from a poll of the trees. [26]. The same combination of features were tested with this model, under the same assumptions and descriptions as noted above.

The models will be tested on the same subsets of data sets. The general regression models will be applied to all data sets, as well as the unbiased data for the articles that spread on the network. For all models only the articles that spread on twitter will be tested. The Mueller Report data was omitted from the test due to the size of the data set. Each of the four methods will be tested on individual articles.

Feature importance will also be consider for the BRR and RF Regression. There are two general methods that are used for this task, weight of coefficient and feature importance. Those values are given by functions from the sklearn package.

3.1.6 Measure of Error

Error will be measured separately per event. We looked at three different error measures, MSE, MAE and the R-Squared value. The error for the BRR and the RF Regression will consider test error as well as validation error. The error for the the LMS will consider error only for the test set. The error for the general regression models will consider the test set error, as well as full set error, measured over all articles, articles with spread on Twitter and articles with spread on Twitter with the bias removed, for a better functional fit.

Mean square error is given as

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Mean Absolute Error is given as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

R squared value is given as

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where N is the length of $\mathbf{Y}$, y_i is the ith value of $\mathbf{Y}$, $\hat{y}_i$ is the predicted value for the ith value, and $\bar{y}$ is the mean of $\mathbf{Y}$.

Because the variance of $\mathbf{Y}$ is zero for articles that did not spread on Twitter, and for articles that had limited spread on Twitter, the variance was low, some of the R-squared values were either not a number, in the case of dividing by zero variance, or negative, in the case of a small variance and a large MSE. In those cases the value was set to zero to indicate that the fit was poor. Additionally the output of the change predictor was rounded to the nearest whole number. If the value of the error exceeded a larger threshold it was limited to that threshold to avoid infinite error.

Chapter 4

Experiment

4.1 Overview

59,401 articles were collected from 2,101 unique news sources over the five individual events. After the articles were filtered, 54,604 articles remained from 1,960 unique news sources. Those stories were shared 5,419,470 times on Twitter. Table (4.1) breaks down the data by individual news story. The values for the article quantity and news sources are limited to two months, while the Twitter exposure is considered over a period of 3 months, 1 month past the last article collection date from the Internet.

News Event	Unfiltered Article Quantity	Filtered Article Quantity	Unfiltered News Sources	Filtered News Sources	Filtered Tweets
Jussie Smollett	8808	7199	1011	910	341864
New Zealand	8855	8281	1114	1063	470956
Mueller Report	31091	30423	1361	1336	4170401
Nipsey Hussle	8216	6583	936	816	203942
Rep. Omar	2431	2118	573	526	232307

TABLE 4.1: Article, Source and Tweet Volume by Event

Figure (4.1)-Figure (4.4) are the distribution of readability, sentiment and the source bias of the articles. In Figure (4.1) the threshold for the readability is at 9.0. In Figure

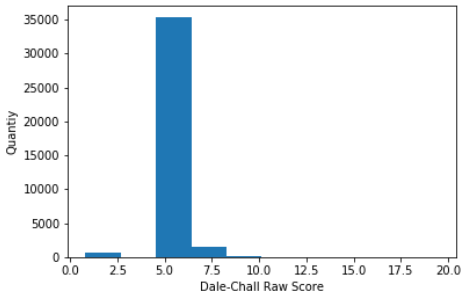

FIGURE 4.1: D-C Readability Dist.

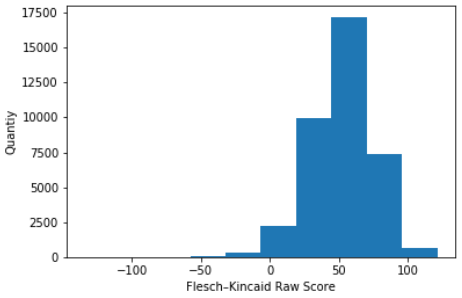

FIGURE 4.2: F-K Readability Dist.

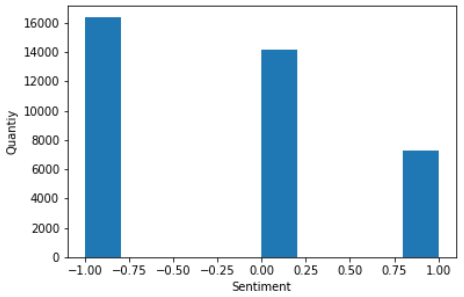

FIGURE 4.3: Sentiment Dist.

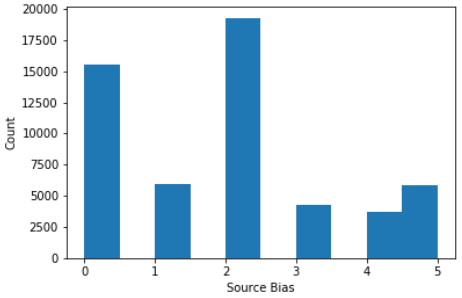

FIGURE 4.4: Source Bias Dist.

(4.2) the threshold for the readability is at 50.0. For Figure (4.3) negative sentiment is -1, neutral sentiment is 0 and positive sentiment is 1. Figure (4.4) shows source bias where the labels are given as 0 for missing, 1 for extreme left, 2 for lean-left, 3 for least-baised, 4 for lean-right and 5 for extreme right

Figure (4.5)-Figure (4.13) show three different articles through time. The y axis was new Tweets, or total Tweet volume for the log transformation of the Tweets and the x-axis is increments of 15 minutes. Figures (4.5)- (4.7) show the output for an article from the Nipsey Hussle event set, titled "Nipsey Hussle Dead at 33, Cause of Death Gunshots to Head and Torso". It was published by TMZ, labeled with non-objective sentiment, non-center source bias, non-extreme source bias, simple F-K readability and difficult D-C readability. It was published on the first day of the event, and was one of the most Tweeted articles. Figures (4.8)-(4.10) are from The New Zealand data set, from an article titled "Kellyanne Conway says 'shut up and pray' after New Zealand massacre" published on March 17th, three days after the shooting. It was published by AOL, and was labeled with non-objective sentiment, non-central source bias, non-extreme source bias, hard to read by the F-K standard and easy to read by the D-C standard . Figures (4.11)-(4.13) were from the Jussie Smollett data set. The title was "Chicago says Smollett owes city $130K for investigation". It was published by Times of Israel, labeled with objective headline sentiment, labeled with non-center source bias,

non-extreme source bias, with difficult readability from both measures.

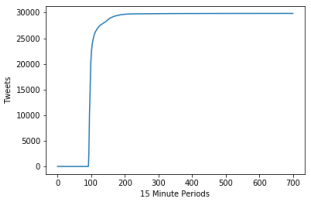

FIGURE 4.5: Total Volume of Tweets

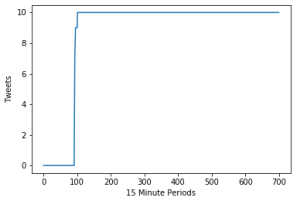

FIGURE 4.6: Log Transform of Tweet Volume

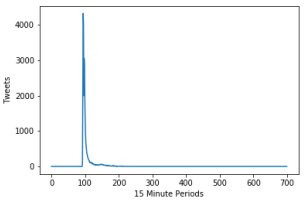

FIGURE 4.7: New Tweets

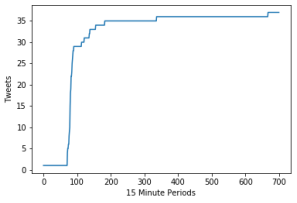

FIGURE 4.8: Total Volume of Tweets

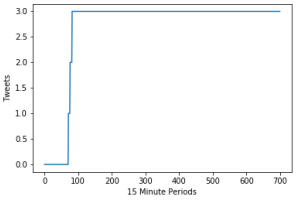

FIGURE 4.9: Log Transform of Tweet Volume

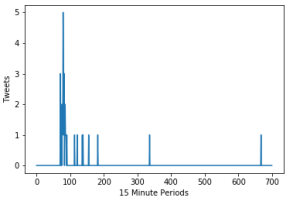

FIGURE 4.10: New Tweets

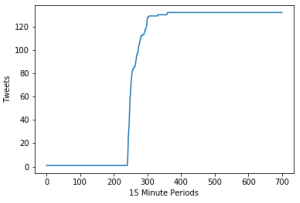

FIGURE 4.11: Total Volume of Tweets

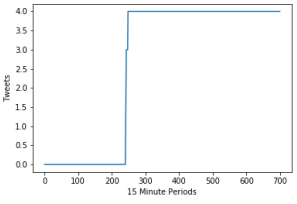

FIGURE 4.12: Log Transform of Tweet Volume

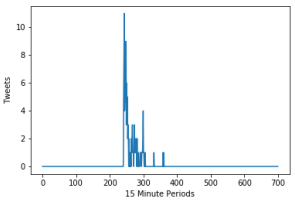

FIGURE 4.13: New Tweets

There were a number of articles that were note published on Twitter. Per event there were 3480, 2659, 12496, 3262, and 704 articles that were not published on Twitter. As a percentage, 48.43%, 32.11%, 41.07%, 49.55%, and 33.24%, for the JS, NZ, MR, NH and RO events, respectively.

Figures (4.14)-(4.23) compare the volume of articles published per day on the internet compared to the volume of Tweets about those articles. Table (4.2) shows the top ten news sources by event and the normalized frequency of articles published on the Internet. Table (4.3) shows top ten news sources by event and the normalized frequency of Tweets. Table (4.4) shows the top ten keywords by event and there normalized frequency of published on the Internet. Table (4.5) shows top ten keywords by event and

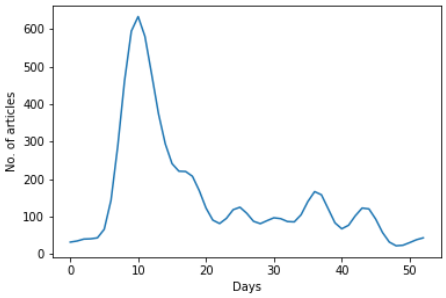

FIGURE 4.14: Web Articles for JS

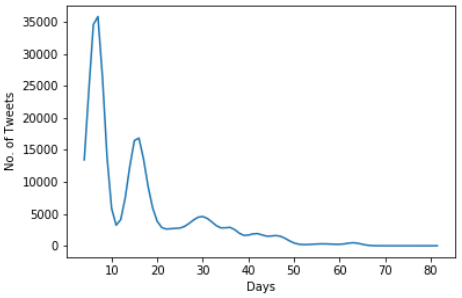

FIGURE 4.15: Article Tweets for JS

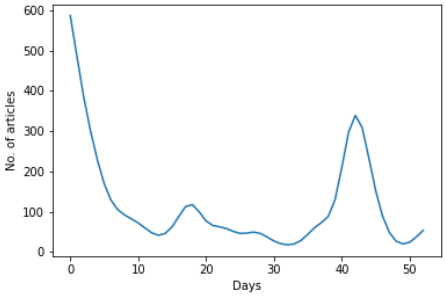

FIGURE 4.16: Web Articles for NZ

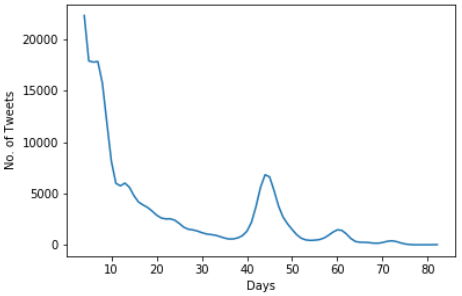

FIGURE 4.17: Article Tweets for NZ

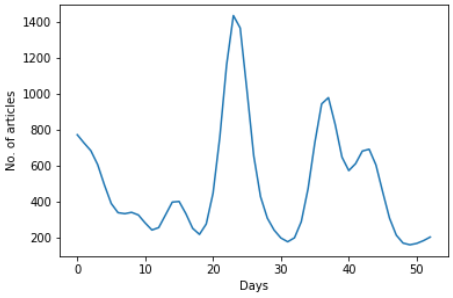

FIGURE 4.18: Web Articles for MR

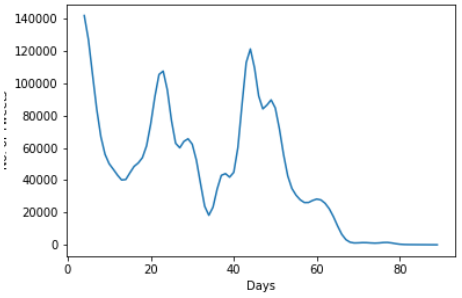

FIGURE 4.19: Article Tweets for MR

the normalized frequency of Tweets.

Figures (4.24)-(4.33) are examples of frequency changes over time for the articles published on the Internet. Figures (4.34)-(4.39) are the correlation matrices of the different feature sets to the change and growth. The predicted values are the last two elements of the matrices.

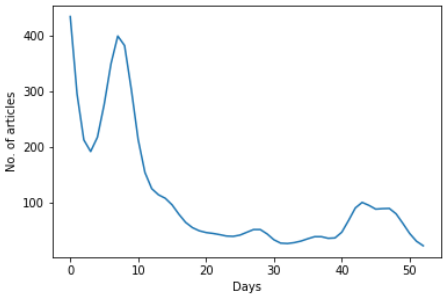

FIGURE 4.20: Web Articles for NH

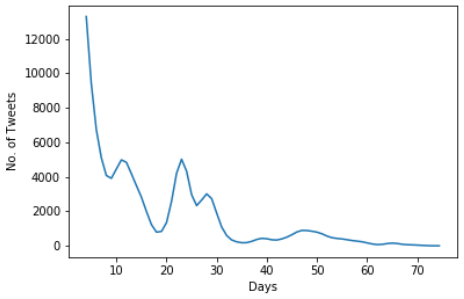

FIGURE 4.21: Article Tweets for NH

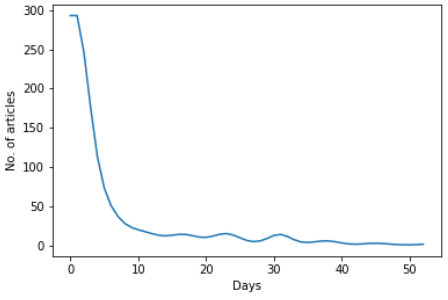

FIGURE 4.22: Web Articles for RO

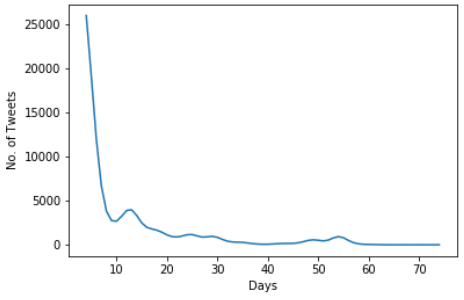

FIGURE 4.23: Article Tweets for RO

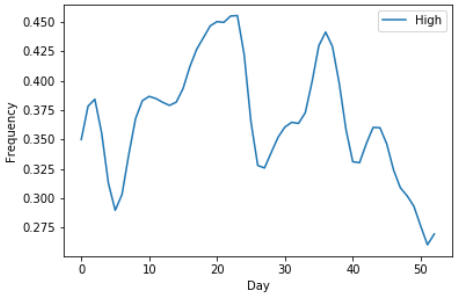

FIGURE 4.24: F-K Readability Through Time for JS

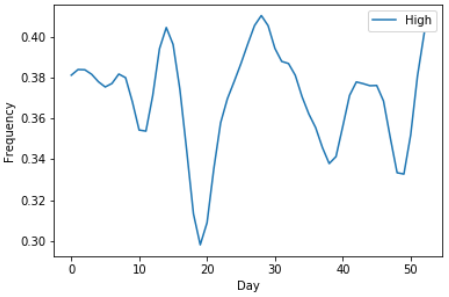

FIGURE 4.25: F-K Readability Through Time for MR

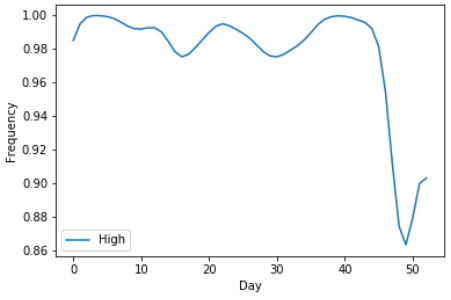

FIGURE 4.26: D-C Readability Through Time for JS

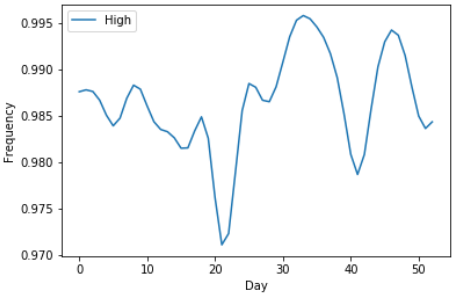

FIGURE 4.27: D-C Readability Through Time for MR

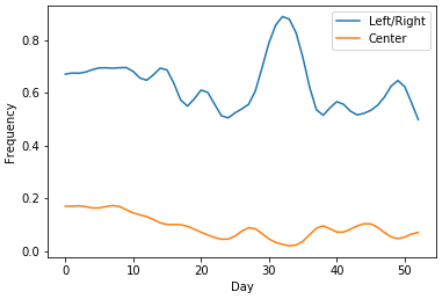

FIGURE 4.28: Center Bias Through Time for NZ

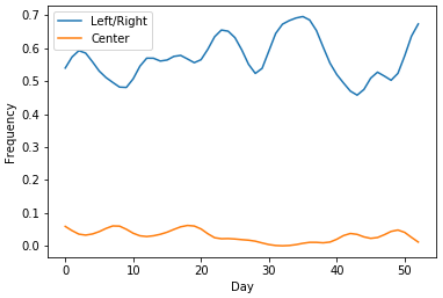

FIGURE 4.29: Center Bias Through Time for NH

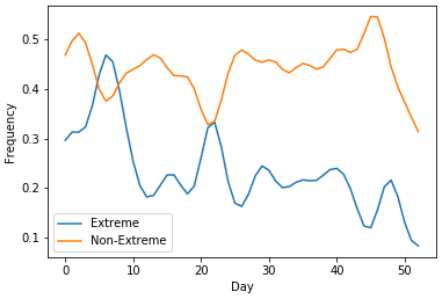

FIGURE 4.30: Extreme Bias Through Time for JS

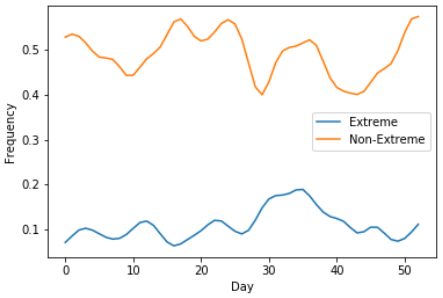

FIGURE 4.31: Extreme Bias Through Time for NH

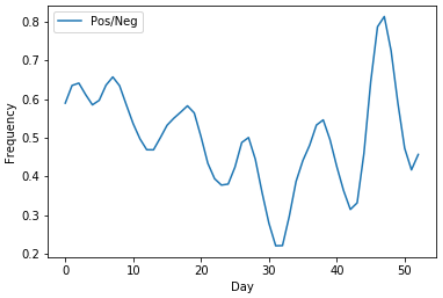

FIGURE 4.32: Sentiment Through Time for RO

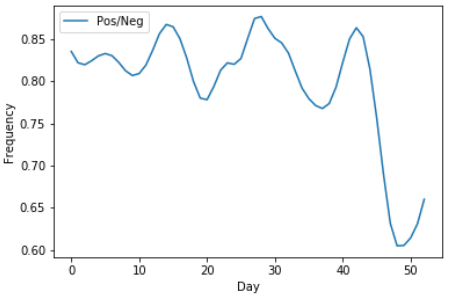

FIGURE 4.33: Sentiment Through Time for NZ

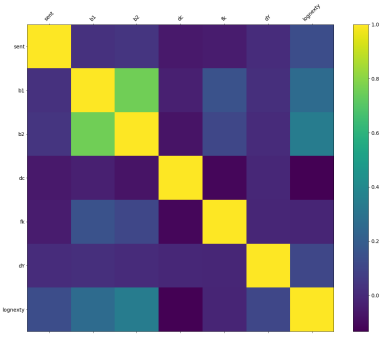

FIGURE 4.34: X Feature Correlation for RO

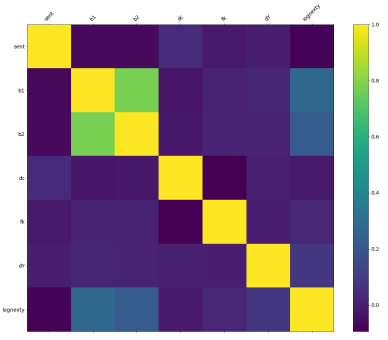

FIGURE 4.35: X Feature Correlation for NZ

	JS	Freq.	NZ	Freq.	MR	Freq.	NH	Freq.	RO	Freq.
1	Chicago Tribune	1.819	Stuff	6.376	The Hill	3.001	Complex	2.020	Business Insider	2.833
2	Daily Mail	1.639	Independent	3.840	Business Insider	2.038	Uproxx	1.549	Mediaite	1.747
3	Fox News	1.264	NZ Healrd	2.355	Fox News	1.841	Sohh	1.534	Twitchy	1.700
4	FreeRepublic	1.236	Daily mail	2.307	Raw Story	1.624	XXL Mag	1.458	BizPac Review	1.700
5	Breitbart	1.111	Yahoo	1.811	Washington Post	1.519	Rolling Out	1.230	Fox News	1.653
6	Daily Caller	1.070	First Post	1.534	Haaretz	1.512	The Grio	1.109	Daily Caller	1.605
7	The Grio	0.945	Reuters	1.328	Washington Examiner	1.328	Metro	1.079	Daily Wire	1.511
8	Page Six	0.8751	Haaretz	1.292	DDaily Mail	1.068	Vibe	1.079	NY Post	1.322
9	SLT Today	0.8751	Business Insider	1.232	CNN	0.963	TMZ	0.957	Washington Examiner	1.322
10	The Wrap	0.847	India Times	1.159	India Times	0.924	Hophopdx	0.957	Town Hall	1.086

TABLE 4.2: Top Ten Web Sources by Frequency

	JS	Freq.	NZ	Freq.	MR	Freq.	NH	Freq.	RO	Freq.
1	Breitbart	13.81	NZ Hearld	6.906	Washington Post	15.99	TMZ	40.41	Fox News	12.90
2	Fox News	10.41	Gateway Pundit	6.772	NY Times	12.78	XXL Mag	12.67	NY Post	9.896
3	Gateway Pundit	8.267	NY Times	5.608	The Hill	8.735	USA Today	10.49	Washington Post	8.809
4	Chicago Tribune	5.271	CNN	5.559	Fox News	6.318	Vibe	3.302	Breitbart	5.144
5	TMZ	4.911	Indpendent	4.132	CNN	5.835	LA Times	3.042	NY Mag	5.124
6	National Review	4.370	Al Jazeera	4.070	NBC News	3.546	NBC News	1.937	Daily Caller	2.235
7	Daily Mail	3.776	The Hill	3.901	The Atlantic	2.787	The Source	1.844	News Max	3.101
8	NBC News	3.239	BBC	3.437	Raw Story	2.294	NY Times	1.677	NY Times	2.722
9	Daily Caller	2.737	Think Progress	3.350	The Daily Beast	2.007	Rap-Up	1.389	The Atlantic	2.285
10	NY Post	2.699	Go	3.204	Breitbart	2.002	complex	1.356	NBC News	2.277

TABLE 4.3: Top Ten Twitter Sources by Frequency

	JS	Freq.	NZ	Freq.	MR	Freq.	NH	Freq.	RO	Freq.
1	Smollett	5.938	Zealand	3.807	Trump	4.992	Nipsey	5.904	Omar	6.156
2	Jussie	4.501	Mosque	3.088	Report	4.964	Hussle	5.166	Trump	3.900
3	Chicago	3.190	Shooting	2.806	Mueller	4.642	Rapper	2.876	Ilhan	3.199
4	Case	2.320	Christchurch	2.397	Barr	2.980	Los	2.630	Video	2.698
5	Charges	2.243	Attack	2.000	President	2.843	Hussles	2.223	911	2.576
6	Smolletts	1.630	Synagogue	1.094	House	2.226	Angeles	2.162	President	2.576
7	Empire	1.589	Shootings	0.897	Muellers	2.027	Shot	1.210	Rep	2.427
8	Foxx	1.290	Gun	0.802	Special	1.523	Store	1.140	Omars	1.707
9	Attack	1.175	Victims	0.802	Investigation	1.476	Holder	1.065	Trumps	1.673
10	actor	1.105	gunman	0.795	justice	1.470	memorial	1.033	comments	1.598

TABLE 4.4: Top Ten Web Keywords by Event and Relative Frequency

4.2 Prediction of Tweets Based on Time

Table (4.6) is the MSE for the MMSE fit regressors for change and growth for all articles in the dataset and for articles that were published on Twitter.

	JS	Freq.	NZ	Freq.	MR	Freq.	NH	Freq.	RO	Freq.
1	Smollett	5.645	Zealand	4.378	trump	5.094	Nipsey	6.6403	Omar	5.642
2	Jussie	4.843	Mosque	3.448	Mueller	4.545	Hussle	5.387	ilhan	4.002
3	Chicago	3.319	Christchurch	2.736	Report	4.285	Told	3.020	911	3.372
4	Case	3.106	shooting	2.290	President	2.863	Hussles	2.285	Trump	2.386
5	Charges	2.429	Attack	1.945	Barr	2.723	La	1.612	Rep	2.236
6	Foxx	2.362	White	1.739	House	1.580	Death	1.592	Omars	2.083
7	Smolletts	2.111	Shootings	1.678	Investigation	1.531	Rapper	1.552	Muslim	1.981
8	Kim	1.764	Killed	1.511	Justice	1.501	Life	1.422	President	1.288
9	States	1.5232	Muslim	1.114	muellers	1.423	shot	1.338	Attacks	1.245
10	dropped	1.275	Attacks	1.080	Special	1.372	memorial	1.295	Video	1.202

TABLE 4.5: Top Ten Twitter Keywords by Event and Relative Frequency

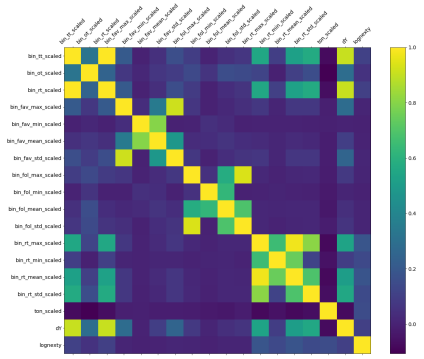

FIGURE 4.36: W Feature Correlation for RO

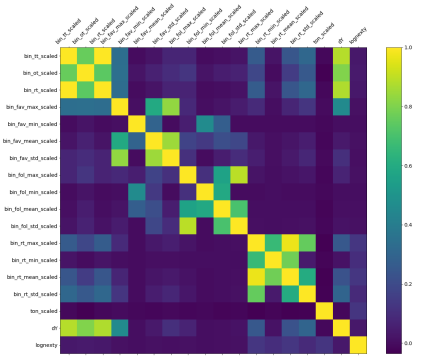

FIGURE 4.37: W Feature Correlation for NH

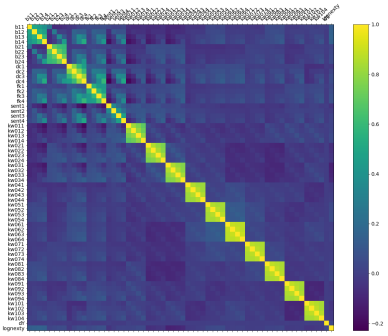

FIGURE 4.38: Z Feature Correlation for NH

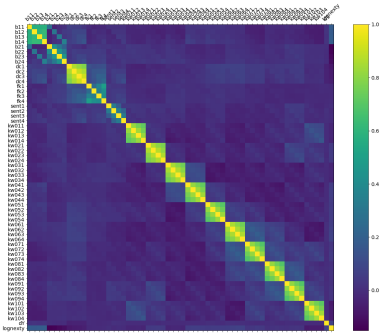

FIGURE 4.39: Z Feature Correlation for JS

Table (4.7) is the table of MSE from the MMSE Fit Regressors after the bias was removed.

Figure (4.40) is the distribution of the time it takes for a tweet to reach maximum exposure as measure over the 30 days.

Table (4.8) is the resulting mean square error after restricting the time for the MMSE fit Regressors, where b is the number of 15 minute periods.

Model	Change MSE	Log Growth MSE	Sub Change MSE	Sub Log Growth MSE
Linear	23206.4265603298	5.02994354354815	39775.369323254	8.62122660719139
Poly, n=2	16698.5951978501	4.7357517830892	28621.0713850024	8.1169875812607
Poly, n=3	11169.5535425991	4.5038352970299	19144.400202149	7.7194871793278
Logarithmic	23606.0955011159	9.210517238094	40460.394210023	15.786649609796
Exponential	854114.848100904	578.682219961997	1463890.41633624	88.834261263117

TABLE 4.6: Average MSE for MMSE Fit Regressors for All Data and Data That Spread

Model	Change MSE	Log Growth MSE
Linear	8745.693057445598	5.244634421324561
Poly, n=2	6441.715847566077	4.8222227718890744
Poly, n=3	4836.755705080069	4.385039516018547
Logarithmic	10880.813604644649	9.051535408269256
Exponential	859395.8057239257	52.81560313120043

TABLE 4.7: Average MSE for MMSE Fit Regressors for Non-Biased Set

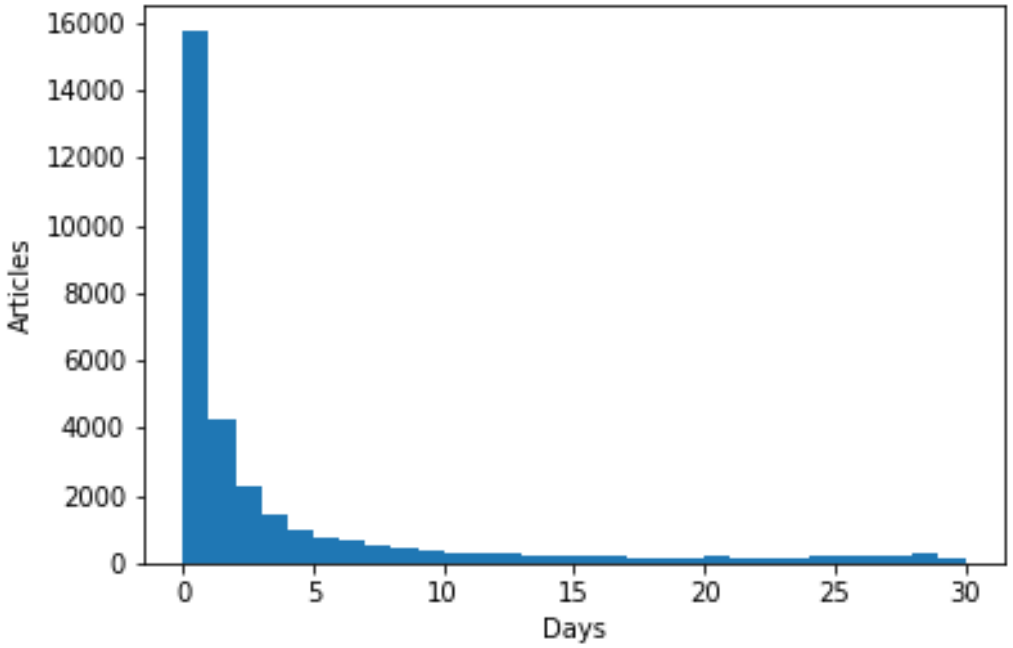

FIGURE 4.40: Distribution of Days to Reach Maximum Twitter Exposure

Figures (4.41)-(4.46) are the MSE, MAE and R2 box plots for the MMSE fit regressors. Figures (4.47)-(4.52) are the MSE, MAE and R2 box plots for the MMSE fit regressors without the exponential fit.

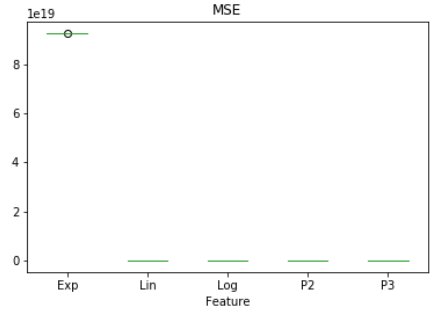

FIGURE 4.41: MSE for MMSE Fit Models for Change

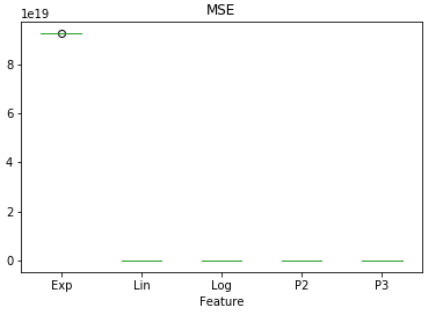

FIGURE 4.42: MSE for MMSE Fit Models for Log Growth

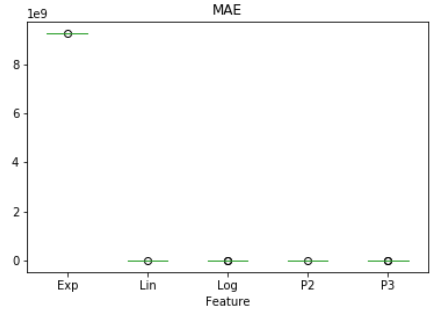

FIGURE 4.43: MAE for MMSE Fit Models for Change

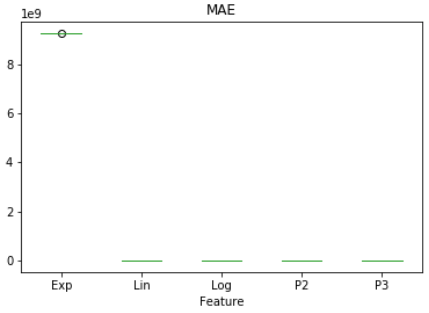

FIGURE 4.44: MAE for MMSE Fit Models for Log Growth

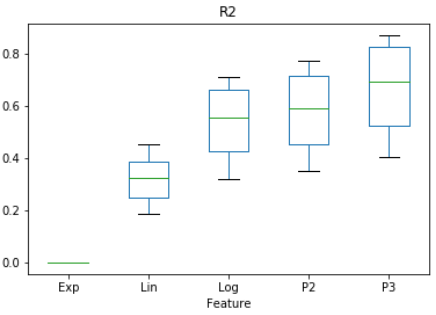

FIGURE 4.45: R^2 for MMSE Fit Models for Change

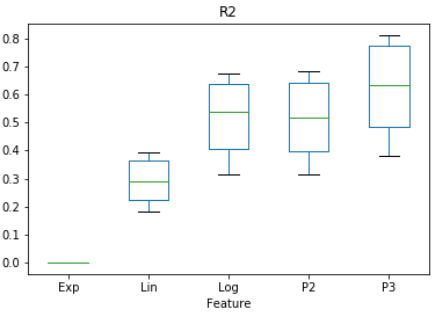

FIGURE 4.46: R^2 for MMSE Fit Models for Log Growth

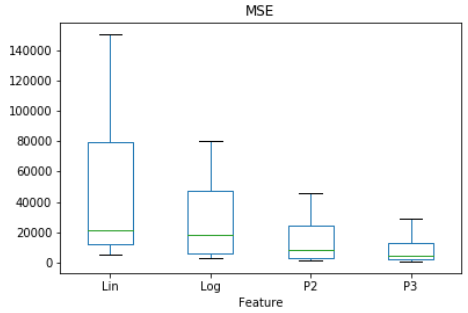

FIGURE 4.47: MSE for MMSE Fit Models for Change

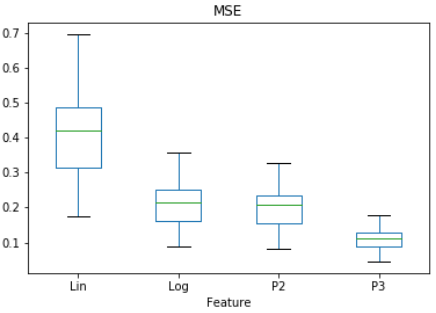

FIGURE 4.48: MSE for MMSE Fit Models for Log Growth

Model	Change b=2001	Log Growth b=2001	Change b =1001	Log Growth b=1001	Change b=11	Log Growth b=11
Linear	1171.36258	0.4556328	447.529	0.2260587	0.4506	0.00200
Poly, n=2	849.090185	0.4393559	330.504	0.219353	0.3555	0.00200
Poly, n=3	574.90159	0.4257199	228.567	0.2136236	0.2676	0.00200
Logarithmic	1215.069	0.8604626	474.0563	0.4310297	0.5411	0.00399
Exponential	44482.372	65.5089198	17481.544	64.91282	29.5469	64.4560

TABLE 4.8: MSE for Smaller Time Frame

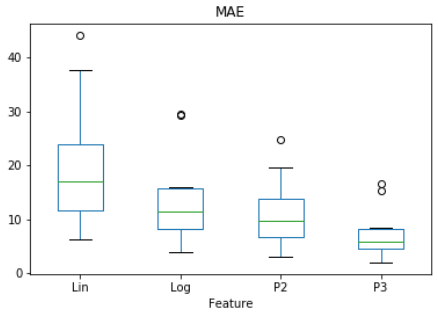

FIGURE 4.49: MAE for MMSE Fit Models for Change

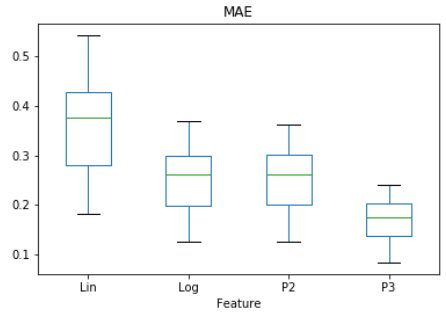

FIGURE 4.50: MAE for MMSE Fit Models for Log Growth

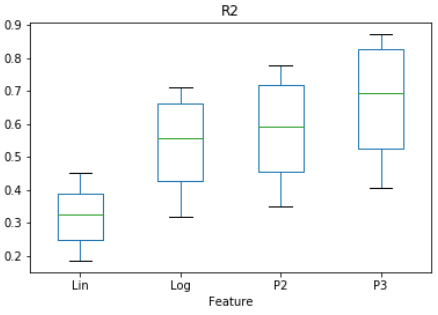

FIGURE 4.51: R^2 for MMSE Fit Models for Change

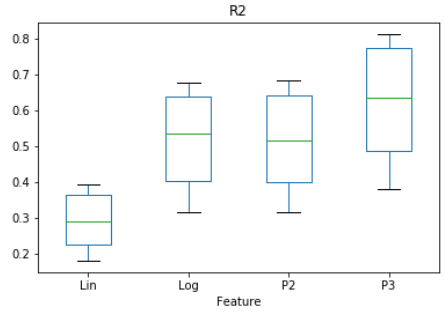

FIGURE 4.52: R^2 for MMSE Fit Models for Log Growth

4.3 Models Fit with Features

4.3.1 Least Mean Squares Adaptive Filter

Figures (4.43)-(4.58) are the MSE, MAE and R^2 error values for the LMS model for each of the combinations of all feature sets.

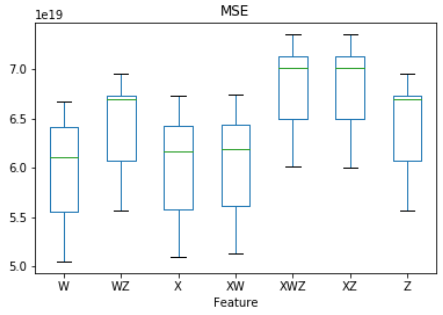

FIGURE 4.53: MSE for LMS Models for Change

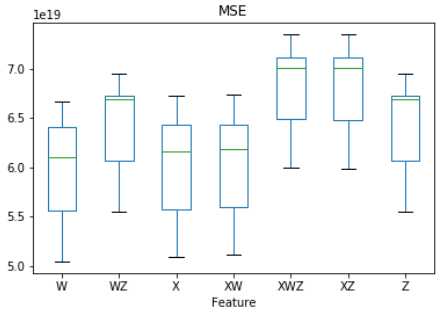

FIGURE 4.54: MSE for LMS Models for Log Growth

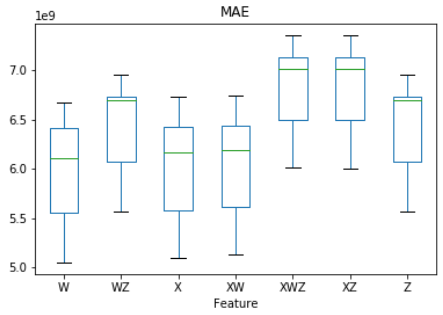

FIGURE 4.55: MAE for LMS Models for Change

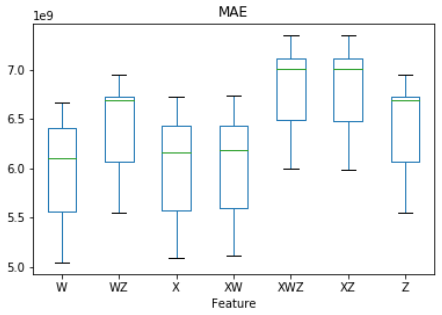

FIGURE 4.56: MAE for LMS Models for Log Growth

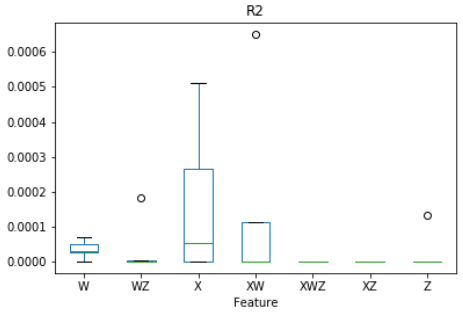

FIGURE 4.57: R^2 for LMS Models for Change

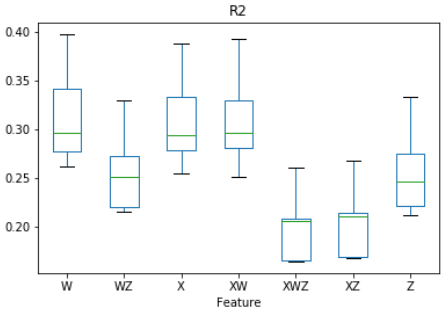

FIGURE 4.58: R^2 for LMS Models for Log Growth

4.3.2 Bayesian Ridge Regression

Figures (4.59)-(4.64) are the MSE, MAE and R^2 error values for the BRR model for each of the combinations of all feature sets.

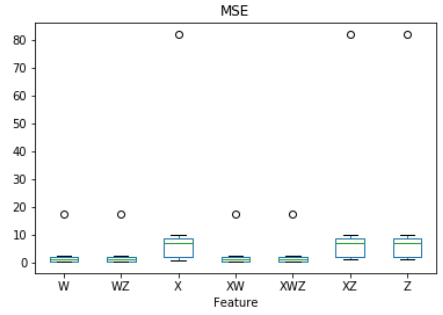

FIGURE 4.59: MSE for BRR for Change

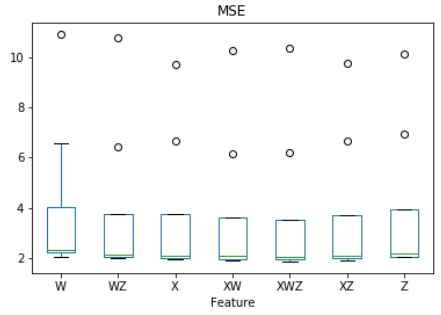

FIGURE 4.60: MSE for BRR for Log Growth

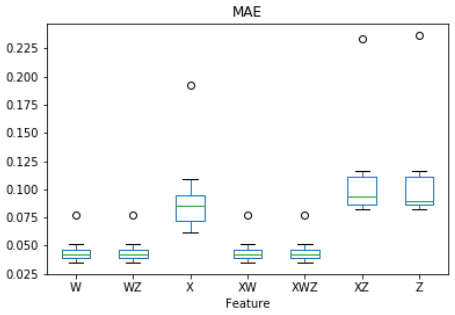

FIGURE 4.61: MAE for BRR for Change

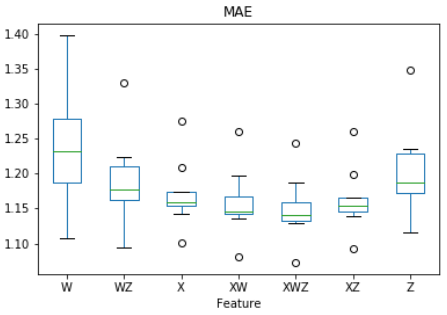

FIGURE 4.62: MAE for BRR for Log Growth

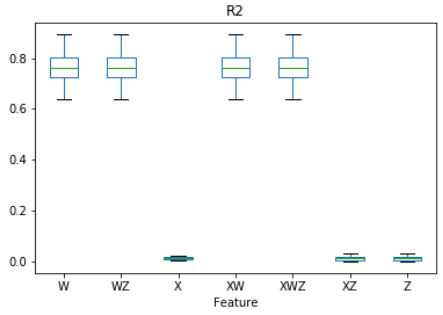

FIGURE 4.63: R^2 for BRR for Change

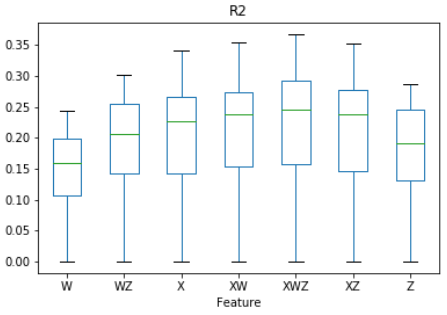

FIGURE 4.64: R^2 for for BRR for Log Growth

4.3.3 Random Forest Regression

Figures (69)-(74) are the MSE, MAE and R^2 error values for the RF Regression model for each of the combinations of all feature sets. Tables (4.9)- (4.12) are the importance for the different features for models trained for the individual feature sets and the combination of the all features. The sum of all of values is equal to one.

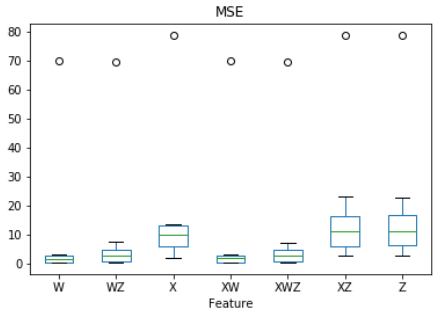

FIGURE 4.65: MSE for RF Regression for Change

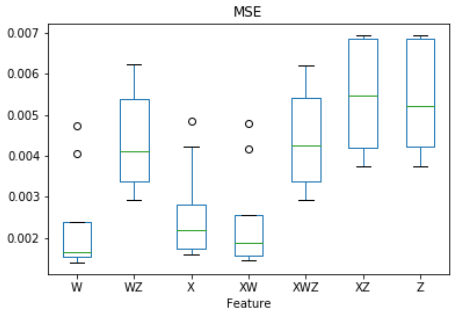

FIGURE 4.66: MSE for RF Regression for Log Growth

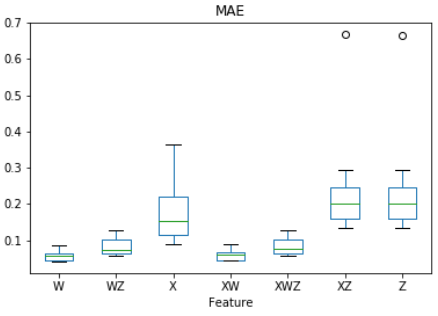

FIGURE 4.67: MAE for RF Regression for Change

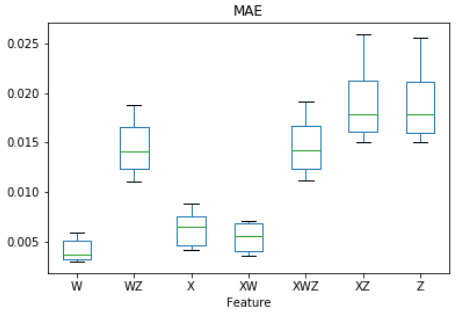

FIGURE 4.68: MAE for RF Regression for Log Growth

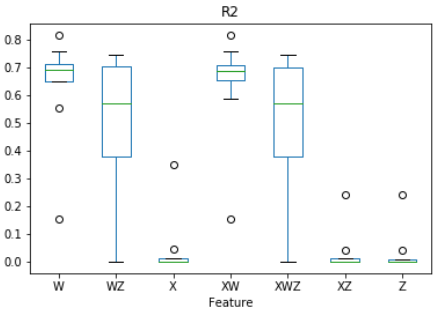

FIGURE 4.69: R^2 for RF Regression for Change

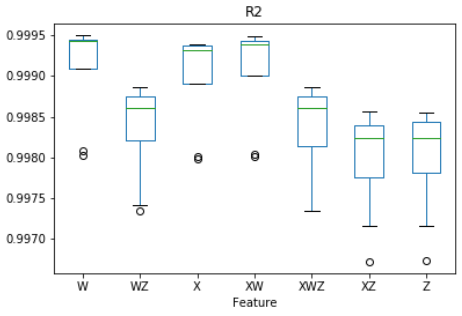

FIGURE 4.70: R^2 for for RF Regression for Log Growth

Feature	Output	JS1	NZ	NH	RO	Mean
Sentiment	Δy	0.061481317	0.05778648	0.047987214	0.048570561	0.053956393
Center Bias	Δy	0.005692892	0.076410575	0.006261107	0.002979955	0.022836132
Extreme Bias	Δy	0.037368463	0.062618023	0.014092948	0.052956582	0.041759004
F-K	Δy	0.000265287	0.001479341	0.00061466	0.005970085	0.002082343
D-C	Δy	0.02675865	0.023994172	0.033389432	0.068581083	0.038180834
$y(t)$	Δy	0.444604196	0.34871472	0.349523774	0.473235433	0.404019531
Time	Δy	0.423829195	0.428996689	0.548130864	0.347706301	0.437165762
Sentiment	$log(y(t))$	1.20E-05	6.60E-06	1.13E-05	1.73E-05	1.17961E-05
Center Bias	$log(y(t))$	5.04E-06	5.95E-06	6.09E-06	5.01E-06	5.5243E-06
Extreme Bias	$log(y(t))$	1.02E-05	6.41E-06	9.06E-06	1.14E-05	9.28058E-06
F-K	$log(y(t))$	1.21E-06	9.62E-07	2.57E-06	3.67E-06	2.10118E-06
D-C	$log(y(t))$	1.12E-05	8.35E-06	1.10E-05	1.81E-05	1.21811E-05
$y(t)$	$log(y(t))$	0.99989551	0.999924317	0.999902483	0.999847073	0.999892346
Time	$log(y(t))$	6.48E-05	4.74E-05	5.75E-05	9.74E-05	6.67708E-05

TABLE 4.9: Feature Importance for **X**

Feature	Output	JS1	NZ	NH	RO	Mean
Total Tweets	Δy	0.689766346	0.728155126	0.306918326	0.803256707	0.632024126
Retweets	Δy	0.160591749	0.151795155	0.465479779	0.021115398	0.19974552
$y(t)$	Δy	0.017695544	0.01166876	0.02826776	0.026842839	0.021118726
Time	Δy	0.009846429	0.0149956	0.032098019	0.025730677	0.020667681
Time on Net	Δy	0.012446783	0.010803062	0.020081969	0.020414393	0.015936552
Mean RT	Δy	0.010698464	0.008620961	0.019318124	0.013024517	0.012915516
Max Fav	Δy	0.016150483	0.005789791	0.011941012	0.009481393	0.01084067
STD RT	Δy	0.012501461	0.01109985	0.00866961	0.008608458	0.010219845
Mean Fol	Δy	0.010273154	0.006856204	0.013494926	0.010070023	0.010173577
Max Fol	Δy	0.008968901	0.007558515	0.012434809	0.011138496	0.01002518
$y(t)$	$log(y(t))$	0.996816914	0.997080966	0.991855889	0.99772981	0.995870895
Time on Net	$log(y(t))$	0.003052974	0.002803692	0.008028214	0.002132862	0.004004435
Time	$log(y(t))$	3.91E-05	3.44E-05	4.51E-05	4.11E-05	3.99292E-05
Total Tweets	$log(y(t))$	1.25E-05	1.10E-05	6.58E-06	1.28E-05	1.0695E-05
Min Fol	$log(y(t))$	9.99E-06	9.93E-06	9.67E-06	1.23E-05	1.04727E-05
Mean Fol	$log(y(t))$	9.00E-06	8.80E-06	9.50E-06	1.10E-05	9.58081E-06
Max Fol	$log(y(t))$	8.27E-06	8.12E-06	8.55E-06	1.17E-05	9.16662E-06
STD Fol	$log(y(t))$	7.64E-06	9.51E-06	4.66E-06	6.49E-06	7.07491E-06
Max RT	$log(y(t))$	5.98E-06	6.50E-06	4.80E-06	5.95E-06	5.80816E-06
Max Fav	$log(y(t))$	4.49E-06	4.23E-06	5.38E-06	6.01E-06	5.02558E-06

TABLE 4.10: Top Ten Important Features for **W**

Feature	Output	JS1	NZ	NH	RO	Mean
Time	Δy	0.255377694	0.244023959	0.382730175	0.259696739	0.285457142
$y(t)$	Δy	0.141180032	0.196709848	0.177231406	0.253297247	0.192104633
Sentiment_2	Δy	0.080440378	0.073777664	0.02622156	0.004286089	0.046181423
$KW10_1$	Δy	0.01026105	0.146431161	0.009316652	0.006754384	0.043190812
Extreme Bias_2	Δy	0.08395629	0.010115043	0.002707445	0.012089689	0.027217117
$KW3_4$	Δy	0.012218432	0.012422769	0.072539584	0.007435257	0.026154011
$KW4_1$	Δy	0.005953303	0.033046947	0.020763492	0.010922256	0.0176715
$KW6_1$	Δy	0.022817886	0.017344316	0.006126836	0.020338486	0.016656881
$KW8_1$	Δy	0.028123346	0.011330859	0.00850291	0.01838132	0.016584609
$KW9_1$	Δy	0.002723215	0.036542045	0.015527644	0.011202767	0.016498918
$y(t)$	$log(y(t))$	0.999454171	0.999444629	0.999373465	0.999497327	0.999442398
Time	$log(y(t))$	0.000215237	0.000229921	0.000260167	0.000200191	0.000226379
$KW9_1$	$log(y(t))$	8.50E-06	1.36E-05	1.25E-05	9.03E-06	1.09214E-05
$KW10_1$	$log(y(t))$	8.53E-06	1.27E-05	1.45E-05	7.53E-06	1.08058E-05
$KW8_1$	$log(y(t))$	7.98E-06	1.31E-05	1.35E-05	8.08E-06	1.06734E-05
$KW7_1$	$log(y(t))$	7.68E-06	1.15E-05	1.35E-05	9.37E-06	1.05137E-05
$KW4_1$	$log(y(t))$	7.19E-06	1.19E-05	1.20E-05	1.01E-05	1.02846E-05
$KW5_1$	$log(y(t))$	1.01E-05	1.20E-05	1.17E-05	7.20E-06	1.02416E-05
$KW6_1$	$log(y(t))$	7.83E-06	1.18E-05	1.18E-05	9.47E-06	1.0209E-05
$KW2_1$	$log(y(t))$	7.55E-06	1.12E-05	1.10E-05	8.13E-06	9.45894E-06

TABLE 4.11: Top Ten Important Features for **Z**

Feature	Output	JS1	NZ	NH	RO	Mean
Total Tweets	Δy	0.665212147	0.722414691	0.297268679	0.794844326	0.61993496
Retweets	Δy	0.154695497	0.15000089	0.45085842	0.020283182	0.193959497
Time	Δy	0.021510607	0.013258726	0.03159414	0.019032924	0.021349099
$y(t)$	Δy	0.013158808	0.007881039	0.0155098	0.01629721	0.013211714
Time on Net	Δy	0.009306778	0.008074881	0.015335144	0.014936488	0.011913322
Max Fav	Δy	0.014710623	0.005148384	0.010762947	0.008570719	0.009798169
Mean RT	Δy	0.007693576	0.006476862	0.015701937	0.00787984	0.009438054
STD Fav	Δy	0.010227993	0.005194996	0.011543016	0.006193961	0.008289992
STD RT	Δy	0.010528176	0.009136205	0.006591462	0.00620147	0.008114328
Mean Fav	Δy	0.010082427	0.006386463	0.008857753	0.006246101	0.007893186
$y(t)$	$log(y(t))$	0.996459711	0.996657626	0.991294424	0.997354057	0.995441455
Time on Net	$log(y(t))$	0.003013316	0.002807399	0.008100595	0.002166391	0.004021925
Time	$log(y(t))$	0.00019029	0.000203743	0.000233449	0.000171978	0.000199865
Total Tweets	$log(y(t))$	1.20E-05	1.07E-05	6.44E-06	1.23E-05	1.03733E-05
$KW10_1$	$log(y(t))$	7.78E-06	1.09E-05	1.25E-05	6.15E-06	9.33583E-06
$KW9_1$	$log(y(t))$	7.66E-06	1.09E-05	1.11E-05	7.06E-06	9.1808E-06
$KW5_1$	$log(y(t))$	8.77E-06	1.05E-05	1.05E-05	5.76E-06	8.89424E-06
$KW7_1$	$log(y(t))$	6.45E-06	9.70E-06	1.15E-05	7.79E-06	8.85201E-06
$KW8_1$	$log(y(t))$	6.54E-06	1.06E-05	1.16E-05	6.38E-06	8.75978E-06
$KW4_1$	$log(y(t))$	6.24E-06	1.04E-05	1.07E-05	6.32E-06	8.40874E-06

TABLE 4.12: Top Ten Important Features for **XWZ**

Chapter 5

Discussion

Table (4.1) shows that there is not necessarily a simple linear relationship between the number of articles published online and the number of Tweets that those article receive in the data set. There also does not appear to be a simple relationship between the articles published and the news sources reporting on them. It does illustrate that there are varying amount of interest in a story, based on topic and subject, with the top interest focusing around the Mueller Report and the New Zealand Shooting. Because of the global impact of those two events it follows that they would receive more attention in the news media and on social media.

The distribution of the readability of headlines for the D-C measure in figure (4.1) shows that almost all of the headlines are below a high school graduates reading level. This may suggest that this feature may not be informative for prediction with the chosen threshold. The F-K distribution in figure (4.2) is nearer to normal, with a mean that would be below high school graduates reading level, but enough variance that there would be headlines that would be considered above a high school graduates reading level. The sentiment distribution in figure (4.3) is not uniform, or dominated by neutral sentiment, which may suggest that the headlines collected are not necessarily objective, balanced and purely informative, but may be written in a way that is meant to persuade or appeal to the emotion of the reader. There were approximately 38.47% missing labels for sources, but of the articles only 28.48% were missing their source bias labels. This is likely from the independent sources that are not common enough to have received a score from the two services or fit with the alternative labeling methods. There are more non-centered and non-extreme sources overall in figure (4.4), which likely play a role in frequency differences we will evaluate in a later section of this study.

Figures (4.5) - (4.7) illustrates the general shape of the two outputs that were being predicted. We can see that the shape of the total volume of tweets is similar to a population growth curve, and that the log transformation of the growth is not as linear, as the point of saturation occurs rather quickly. The article did not gain traction on Twitter early in the day, which follows considering the man was shot near midnight of the first day. There does not appear to be a long delay for this article to reach Twitter. The time it took for this article to reach maximum growth was around a day.

Figures (4.8)-(4.10) are examples of one of the more consistent articles, with the growth taking a longer time to saturate on the network. We can also see that the saturation point was much lower than the first article. This article did appear to reach Twitter on the day that it was published online, and continued to receive Tweets consistently through the following day. Then it was took approximately five and three quarter days to reach maximum growth.

Figures (4.11)-(4.13) show an article that receive attention for about a day, but did not reach Twitter for about 2 and a half days. It was published on March 29th, which was two months after the initial incident. This article event was dynamic with multiple developments over time, especially compared to the other two stories that the other figures represented. These show that there is no general way that the articles are being published on Twitter, and illustrates the challenges in developing a simple method for prediction based on a set of features. With the differences in feature sets the ML models should perform better than a simple linear method.

Reviewing figures (4.14) - (4.23) we can see that there is not a clear one to one relationship between articles published online and articles that were published on Twitter, even considering that the graphs for Twitter were 30 days longer than the graphs for the volume of articles published on the Internet. They all generally follow an exponentially decaying process, one that appears to be similar to the Hawkes Process, but it does not appear that the key driving factor for Twitter is the amount of articles published online. We would expect that developments in the story would cause spikes in volume of articles published online and the amount of exposure the event is getting on Twitter. The Mueller Report Tweet volume best illustrates the lack of staying power for articles on Twitter. A couple days after 60 we see a sharp drop off in volume of Tweets. We also see that in every other event, though there is not as a significant of a decline as there is for the Mueller Report. For Rep. Omar there was practically no Tweets after 60 days.

Comparing Table (4.2) to (4.3) we see that articles published on the internet are not dominated by any particular source, but there are important sources on Twitter. This may suggest that the prestige of a particular source is significant in developing the model for prediction. There was no measure of prestige of a source in this study, and that will be discussed in a later section of this study. This would suggest that sources are important factors when predicting spread of information on Twitter, as previous research has found. It also suggests that source bias may be informative in the prediction process.

Tables (4.4) and (4.5) show that there is not much difference in the keywords that are being written about online and being shared on Twitter. This would suggest that the news media is in sync with the tastes of social media users.

The frequency of the features through time shown in figures (4.24) - (4.33) are considered complements, for readability and sentiment, but bias is missing the frequency of the missing sources. In the case where any single day does not sum to one, the missing portion of the sum of the frequency of the missing source bias labels for the day. The figures chosen are best and worst examples of the dynamics of the frequencies through time, compared visually. The F-K measure was more dynamic through time compared to the D-C measure, as we would expect from the distributions seen in the earlier figures. The center bias measure does not vary as much as the extreme bias measure, as we should expect from the distribution that was heavy for lean-left, and there is a low volume of sources with extreme bias labels.

The correlation of the features from each data set were also examined to understand the potential for linearity of the features to the two values of interest through figures (4.34) - (4.39). There were few features throughout all of the considered features sets, **X**, **W** and **Z** that showed potential for a linear relationship through their correlation. Bias measures had the highest correlation to the log transformation of the growth of the Tweet. The change in Tweets over time had the best correlation with past tweeting activity. The correlation of article bias and past bias could be explained by examining the differences in source distribution on Twitter compared to the internet as noted by tables (7) and (8). The correlation between change in time and the past history of total Tweets, Retweets and original Tweets follows natural from the fact that there should be a higher likelihood that a user would Retweet a Tweet they recently saw. It also supports the idea that information will spread when there is more exposure to it.

5.1 Models Based on Time Alone

Table (4.6) illustrates that simplifying the problem into a prediction problem based solely on previous growth and change in time is not a feasible approach to the problem. The nature of the two different outputs, change vs. growth are also illustrated through this table. The log transform of the growth performed better overall in all cases. One reason for the lower relative error for the entire dataset is that many of the articles did not spread, so minimizing the error across those would be a constant zero, which would result in no error at all. The better measure is the subset of articles that were published on Twitter. In every case that subset performed worse than the entire set. Despite that, similar trends are seen with both sets in table (4.6), where the best fit was a third-degree polynomial. This is counter to what was originally thought, given the nature of the growth. The log transform was expected to be linear in nature and the linear fit would have been expected to perform the best for that set, but the rapid change was not modeled well with the linear function. There was no expected good fit for the change, as that is just a set of non-periodic discrete weighted impulses through time for most articles.

Because not all articles that published on Twitter were published at the first time period of the day they were published online, Table (4.7) reported the MSE for the general regressors for that case. Overall all models performed better, which suggests that fitting the initial spike of activity within the network is a difficult task for the general regressors, based solely on time.

Figure (4.40) illustrates that most articles reach there maximum exposure relatively quickly within the month period. The mean value for all the articles that were published on Twitter was approximately 4.08 days, with a standard deviation of approximately 6.85 days. That suggests that the MMSE fit models would do better for shorter duration of time. Table (13) verifies that as the shortening the duration of time results in better error, for all of the fit types. Reducing the days to just under 21, we see a marked decrease in the error across all fits for the change prediction, and MSE's below one for log growth prediction. For a duration of approximately 10.42 days, we see error values decrease by more than half. Restricting the duration to 2.75 hours we see very low error, but this does not achieve the desired duration of prediction. Likely there is little activity during a duration this small and the low error is a product of flat or little growth, where the optimal fit line is a constant value. The similar error for the linear and second and third degree polynomial fits suggest that this is the case.

Examining the MSE, MAE and R^2 values in the box plot figures from figures (45)-(56) we see that despite having somewhat lower values for MSE, the range for the error is large. This is most evident in the R^2 scores, where we see that the exponential fit has a score of zero, and other fits, such as linear may vary by 0.2 or more.

5.2 Models Based on Features

5.2.1 Least Mean Squares

Figures (4.53) - (4.58) illustrate that this was the worst of the three methods to predict growth and change based on the features chosen. It performed worse overall, even in comparison to the MMSE fit regressors based on previous growth and change. The general shape of the graphs shown in figures (57) - (62) would explain why this method would not prove to be a good prediction method for finding growth or change. The results do illustrate the the best solo feature set was **W**, and the worst was the **Z** feature set. In general adding in more features from these sets did not results in any notable improvement. The error for the growth and change were similar from MSE and MAE, but the R^2 values vary, showing that the growth was a much better fit that the change. Because the MSE were similar, this result is due to the fact that the variance in lower for the growth models in general in comparison to the variance in the change. This follows as the range for the log growth would always be lower than the change.

5.2.2 Bayesian Ridge Regression

Figures (4.59) - (4.64) show that this method of prediction proved to be better than the LMS Adaptive filter, but was not the best overall. It was the best for predicting the change, yet the log growth performed better than the change. This should be expected as the log growth is more linear in nature compared to the change, and this learner is based on linear function assumptions. We could argue that **W** provides the best performance for the change, and **X** provides the best performance for the log transformation of growth, from looking at MAE. **Z** performs worse in both cases. We do see improvement when combining the predictors sets, where the change prediction is driven by combinations with **W** features, and **X** decreases error for combination of the log transformation of growth. The R^2 values are low for everything outside of combinations of **W** for change.

5.2.3 Random Forest Regression

Figures (4.65) - (4.70) illustrate that this method provides the best prediction for the log growth. Again the frequency features alone proved to be a poor set of features. Looking at the prediction of change we see that **W** drives the error down the most, especially when coupled with **X**. Looking at the R^2 values we see that **W** alone is better than any other combination at all. Looking at the error reported by the log growth we see that **W** again is the best feature set. Overall the error measured for this method were among the best of the three methods examined.

Table (4.9) illustrates that time and Tweet total are also dominate features for this learner, where Tweet total almost completely dominates the decision for growth prediction. We see that exposure dominates for **W** in table (4.10), where change is more dependent on the current amount of Tweets on the network for that time period, and growth is more dependent on the total amount of Tweets that are on the network over all time. In table (4.11) we see similar trends where the only features that seem to be informative is the time and $y(t)$. In table (4.12) we see that this feature set is dominated by features from **W** again, and there is little change from that feature set and the whole feature set. The small drop in the importance values follows given that all values must sum to one and there are many more features for the entire feature set, compared to any single set alone.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Overall we can see that the frequency features did not improve the error. Alone it proved to be one of the worst set of features out of the three. The best set was the dynamic measure of Twitter itself. Considering that anywhere from one third to almost half of the articles published online do not make it to Twitter, it is understandable that the frequencies of the Internet would not predict activity on Twitter well. Going forward it would be better to measure the frequencies as they appear on Twitter. From the perspective of linear models the results presented follow from the fact that the features that were most correlated were in the $\mathbf{W}$ feature set. Likewise the $\mathbf{Z}$ shows some of the worst correlation to both change and the log growth. The improvement with the combinations of the features was expected, but the small improvement may be offset by the increase in complexity of the problem.

Not being able to see any improvement in the error from the addition of the feature set in the empirical example may be due to factors and simplifying assumptions that were made in the process. Not considering the nature of the structure of the graph, and the underlying algorithms that drive user engagement, may have lead to errors in feature selection that were meant to encapsulate the selection process. Simplifying the articles features into binary classes may have also not accurately captured the importance of the frequencies distribution and selection process for articles. Developing a better method to maintain the temporal aspect of the data when training the machine learning models may also see improvement in the prediction. That may be accomplished in either collecting data for a longer period of time, in order to have enough samples to train on single articles, then transfer the weights as the initialization of weights for the next training

period. One issue with that is the typical article reaches saturation rather quickly, so capturing long term trends remains a challenge.

6.2 Future Work

Going forward it would be beneficial to understand the structure of the network, and the method that Tweets are presented to the users. Being able to model the network as a whole and label the users with basic data, such as follower lists, followed lists and time zone would help inform on the selection process better. Changing $\mathbf{Z}$ to the frequencies of articles that are on Twitter may prove to better predict that the articles that are never published on Twitter. This may be estimated by building the binary classifier that predicts if the article will be published on Twitter based on article features. Then the output of the frequencies can be determined from the articles that are predicted to be published on Twitter. Adjusting parameters for the learning models and using feature selection methods for the feature sets may see improvement in the error. Developing a better way to maintain the temporal properties of the data while training may also improve the model performances.

Appendix A

Appendix

A.1 Data Set Description

The data set that was collected was centered on five different news stories that were thought to be popular on social media. Four of the five topics were considered to be polarizing in the political realm, meaning likely different bias news sources would report on the stories in different ways, which should show a different spread on the network, assuming that there is an equal distribution of users on the network. An interesting note, is during the filter review process, there were multiple occasions in which the events crossed. This suggests that these stories were important enough to be linked to or referenced by the other stories. This may suggest that the importance is either driven by the news editor at the broadcast media or by the media consumers, who drive the news cycle. Robot and Subscribe were used as blanket terms, to catch stories from Bloomberg, which had a splash page to catch automated crawling services, and Financial Times, which had a landing page for subscriptions.

A.2 Jussie Smollett

Jussie Smollett, a homosexual, African American actor, filed a police report about an attack that would be classified as a hate crime. The report stated that at 2am on January 29, 2019, in Chicago, Mr. Smollett was attacked by two white men who called him different slurs, referring to his race and sexual orientation, and also made references to support of the current President. They poured an unknown substance on Mr. Smollett and hung a noose around his neck during the attack. On February 20th Mr. Smollett was charged with disorderly conduct for allegedly paying two Nigerian brothers to stage the attack. A deal was reached by Mr. Smollett's defense team and the state on March

26th, and charges against him were dropped. The following day the FBI announced that an investigation into why the charges were dropped was to begin. On April 12 the city filed a lawsuit against Mr. Smollett for the cost of the investigation. The initial public response was polarized with both sides holding firm throughout most of the unfolding story [27]. Data was collected from March 14 to May 14. The final collection date for the Twitter activity was June 13th.

The keyword used to generate the list was "Smollett". The keyword list used in the filter:

[robot , Subscribe , Juicy , Jussie , attack , smolletts , smollet , jussie , arrest , justice , smollett , Smollett , police , Police , Empire , empire , crime , Foxx , hate , charged , charges , hoax , alleged]

Tangential stories that surrounded and may have been included in this discussion could have ranged from the future of Mr. Smollett's role in his show Empire , connections to influential people who helped guide the eventual countersuit in Mr. Smollett's favor. From Figure (4.14) we can see that there were multiple spikes in the volume of reports over the days. This is most likely attributed to the continuing development of the story over the collection time period.

Possible discussions that surround this event were; the impact and prevalence of white nationalism, hate crimes, homophobia and racism in the United States, presumption of innocence, classist judicial systems, and consequences to reporting false crimes.

A.3 New Zealand Shooting

On March 14th, in Chirstchurch, New Zealand, two gunmen entered two different mosque's and murdered 51 Muslim worshippers, and injured 49 more. The shooting was live-streamed over Facebook Live. The media described the two shooters as alt-right and white supremacists. Along with the discussion about white supremacy and the global impact that the ideology is having, it also sparked a response from the New Zealand government to tighten existing gun laws. One of the weapons that was used was an AR-15 style rifle, which is a common rifle that many people in the United States who are lobbying for increased gun laws focus on. The manifesto published by one of the shooters included anti-immigration sentiment, as well as white supremacist references,

and calls to memes connected to Muslim genocides, Norwegian terrorists and the President of the United States. Some media sources reported it as a trolling effort [28]. Data was collected from March 14th to May 14th. The final collection date for the Twitter activity was June 13th.

The keywords used to generate the list were: "New Zealand, shooting, mosque." The keyword list used in the filter:

[robot , Subscribe , Zealand , mosque , firearms , synagogue , mass , bomb , zealanders , coast , guard , ideologies , sri , Sri , lankan , Lanka , lankas , thoughts , prayers , deadly , right , extreme , islamophobia , NZ , terror , guns , gun , killer , supremacists , muslims , massacre , nationalist , shooter , racist , terrorism , crime , violent , zealand , zealands , nationalism , christchurch , hate , shooting , muslim , Muslim , Christchurch , supremacy , shooting]

Tangential stories that were included were centered around this event and may have been included, include a San Diego attack, a Sri Lankan attack, and the arrest of a Coast Guard officer who was planning an attack. As with the rise in reporting that were seen with the Jussie Smollett story, there were spikes in the articles written over time when the other stories related to the shooting broke. Again this is visible in Figure (4.16).This story did not have continuing coverage like Smollett, so there were less days in which we saw a spike, the spikes were instead centered around the tangential stories.

The potential discussion points centered around gun control and white nationalism. This was reported as an action of white nationalism, which was a trending topic in the United States public discussion at the time. It also had aspects of impacting the gun control debate. I did not include articles that would necessarily describe the role of social media in filtering out content. That was a diversion that I did not explore as it did not fit the general narrative of gun control , terrorism and white nationalism. It seemed to be a marginally larger discussion that spawned from this event, as there were a larger trend in the filtered articles that were concerned with this topic.

A.4 Mueller Report

Robert Mueller was in charge of a two plus year investigation into Russian interference in the 2016 election. At the conclusion of his investigation he released a report outlining his findings. It included information on grand jury proceedings, as well as information that was classified, and gave insight into national security methods. For that reason it was reported that a redacted version would be made available to the public, and to Congress. Mr. Mueller's investigation garnered a lot of publicity from media outlets, as some reported that it would lead to the evidence necessary to impeach the President. The report was submitted to the Attorney General on March 22, 2019, and publicly released on April 18, 2019. On May 1, 2019, AG Barr testified. The release of the report and the subsequent summary from the Attorney General, gained a lot of traction in the media, and throughout social media, in response to the ongoing online discussion about impeachment of a President that had mixed support from the general public, and a reported bad public image in the rest of the globe [29]. Data was collected from March 22 to May 22. The final collection date for the Twitter activity was June 22.

The Keywords used to generate the list include: mueller report, Mueller report and Barr. The keyword list used in the filter:

[general , Russian , attorney , crisis , constitutional , robot , subpoenas , manafort , impeach , rosenstein , rod , investigations , Subscribe , postmueller , exonerate , exoneration , russiagate , trumprussia , trump , AFP.com , Bill , Barr , barrs , William , william , intelligence , burger , comey , Comey , Burger , nadler , Nadler , russia , Russia , schiff , theory , investigation , muellers , conspiracy , muellers , findings , muellerrelated , committee , collusion , report , barr , Barr , Mueller , president , mueller , robert , Robert , obstruction , impeachment]

There were not many tangential stories, as this was relatively focused on the report and the coverage of the report and the people's responses and action to the report. Figure (4.18) shows spikes in reporting centered around important dates mentioned above. The largest spike in articles came on April 18, with over 3000 articles collected that day. This was the only event that consistently had large number of articles throughout the collcetion period.

Potential discussion points included impeachment proceedings, the OLC ruling about not charging a sitting President, obstruction of justice, both parties roll in the 2016

election and ties to Russia, general actions of the current and past administrations.

Data began when the report was released, but the overarching story is something that was ongoing at the time of the collection

A.5 Nipsey Hussle Shooting

Ermias Joseph Asghedom, known as rapper Nipsey Hussle, was murdered in front of Marathon Clothing on March 31st. Outside of his music career he was also a business owner, actor and community activist. He began his music career in 2005, and had his last release, his debut studio album in 2018. His music denounced gun violence and spoke about how gang culture influenced his life. He gave money and time back to the community, and started a co-working space that not only was targeted for work, but was also meant to provide education for children in the community. He planned a meeting with LAPD to discuss prevention of gang violence on April 1, 2019, the day after he was murdered. Another rapper claimed credit for the murder, but police arrested Eric Holder, who possible had ties to the rapper claiming credit, on April 2, 2019 [66]. Many people showed up to his memorial, and there were altercations reported in the wake of his death, that may have been in connection to his murder. Data was collected from March 31 to May 31. The final collection date for the Twitter activity was July 1.

There were a few articles that were filtered out that were related to entertainment in general, whose artists could reasonably be related to the late rapper, mostly through the genre of rap. The filtered results removed those articles. The reason behind removing those was to preserve the main subject. This was done to remove general topic concerning entertainment and sports.

The keyword used to generate this list was "Nipsey Hussle". The keyword list used in the filter:

[]robot , Subscribe , Nipsey , nipsey , death , hussle , Hussle , rappers , rapper , shot , killed , hussles]

Most tangential stories were removed. Overall the volume of stories for this event was small. The initial peak surrounded his murder, and the second peak may be attributed

to reports about the suspect who was arrested and charged with the murder, as well as stories about the memorial services. There were also random peaks that may not necessarily seem to be related to anything that was noted about the subject. Figure (4.20) shows a drop off in reporting very soon after peak news days.

Data Collection began at the beginning of the story. It was not initially believed to be polarizing, but a large volume of results on social media news feed early on were noted, so collection of this event began.

A.6 Comments by Rep. Omar at a CAIR conference

March 23rd, 2019 at a Council on American-Islamic Relations meeting, Rep. Ilhan Omar was talking about the response of the US to the attacks of 9/11 as they were experienced by the Muslim-American community. She was speaking of the impact of the event to the general community. In response to a sound byte from the speech, there was general blowback from conservative opponents, including but not limited to the President of the United States. Although the comments may be seen as the center of the story, this collection was meant to capture the back and forth actions and reactions between the two parties. Data was collected from April 10 to June 10. The final collection date for the Twitter activity was July 11.

The keywords used to generate the list are "ilhan omar, 9/11". The keyword list used in the filter included:

[Robot, islamophobia, Subscribe, Omar, 9/11, Ilhan, ilhan, tlaib, omars, 911 , omar , Holocaust , Tlaib , 9 , 11]

Tangential stories that were included were centered around this event may include comments by Rep. Tlaib about the Palestinian people and their role in the Holocaust, as a means to provide respite for the displaced Jewish peoples. This story generated very little traction in the news media, and there were days where there was not reporting at all. The fall off is quite sudden, as shown in figure (4.1), and despite being centered around a potential Twitter feud between government officials did not garner much different activity levels on Twitter.

A.7 Methods Used in This study

Four different methods will be explored for the prediction of Tweets over time, general MMSE fit regression; including linear, polynomial, exponential and logarithmic fits. These will server as a baseline models, where the features fro prediction are the historically observed output. The output will also be predicted with three separate methods adaptive LMS filtering, BRR and RF.

A.7.1 Least Means Squares Adaptive Filter

The LMS filter is an adaptive filter that seeks to minimize the MSE of the estimation iteratively by adjusting weights as each time step, based on the current error at that time step. It is a two step process, guided by the following equations

$$y[n] = \mathbf{x}[n]\mathbf{w}[n]$$

$$e[n] = d[n] - y[n]$$

$$\mathbf{w}[n+1] = \mathbf{w}[n] + \mu e[n]\mathbf{x}[n]$$

where $y[n]$ is the estimate at time n, $\mathbf{x}[n]$ is the observations at time n, $\mathbf{w}[n]$ are the weights for time n, calculated at the previous time step, $d[n]$ is the desired output, and μ is the step size for the process.

For each time step the equations are solved in the given order to prepare for the next time step. Some design parameters include the initialization of the weights and the value of μ, either as a constant scalar value, or a function of other variables.

A.7.2 Bayesian Ridge Regression

BRR is a variation of ridge regression, which is a modification of the solution to the OLS problem. We can consider the general linear model

$$y = \beta x + \epsilon$$

where ϵ is some unknown noise and β is a scalar value that relates x to y. Extending the case into multiple regressors and observations we can build a matrix notation of the

problem and estimate the output of the system as

$$\mathbf{Y} = \mathbf{X}\beta$$

where $\mathbf{Y}$ is a n by 1 matrix, $\mathbf{X}$ is a n by p+1, and β is a p+1 by 1 matrix, where n is the number of observations and p is the number of regressors. The first column of $\mathbf{X}$ is a column of 1's to consider any bias in the equation. We then can find the solution for the β values by solving an optimization problem for a cost function. Consider the square error cost function and the optimization problem from it

$$C(\beta) = ||\mathbf{Y} - \mathbf{X}\beta||^2$$

$$\hat{\beta} =_{\beta} C(\beta)$$

In general the solution would be of the form

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

One assumption in this model is that all the features in $\mathbf{X}$ are linearly independent, in order to find a unique solution for $\hat{\beta}$. If we wish to relax that assumption on some level, and consider some level of collinearity between the features, we can adjust the model by introducing bias or variance to account for prediction error from the OLS solution in these cases. In the case of Ridge Regression a small amount of bias is introduced to improve the performance of the method. We would still consider the solution as the optimization of a given cost function, but modify the cost function to emphasize weights with lower values, and penalize weights with large values. The new cost function is given as

$$C_1(\beta_1) = ||\mathbf{Y} - \mathbf{X}\beta_1||^2 + \lambda||\beta_1||^2$$

The solution then becomes:

$$\hat{\beta_1} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$$

Here λ is the regularization factor. We see that if we allow $\lambda = 0$ the solution reduces to the OLS solution. This allows for a solution in the case $\mathbf{X}^T\mathbf{X}$ is not invertible. To chose λ, cross validation is commonly used. Another approach is assume the solution is the mean of a posterior distribution such that $\beta \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda})$ for each weight in the vector. [30]

A.7.3 Random Forests

RF Regression is built from the concepts of a decision tree and bagging. A decision tree is a learning method that creates subsections of the feature space in order to isolate classes or outputs of the system. A random forest will generate a set amount of trees based on a i.i.d. randomly generated vector. The final output of the learner is based on a vote from all the trees in the forest. The argument against overfitting for this case relies on the Strong Law of Large numbers, which would suggest that for this method to work a sufficiently large number of trees must be grown. The element that makes this process random can be designed based on a variety of ways. Some common methods include bagging, split selection, and selection of training set. Bagging is based on a random selection without replacement made from the samples. Split selection randomly selects the split point of a tree based on the K best split points. Randomly selecting the training set can be done based on randomly assigning weights to the examples from the training set and selecting those examples [31].

Bibliography

[1] Wang S. Han E. Ramakrishnan N. Keneshloo, Y. Predicting the popularity of news articles. pages 441–449, 2016. doi: 10.1137/1.9781611974348.50.

[2] Cambazoglu B. Arapakis, I. On the feasibility of predicting popular news at cold start. *Journal of th eAssociation for Information Science and Technology*, 68:1149–1164, 2017.

[3] Fox E. Zaman, T. and E. Bradlow. A bayesian approach for predicting the popularity of tweets. *The Annals of Applied Statistics*, 8:1583–1611, 2014.

[4] Xie L. Sanner S. Cebrian M. Yu H. Henteryck P. Rizoiu, M. Expecting to be hip: Hawkes intensity processes for social media popularity. 2017. URL https://arxiv.org/abs/1602.06033v8.

[5] E Shearer. Social media outpaces print newspapers in the u.s. as a news source. pew research center. 2018. URL https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/.

[6] Omnicore. Twitter by the numbers: Stats, demographics & fun facts. 2019. URL https://www.omnicoreagency.com/twitter-statistics/.

[7] Sakurai Y. Prakash B. Li L. Faloutsos C. Matsubara, Y. Nonlinear dynamics of information diffusion in social networks. *ACM Transactions on the Web*, 11, 2017. doi: 10.1145/3057741.

[8] Yang J. Ye X. Xu H. Trivedi R. Khalil E. Li S. Song L. Zha H Farajtabar, M. Fake news mitigation via point process based interventions. URL https://arxiv.org/abs/1703.07823.

[9] Tabibian B. Oh A. Schoelkopf B. Kim, J. and M. Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. URL https://arxiv.org/abs/1711.09918.

[10] Guerraoui R. Kermarrec A. Maurer A. Pavlovic M. Balmau, O. and W Zwaenepoel. Limiting the spread of fake news on social media platforms by evaluating users' trustworthiness. URL https://arxiv.org/abs/1808.0992.

[11] Ozdaglar A. Acemoglu, D. and A. ParandehGheibi. Spread of (mis)information in social networks. *Games and Economic Behavior*, 70:194–127, 2010. doi: 10.1016/j.geb.2010.01.005.

[12] K. Kumar and G Geethakumari. Information diffusion model for spread of misinformation in online social networks. *Paper Presented at 2013 International Conference on Advances in Computing, Communications and Informatics*, 2013. doi: 10.1109/ICACCI.2013.6637343.

[13] Kumar K. and Geethakumari G. Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences*, 4, 2014. doi: 10.1186/s13673-014-0014-x.

[14] M. Carlson. Embedded links, embedded meanings. *Journalism Studies*, 17:915–924, 2016. doi: 10.1080/1461670X.2016.1169210.

[15] L. Lee, C. S. Ma. News sharing in social media: The effect of gratifications and prior experince. *Computers in Human Behavoir*, 28, 2011. doi: 10.1016/j.chb.2011.10.002.

[16] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. *In Proceedings of the SIGCHI conference on human factors in computing systems*, 2007.

[17] D. Chua A., Goh and Lee C. Mobile content contribution and retrieval: An exploratory study using the uses and gratifications paradigm. *Information Processing Management*, 48:13–22, 2012.

[18] Sakurai Y. Prakash B. Li L. Faloutsos C. Matsubara, Y. Understanding news sharing in social media. *Online Information Review; Bradford*, 38:598–615, 2014. doi: 10.1108/OIR-10-2013-0239.

[19] de Amorim M.D. Fdida S. Tatar, A. and P. Anatonidis. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, 5, 2014. doi: 10.1186/s13174-014-0008-y.

[20] H.S. Kim K. Lee M.Y. Choi J. Ko, H.W. Kwon. Model for twitter dynamics: Public attention and time series of tweeting. *Physica A*, 404:142–149, 2014.

[21] Stack Overflow. Count the number of syllables in a word. 2013. URL https://stackoverflow.com/questions/14541303/count-the-number-of-syllables-in-a-word.

[22] LLC. Media Bias Fact Check. Methodlogy. 2019. URL https://mediabiasfactcheck.com/methodology/.

[23] AllSides. How allsides rates media bias: Our methods. 2019. URL https://www.allsides.com/media-bias/media-bias-rating-methods.

[24] E. Hutto, C. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the 8th International Conference on Weblogs and Social Media.*, 2015.

[25] C Matous. Padasip. 2016. URL http://matousc89.github.io/padasip/.

[26] S. Ronaghan. The mathematics of decision trees, random forest and feature importance in scikit-learn and spark. *Toward Data Science url =*, 2018.

[27] Wikipedia contributors. Jussie smollett alleged assault. in wikipedia, the free encyclopedia. 2019. URL https://en.wikipedia.org/w/index.php?title=Jussie_Smollett_alleged_assault&oldid=925143721.

[28] Wikipedia contributors. Christchurch mosque shootings. in wikipedia, the free encyclopedia. 2019. URL https://en.wikipedia.org/w/index.php?title=Christchurch_mosque_shootings&oldid=926488036.

[29] Wikipedia contributors. Mueller report. in wikipedia, the free encyclopedia. 2019. URL https://en.wikipedia.org/w/index.php?title=Muelle\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{_\global\mathchardef\accent@spacefactor\spacefactor}\accent23_\egroup\spacefactor\accent@spacefactorReport&oldid=925378317.

[30] Holmström L . Pasanen, L. and M. Sillanpää. Bayesian lasso, scale space and decision making in association genetics. *PloS one*, 10, 2015. doi: 10.1371/journal.pone.0120017.

[31] L . Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324.