Department of Economics, University of Warwick

Monash Business School, Monash University

as part of

Monash Warwick Alliance

Do Political Actors Engage in Strategic Deception on Social Media?

Simon Ricketts

Warwick-Monash Economics Student Papers

September 2021 No: 2021-16

ISSN 2754-3129 (Online)

https://warwick.ac.uk/fac/soc/economics/research/wmesp/

The Warwick Monash Economics Student Papers (WM-ESP) gather the best Undergraduate and Masters dissertations by Economics students from the University of Warwick and Monash University. This bi-annual paper series showcases research undertaken by our students on a varied range of topics. Papers range in length from 5,000 to 8,000 words depending on whether the student is an undergraduate or postgraduate, and the university they attend. The papers included in the series are carefully selected based on their quality and originality. WM-ESP aims to disseminate research in Economics as well as acknowledge the students for their exemplary work, contributing to the research environment in both departments.

"*We are very happy to introduce the Warwick Monash Economics Student Papers (WM-ESP). The Department of Economics of the University of Warwick and the Economics Department at Monash University are very proud of their long history of collaboration with international partner universities, and the Monash Warwick Alliance reflects the belief in both Universities that the future will rely on strong links between peer Universities, reflected in faculty, student, and research linkages. This paper series reflects the first step in allowing our Undergraduate, Honours, and Masters students to learn from and interact with peers within the Alliance.*"

Jeremy Smith (Head of the Department of Economics, University of Warwick) and Michael Ward (Head of the Department of Economics, Monash University)

Recommended citation: Ricketts, S. (2021). Do political actors engage in strategic deception on social media? *Warwick Monash Economics Student Papers* 2021/16

[1] Warwick Economics would like to thank Lory Barile, Gianna Boero, and Caroline Elliott for their contributions towards the selection process.

Do Political Actors Engage in Strategic Deception on Social Media?

Simon Ricketts*

September 2021

Abstract:

We examine whether political actors engage in strategic deception on social media. We find evidence that certain groups of politicians engage in deception in response to an election. To infer deception, we construct a novel wealth inference model from text of political social media accounts. We use machine learning and natural language processing, which is accurate to within half an order of magnitude when compared to real wealth disclosures as required by law in the United States. Wealth exaggeration is not homogenous; in an election year, the wealthiest political actors minimise their perceived wealth, while the poorest exaggerate their perceived wealth. We do not find evidence that there are differences in exaggeration due to sex, party or experience.

JEL Classification: C55, D72

Keywords: Strategic deception, wealth-inference, machine-learning, natural language processing, social media, election

* Contact: ricketts.simon1@gmail.com

I would like to thank Simon Angus as supervisor for this paper.

Chapter 1

Introduction

The underpinnings of democracy is the contest for representation and power. Who voters choose, collectively, can change the direction of policy and politics. These political contests are high stakes, both for candidates and their constituents. For candidates in US elections, being an elected member of Congress means that they can influence legislation and directly affect constituent outcomes throughout their tenure. But what of the voters? How do they make up their minds and to what extent does media influence their decisions? The traditional role of media has been to inform the public, including on political issues (Dautrich and Hartley, 1999; Owen and Hughes, 1991). Opinion and analysis was limited to a small number of political analysts (Gerstl-Pepin, 2007; Fraser, 1990). Contrasted with the current era of peer-to-peer communication and social media, voters are presented more information, more frequently, by a wider variety of sources, including from politicians. Social media has lowered barriers to entry for production and consumption of media and at the same time, the media environment is becoming increasingly polarised (Bail et al., 2018; Hong and Kim, 2016).

This paper investigates whether political actors engage in strategic deception on social media in response to incentives. By strategic deception, we refer to distinct changes in communication characteristics that either minimise or exaggerate how wealthy they appear to voters. The hypothesis is that during an election year, political actors will demonstrate communication characteristics distinctly different to a non-election year. To test this hypothesis we do three things: firstly, we build a model to accurately estimate wealth from tweets in a non-election year; secondly, we apply the model to estimate wealth from tweets in an election year; thirdly, we compare differences in *error* between an election year and a non-election year. If no deception were to exist, we would expect error to be distributed similarly between these groups, regardless of whether candidates face an election. The wealth estimation model will use United States politicians' net worth data from the Center for Responsible Politics (Center for Responsible Politics, 2020), as well as tweets from politician twitter accounts (Littman, 2017). Both data sets used cover the years 2017 and 2018.

While there is a growing body of literature on use of social media as a method for politicians to communicate (Golbeck et al., 2018; Hemphill, Otterbacher, and Shapiro, 2013), there is little research on how communication patterns on social media change depending on the incentives that politicians face. We hope to provide insight into whether communication patterns change in response to incentives such as an election and the circumstances under which change exaggeration is likely.

Furthermore we note that previous work on estimating socioeconomic characteristics such as social class and income (Aletras and Chamberlain, 2018; Preoţiuc-Pietro, Lampos, and Aletras, 2015) use *inferred* data rather than information obtained directly about the individual. By inferred data, an example is using the median wage of an occupation, rather than using an individual's real wage. We further the literature on estimating socioeconomic characteristics by using personal finance disclosures that *directly* relate to an individual's wealth.

We use machine learning and natural language processing (NLP) to construct a novel inference tool that can predict wealth of a political actor to within accuracy of half an order of magnitude (RMSLE[1] of ~ 0.6). For context, the range of log wealth varies between 3.9 ($8,000) and 8.5 ($326m), meaning that RMSLE of 0.6 is quite reasonable. This means that a politician worth $1,000,000 would have log wealth of 6, and would have predicted wealth between $250 thousand and $3.98 million. The model uses data from Twitter accounts of United States politicians and their demographic characteristics to infer wealth.

[1]RMSLE is Root Mean Square Log Error

Our results indicate that political actors do not engage in deception as a homogenous group when facing election, rather, they respond to individual circumstances. We find that the top tercile of wealth, when facing an election, has a Twitter inferred log wealth of 0.28 less than they actually have. Similarly, the bottom tercile of politicians facing election have a Twitter inferred log wealth of 0.2 higher than they have in reality. In other words, when faced with an election the wealthier politicians tend to minimise their perceived wealth and poorer politicians seek to exaggerate their perceived wealth. We do not find evidence that exaggeration occurs based on party or experience. The results consolidate existing literature demonstrating that social media and NLP can be used to infer socioeconomic characteristics, including wealth. Furthermore, it underlines an important aspect regarding the rise of social media; the lack of incentives to publish truthful information is fundamentally a problem of asymmetrical information that has little reward for truth and outsized rewards for exaggeration.

Chapter 2

Key Literature

To understand the context for this paper, we firstly discuss the historical role of media and secondly contrast this to the rise of peer-to-peer communication that typifies social media. Thirdly, we discuss the role of incentives and consequences in how social media is susceptible to deception.

Social media allows professionally created content to coexist with amateur user created content (Dijck, 2009), leading to lower barriers to entry for publishing content. While we are in an age of media abundance (Keane, 2013), this does not guarantee information quality, rather, it has the potential to undermine content accuracy (Casero-Ripollés, 2018). It may even even undermine key democratic processes (Bennett and Livingston, 2018). In a political sense, inaccuracies, embellishments or exaggeration in the political process may alter trust in democracy and even influence political outcomes.

2.1 The Traditional Role of Media in Informing Voters

In any democracy, keeping the public well informed of political differences is generally viewed as a positive aspect. Media, such as newspapers, served as the primary way that the public received information on political issues (Dautrich and Hartley, 1999; Owen and Hughes, 1991). Importantly, engaging in dialogue on political candidates was limited to a select few pundits, rather than the general public (Gerstl-Pepin, 2007; Fraser, 1990). Habermas (1991) proposed that the role of media enabled a *public sphere*, where discourse was uninhibited by economic or political interests. A commonality between media platforms prior to social media is that the flow of information was tightly controlled. The choice of who to interview, what to publish and the manner in which information is published was up to the media broadcaster. Therefore, politicians were required to participate in mainstream media in order to persuade and inform voters.

2.2 The Rise of Peer-to-Peer Communication

The rise of peer-to-peer communication has fundamentally changed the way that content is distributed and consumed. We define peer-to-peer communication as online platforms that allow actors to simultaneously produce, distribute and consume content. This definition encapsulates well known companies such as Facebook, Twitter and Reddit. American academic and journalist Jeff Jarvis (Jarvis, 2008) argues that we are entering a phase of participatory media, whereby consumers are also producers of content. He argues that the traditional "one-way street" of mainstream media is fundamentally limited in its capacity to produce and distribute content. This shift in the way that media and content is distributed means that information available on social networks does not have the same professional standards as journalists (Keen, 2007).

This increase in social media use is not just anecdotal; British Twitter use has increased year-on-year in the period 2005 - 2011 (Dutton and Blank, 2011). Media consumption increasingly occurs in the digital space (Statista, 2020), and this means preferences for consumption are changing too; online news consumption is becoming increasingly dominant (Pew Research Center, 2019). Along with an increase in consumption of news content through social networks, it is also changing the incentives that traditional media faces, making them more responsive to the "clickstream" of social network users (Currah, 2009) due to the advertising revenue generated from increasing click counts.

2.3 On Deception and Gaining Strategic Advantage

The role of incentives (and consequences) in changing an agent's behaviour is well established in economic literature. Key papers such as The Market for Lemons (Akerlof, 1978), underline how asymmetrical information can be detrimental to quality in the car market.

Similarly, just as the costs of dishonesty can be detrimental to the car market, leaving only "lemons", in the case of Twitter, this leaves information of poorer quality. Furthermore, the role of consequences are also important in incentivising truthful information (Gneezy, 2005). Given that Twitter features asymmetrical information and a lack of consequences for distributing exaggerated information, we argue that it creates incentives for political actors to deceive.

During the 2016 Presidential Election, in addition to winning the Electoral College, Donald Trump claimed on Twitter that he won the popular vote; claiming millions of illegal votes were cast for Hillary Clinton. One's beliefs in these claims are impacted by information distortion (Carlson, 2018) and whether our network endorses the information (Anspach, 2017). Journalist John Diamond wrote '[on the internet] there is no real way of discerning truth from lies. The net is a repository of facts, statistics, data: unless anything is palpably wrong, we tend to give all facts on our computer screens equal weight.' (As cited in Ball, 2017). Exaggeration has been around well before the advent of social media. However, social media allows political actors to communicate directly with the public, thus potentially circumventing traditional media (Gainous and Wagner, 2013). Given that the environment of social media is susceptible to deception (Tsikerdekis and Zeadally, 2014), individuals may deceive to gain strategic advantage (Buller and Burgoon, 1996). Consequently, political actors have little incentive to produce unflattering content. In other words, they may seek to embellish or exaggerate their personal success in order to present themselves as the more appealing candidate in political contests. Given that Twitter use has near complete saturation by politicians in the United States (Golbeck et al., 2018), this allows unprecedented levels of communication directly to voters, which could be used to influence election outcomes.

2.4 Key Contribution to the Literature

Combining large data sets to estimate economic outcomes, is not new. It has been used to estimate poverty (Blumenstock, Cadamuro, and On, 2015; Blumenstock, 2016; Jean et al., 2016), GDP (Indaco, 2019) and socioeconomic characteristics (Aletras and Chamberlain, 2018; Preoţiuc-Pietro, Lampos, and Aletras, 2015). This study aims to contribute to the literature in two ways. Firstly, it will advance the literature related to inferring socioeconomic status from social media by using a dataset that has real disclosures of net worth of individuals. To our knowledge, there has not been a study that uses a personal net worth dataset to train a model on. It is hoped that this will outline key characteristics of communication on social media that will be useful for future research. Secondly, it will advance our understanding of differences in communication profiles between political actors, including level of experience, political party and incentives they may face. Communication profiles include characteristics such as diversity and structure of language, as well as personal characteristics such as age and location.

Given the lack of analysis of political influence in the digital environment has previously been highlighted (Casero-Ripollés, 2018; Stier et al., 2018), this study will provide clarity on how candidate characteristics influences propensity to exaggerate personal wealth on social media.

Chapter 3

Data

The underlying research question requires accurate estimation of wealth from tweets, and understanding the characteristics of those who over or underestimate their wealth. Given that collecting, cleaning and transforming data has been a multi-step process, with multiple data sources, we will focus on three main data areas in the data analysis process: tweet data, wealth data, and word embedding data. Key tweet and demographic statistics are below in table 3.1 and figure 3.1 provides an overview of the data used in each process.

Table 3.1: *Key tweet and demographic data. NB: Means are calculated on the user level, median values are calculated across all users.*

Feature	Median value
Age (years)	60
Wealth	$1,519,009
Followers	30299
Friends	1133
Favourites (mean)	22.92
Re-tweets (mean)	144.54
Tweet favourites	2869
Total statuses (per user)	4929
Follow-friend ratio	23.6

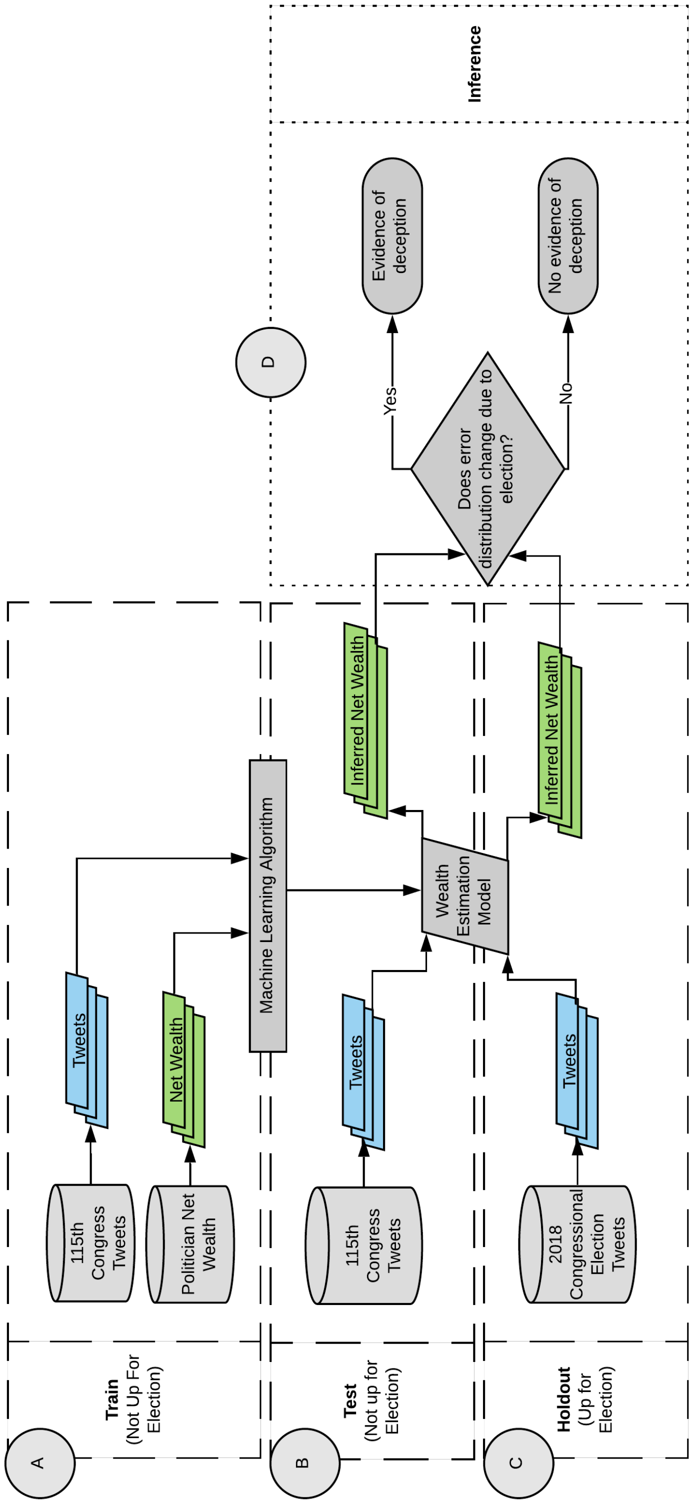

Figure 3.1: *An overview of the data used in each process.*
A: Training data used to build the model.
B: Test data used to validate model and for wealth inference when not facing election.
C: Holdout data used for wealth inference when facing an election.
D: A series of T-tests to test whether error distribution is the same between Test and Holdout predictions.

3.1 Tweet Data

The tweet data is central to estimation of wealth. The initial data set of tweet IDs has been compiled by Littman (2017), and contains over 2 million tweet IDs of the 115th United States Congress. We use a software called Hydrator (*Documenting the Now: Hydrator* 2020) to process the tweet IDs and returns associated text and user information for the corresponding tweet. Furthermore, we also include tweet IDs from the 2018 U.S. Congressional Election (Wrubel, Littman, and Kerchner, 2018), which contains approximately 171 thousand tweet IDs. Of the roughly 2.2 million tweets, we filter for 2017 and 2018, for which we have wealth information, as detailed in Wealth Data. Given that some accounts and tweets may be deleted, it is not possible to retrieve all tweets published. We discuss transformation and processing of the tweet data in the Word Embeddings and Methodology sections.

For each politician they may have multiple accounts from which they publish. Where possible we have attempted to capture this information to avoid data loss, which may weaken estimates of net worth if there is insufficient textual information in the tweets. Therefore, the decision has been made to include this information when computing the word embedding scores, as discussed in the Word Embeddings section. However, with multiple accounts, there is a one-to-many problem when computing individual statistics, such as the number of followers and friends to use. In this instance, we have used the account which has the highest number of followers. The underlying assumption is that this account will carry the most weight when Congress members are communicating with the public.

3.2 Wealth Data

A key contribution of this paper is the use of real wealth information from personal financial disclosures (PFD). The PFD data are publicly available for United States Members of Congress as a result of the *Stop Trading on Congressional Knowledge Act of 2012* or STOCK Act. The STOCK Act requires all Members of Congress to list information about their assets and liabilities. An important caveat is that the Act requires politicians to list the asset and liability *ranges*, rather than the exact value. This means an asset listed as "Real Estate" may have an asset range of $250,001 - $500,000. As a result, exact wealth information is not known. Rather, we can calculate an estimated *range* of assets and liabilities, from which we can then take the average for the range (mean assets - mean liabilities). The Center for Responsible Politics (2020) compiled and computed net worth for politicians. This is the basis of the net worth information that we will use to estimate wealth.

An important factor to note is that PFD data has two formats that politicians may use to file to the *Office of the Secretary of the Senate* or the *Clerk of the House of Representatives*. They may file a handwritten form, or electronically. Filing a handwritten form represents a significant impediment to open and accessible data. The Center for Responsible Politics has previously engaged in competitions to automate processing of these handwritten forms (*PDF Liberation Hackathon* 2014), and it appears that computer-aided techniques were used to assist in processing the data. While the data are not perfect, with missing information for some Congress members, the computed net worth averages were overwhelmingly accurate, as verified through manual testing.

The above information was available as simple averages with a year-by-year breakdown for each politician, giving us the basis for our wealth data. The data was obtained by scraping the information from the Center of Responsible Politics site from the *Personal Finances* section. The resulting data set gives 911 observations. We use data from 2017 - 2018, and we filter out politicians who have wealth negative or missing wealth information. We are left with 370 data points to initially model wealth. Furthermore, we have filtered out politicians that are up for election in 2018, given we want to later estimate whether the underlying patterns of communication change in a year where politicians are up for election.

3.3 Word Embeddings

The third key data component to this analysis is the use of word embeddings. We will discuss further the background of word embeddings in the Methodology. However, briefly, word embeddings allow us to represent text in multi-dimensional space, as vectors. Word embeddings are a broad category of techniques that identify similarities between words.

Given that significant computational power is required to calculate word embeddings accurately and many context-appropriate documents are needed, we are using pre-trained word embeddings trained on Twitter data from Pennington, Socher, and Manning (2014a) in 25 dimensions, along with the pre-processing script that the research team has made available.

Pennington, Socher, and Manning (2014a) use a count-based approach that uses co-occurrence statistics (word-word co-occurrence probabilities) to encode meaning. We will discuss this in more detail in the Word Embedding section.

Chapter 4

Methodology

4.1 An Overview on Machine Learning Approaches

We have now constructed a data set from a variety of sources and can now use this data to estimate wealth accurately, and secondly to test whether politicians exaggerate their wealth when facing an election. Our outcome variable of interest is wealth y_t^i, where i is a person and t represents year. We calculate y_t^i using observed characteristics (features) $x_{t,n,d}^i$, where n represents individual tweets and d represents demographic features d_t^i, which include things such as age and state. We use a multi-method approach that allows models to 'compete' in order to maximise the expected accuracy in predicting wealth from new data. Importantly, this predictive approach differs from many other traditional methods used in the economic toolkit as it does not revolve around *parameter estimation* (Mullainathan and Spiess, 2017). By *parameter estimation*, we mean producing accurate estimates of the relation of parameters β and y and its covariates x. More generally, we use machine learning to minimise a loss function L over function class f subject to complexity restriction $R(f)$ (Mullainathan and Spiess, 2017):

$$\text{Minimise} \sum_{i=1}^{n} L(f(x_i), y_i), \text{ over } f \in F, \text{ subject to } R(f) \leq c$$

This loss function is then used to predict on unseen data. Furthermore, we can scale up or down the *complexity* of a model based on *hyperparameters* and *regularization*. Hyperparameters are values which control the *learning process*, which influence how the algorithm interacts with the loss function (e.g. the number of covariates x to use in a decision tree on each split.) Regularization discourages overly-complex models, which overfits a model to the train set, which will predict poorly on new data.

4.2 Building the Pipeline: From Tweets to Prediction

Given that there is no single "source of truth" for this paper, we will outline the stages involved in getting from tweet IDs to numerical estimations of these tweets used for word embeddings, and combining with wealth data for prediction.

4.2.1 Hydrating Tweet IDs

The first stage involves what is known as *hydrating* tweets. This involves taking a tweet ID and passing it through a program so that it can return information such as the *tweet handle, tweet text, handles tagged in a tweet, hashtags* etc. This process was undertaken using a program called Hydrator (*Documenting the Now: Hydrator* 2020).

4.2.2 Cleaning and Tokenising

A principal aspect of this process is converting text into numerical values that represents language. The complete set of tweets is called a *corpus*. A corpus contains *documents* (individual tweets). A *token* is an atomic piece of a document. By atomic, this refers to the smallest possible piece, and does not necessarily represent a word. We may use the words *term* and *token* interchangeably. For example, the tokenisation of:

- *"I do not like green eggs and ham. I do not like them, Sam-I-Am"* becomes
- *'i' 'do' 'not' 'like' 'green' 'eggs' 'and' 'ham' '.'*

We use a two-stage tokenisation process to ensure that characters such as emojis can be properly processed, and secondly, that all punctuation and hashtags are captured. The two-stage tokenisation has the benefit of separating words, characters, punctuation, repeated letters etc. While this may seem trivial, differences in language are important, as different combinations of characters, punctuation can have different meanings. For example, the following sentences have the same words, however punctuation and emojis alter the meaning:

- Well, this thesis is so interesting!
- Well, this thesis is soooo interesting! 🙃

The assumption is that textual characters, in addition to words, are important to convey meaning and will ultimately improve prediction of wealth. In the Twitter-sphere, non-textual information is more likely to occur than would otherwise be the case. This has been established in textual analysis packages as a standard way to process a corpus, such as in R's Quanteda (Benoit et al., 2018). We implement the tokenisation following from Pennington, Socher, and Manning (2014b) and Mullen et al. (2018), the latter being an R implementation of the Penn Treebank Tokeniser (Taylor, Marcus, and Santorini, 2003).

4.3 Feature Engineering

Feature engineering is the process of extracting features from raw data through various techniques. It is important to keep in mind that the underlying purpose of bringing together disparate datasets is to accurately estimate wealth based on publicly available information such as tweets. As such, we implement a variety of quantitative methods to quantify the textual data for feature engineering. We can group them broadly into three groups: user features, shallow textual features, deep textual features.

4.3.1 User Features

We define user features as information pertaining to the Twitter user and the associated demographic information of the politician. This includes information such as number of followers, number of friends, political party, state.

4.3.2 Shallow Textual Features

We follow the definition of Holmes (1985), whereby we analyse *literary style* by including statistics covering both textual richness, vocabulary and distribution, as well as more advanced computation methods such as textual scaling, which include popular methods such as Wordscores (Laver, Benoit, and Garry, 2003) and Wordfish (Slapin and Proksch, 2008) algorithms.

The following textual statistics use the Quanteda package (Benoit et al., 2018). Where practical, we calculate these features on the document level (each tweet), to avoid mean reversion when concatenating document texts. We take the 10th, 50th and 90th percentiles of the statistics at the document level, which are used as features at the person-year level:

Lexical Diversity

Lexical diversity is a measure to calculate ratios between number of tokens and the length of the document. We implement a Type-Token Ratio: $TTR = \frac{V}{N}$, where V is the number of types (of tokens) and N is the total number of tokens.

Readability

Readability scores are an approximation for how easy sentences are to comprehend. Common parts that make up a reading score include the number of syllables in a word, the length of a sentence etc. We use three readability scores:

1. Flesch's Reading ease score (Flesch, 1948)

$$206.835 - (1.015ASL) - (84.6\frac{n_{sy}}{n_w})$$

   where n_{swy} is the number of syllables, n_w is the number of words and ASL is average sentence length.

2. Simple Measure of Gobbledygook (SMOG) (Mc Laughlin, 1969)

$$1.043\sqrt{n_{wsy} \geq 3\frac{30}{n_{st}}} + 3.1291$$

where n_{swy} is the number of syllables and n_{st} is the number of sentences

3. Mean Sentence Length is calculated as the number of words divided by the number of sentences:

$$\frac{n_w}{n_{st}}$$

Sentiment

Sentiment analysis, or sometimes referred to as polarity score, assigns a value to sentences based on a sentiment dictionary (Jockers, 2017). Each paragraph p_i (for politician i), is broken down into sentences s_n, such that: $p_i = \{s_1, s_2, ...s_n\}$. Sentences are further broken down to words w_n within sentences: $s_i, j = \{w_1, w_2, ...w_n\}$, with w representing each word within a sentence. This approach uses a "ordered-bag-of-words" implementation, where punctuation is largely removed. However, an important aspect of the Jockers (2017) implementation is the use of valence shifters. Valence shifters look at the context directly before and after a word to modify sentiment (polarity). The "look behind" and "look ahead" words are tagged as neutral, negator, amplifier or de-amplifier.

Neutral words do not affect the score. Negator words "invert" the score. For example, without a negator, the phrase *"I do like green eggs and ham."* is positive, however, with a negator, the phrase becomes *"I do not like green eggs and ham."* Likewise, amplifiers and de-amplifiers serve to increase or decrease the polarity of a word. An amplifying example is *"I absolutely do like green eggs and ham."*, whereas a de-amplifying example is *"I barely like green eggs and ham."*

4.3.3 Deep Textual Features

The deep textual features is where we leverage unsupervised modelling techniques to represent text. We use two different categories of deep textual features, the first being text scaling, the second being text embeddings.

Text Scaling

Political text scaling attempts to create a one-dimensional representation of a party based on certain characteristics e.g. left to right, or socialist to capitalist. There are a wide variety of methods to estimate political position, such as expert surveys (Benoit and Laver, 2006; Huber and Inglehart, 1995; Castles and Mair, 1984), manual encoding of political party manifestos (Merz, Regel, and Lewandowski, 2016; Budge et al., 2001) and computational estimation of party/policy position (Laver and Garry, 2000; Laver, Benoit, and Garry, 2003). The early computational approaches relied on manual encoding of a reference (either dictionary or corpus), which requires having a known set of reference texts available, which may be manually intensive. We implement the Wordfish method from Slapin and Proksch (2008), given that this overcomes the manual labour required, and is accepted as the *de facto* method for text scaling (Glavaš, Nanni, and Ponzetto, 2017). The implementation of Wordfish requires two *reference texts* for proper implementation. Typically the reference texts are two documents that are accepted as being on opposite ends of a political spectrum. We use GovTrack's 2018 Ideology Score (*Report Cards for 2018 - Ideology Score - All Senators* n.d.) to determine sides of the political spectrum. We use Senator Kirsten Gillibrand as the "leftmost" position and Senator James Inhofe as the "rightmost" position.

Text as Vectors As we will discuss in the Results section, a large part of the predictive power lies in the use of numerical representations (word and document vectors). These vectors are a way to represent language in multiple dimensions (numerically) to capture meaning. As an example, one approach called "one-hot encoding" is where each word is represented once in a list. This is akin to a dictionary, where words nearby are not related by meaning, rather they are related by their order in the alphabet. A document or word vector can be thought of as representing one aspect as a series of *gradients*. If we take the example of colours, each colour is a combination of three factors: red, green, blue. Under one-hot encoding, we could represent crimson as a 1, where each colour has its own "space" in the "colour dictionary" (Table 4.1). However, under a gradient based model, we can represent crimson as the vector $(0.863, 0.078, 0.2235)$, such as in Table 4.2. Unlike one-hot-encoding, where each colour has its own "space", we can now represented a multitude of colours using a combination of three numbers between 0 and 1.

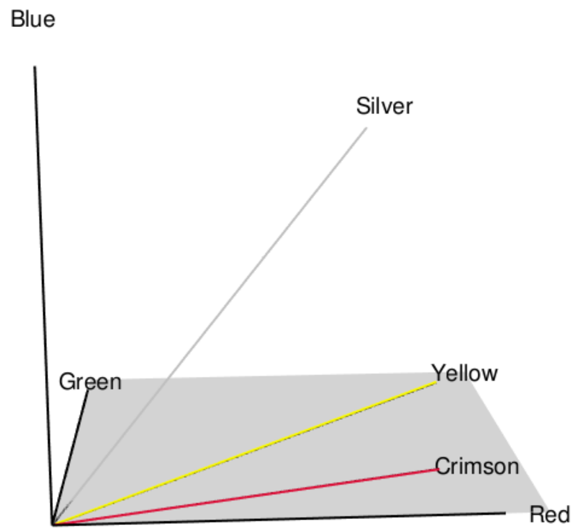

Figure 4.1: *Colours Displayed as vectors*

Table 4.1: *Colours represented in one-hot-encoding.*

Colour	Red	Green	Blue	Lilac	Turquoise	Burgundy	Crimson
Crimson	0	0	0	0	0	0	1

Table 4.2: *Colours represented as vectors.*

Colour	Red	Green	Blue
Crimson	0.863	0.078	0.078
Burgundy	0.520	0.00	0.125
Silver	0.78	0.75	0.75

We see crimson and silver represented as vectors in figure 4.1. Just as colours that are similar have similar RGB values, words and documents that are similar will have similar values along different dimensions. We combine two different approaches to provide a vector based representation of a legislator's words.

Word Embeddings: Global Vectors The first vector representation we use is a word embedding that implements the Global Vector (GloVe) algorithm from the Stanford NLP (Natural Language Processing) Group (Pennington, Socher, and Manning, 2014b). We use a pre-trained word embedding that has been trained using Twitter data. The reason for using a pretrained data set is twofold, firstly it is computationally intensive to train word embeddings and a relatively large corpora is required for accurate word embeddings (Araque et al., 2017; Giatsoglou et al., 2017). The word embeddings are calculated at the person-year level, meaning that individual tweets are concatenated for each person in a year. We use a 25 dimensional vector and weight words by tf-idf [1] (term frequency-inverse document frequency).

Document Embeddings: Gensim The second vector representation this paper uses is a document embeddings that use a *paragraph* approach to creating vectors for each document (or in our case, set of documents), as proposed by Dai, Olah, and Le (2015), using the gensim package (Řehůřek and Sojka, 2010) in Python. Furthermore we calculate each vector at the person-year level using 25 dimensions.

For further reading we recommend Dai, Olah, and Le (2015), Pennington, Socher, and Manning (2014b) Mikolov et al. (2013) for document embeddings, global vector (GloVe) embeddings and word2vec embeddings respectively.

4.4 Modelling Wealth

Following on from definitions listed in the Methodology, our aim is to predict wealth y_t^i using text features $x_{t,n,d}^i$, demographic features d_t^i. We use five model classes to demonstrate both the sensitivity to underlying assumptions about functional form can affect the prediction, as well as to ensure that we can best predict on unseen data. Regression model classes are ordinary least squares, generalised least squares, decision tree models, neural network and support vector machine model. We log transform wealth to ensure that there is a normal distribution, which assists in prediction. Figure 4.2 shows the how the log transformation assists in ensuring a normal distribution.

[1] We use the dot product of a tf-idf weighted document frequency matrix. tf-idf is used to reduce importance of commonly occurring terms, and increase importance of rare/sparse terms. A more technical explanation can be found from Aizawa (2003).

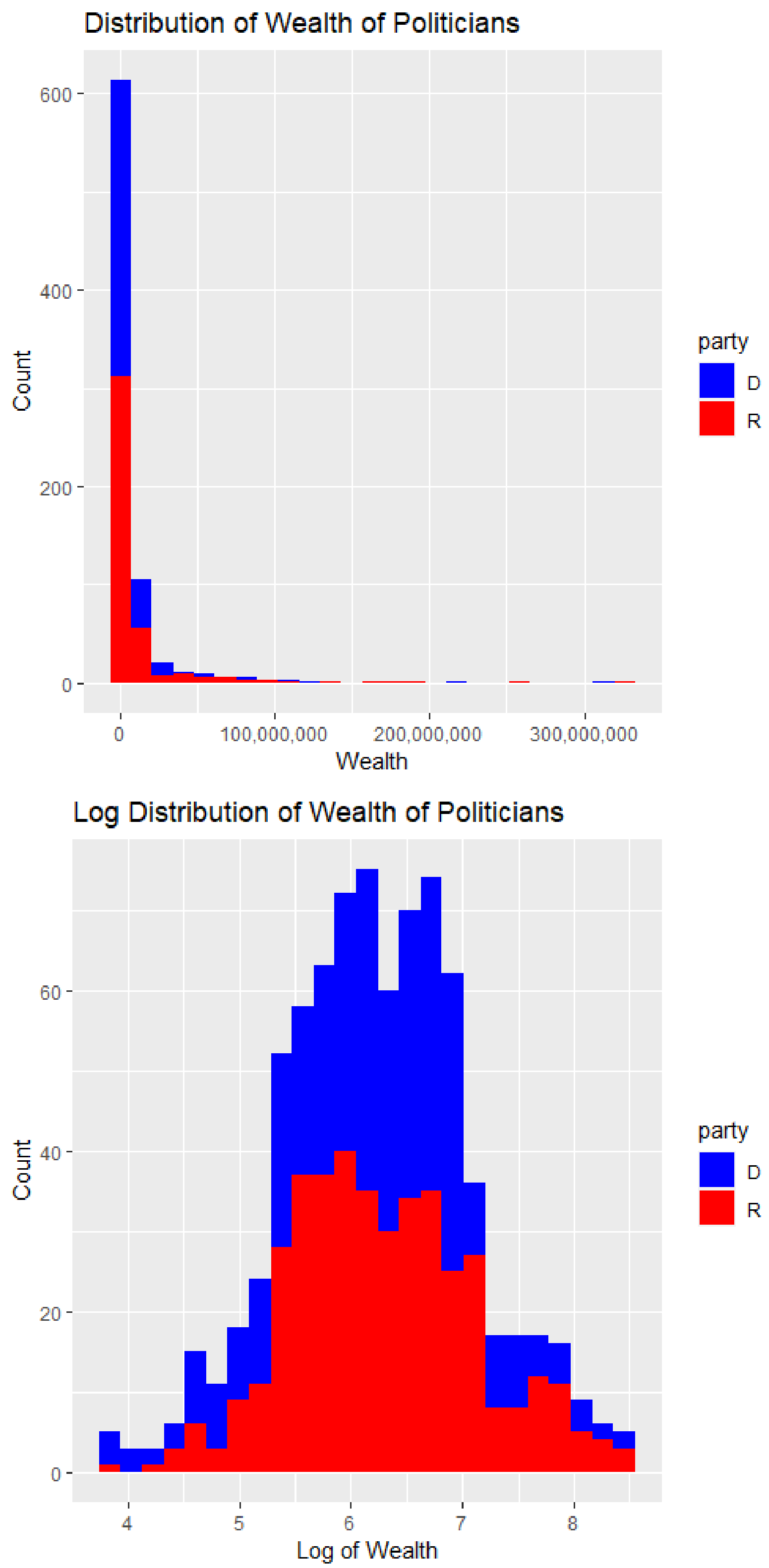

Figure 4.2: *Top: wealth plot without transformation. Bottom: log transformed wealth histogram.*

For more detailed model specifications, including features in data, and model engine used, please refer to the Appendix.

4.4.1 Splitting into Train, Test and Holdout

We split data into three categories: test, train and holdout groups. The holdout group consists of all politicians who are up for election in the 2018 US Congressional Election. From the remaining data, we use an 80/20 train/test split. Furthermore, we process data to normalise around zero and process categorical data as dummy variables. We have 298 rows in our training data, 72 rows in our testing data, and 424 in our holdout data, which is used for inference.

As we will discuss in the Results section, decision tree regression models have the best accuracy (RMSLE), therefore we will devote a larger proportion of the methodology on this class of models.

4.4.2 Linear Regression Models

Firstly, we construct a benchmark ordinary least squares (OLS) model so that we can use this to compare other models to. We use lasso, ridge and elastic net to introduce a penalty for increasing model specification/complexity. We use hyperparameter tuning to find the optimal penalty applied to the regularised models.

4.4.3 Regression Tree Models

Tree based models are a class of models that use two components to make decisions: branches and nodes. Decision trees have branches and then use *splitting criteria* to maximise *information gain* to better predict a class or value. We implement four decision tree models; standard decision tree model, boosted tree model, bagged tree model and random forest model.

Regression tree models broadly consist of five parts[2]:

[2] We note that this largely oversimplifies the algorithmic complexity behind each model, particularly around splitting criteria and regularisation, however, for brevity we have elected not to include detailed model specifications. More information on model selection can be found in the Appendix

- **Trees:** A standard decision tree model only has one tree, however, ensemble methods have more than one tree. More trees will fit data better to the training data, however, too many trees will overfit and perform poorly on unseen data.
- **Sampling procedure:** Decision tree models have a sampling procedure for both *features* and *data points*. Models with multiple trees may use *all* features/data points to build trees, alternatively they may take a *random sample* of features/data to build trees. This avoids overfitting.
- **Depth:** Each tree has a certain depth or number of "splits". More splits means that a particular tree will more tightly classify data.
- **Penalty:** Some models incur a penalty for increasing the complexity of a model. The penalty ensures that the trees are not overfit. Increasing trees or depth increases complexity.
- **Splitting procedure:** All models have criteria in how they split at a branch. Models generally use a variable that gives most information gain. This essentially allows the best "bang for buck" in terms of predictive power, with fewest additional variables needed.

4.4.4 Additional Models Tested

We also use a Single Neural Network (SNN) & Polynomial Support Vector Machine (SVM) models. For more detailed reading on these models we recommend Specht (1991) for neural networks and Smola and Schölkopf (2004) for SVM. We include specifications of all models in the Appendix.

4.4.5 Hyperparameter Tuning

Models hyperparameters were optimised using a bayesian search algorithm. A bayesian search algorithm builds a probability (surrogate) model of the objective function from a Gaussian process. It selects *paramaters* that minimise the difference between the surrogate model and the *true* objective function. Compared to other hyperparameter tuning methods, bayesian search methods are relatively efficient given the use of priors to estimate the surrogate function. See Snoek, Larochelle, and Adams (2012) for further detail on bayesian hyperparameter tuning. We assess hyperparameter results based on root mean squared log error (RMSLE), and choose the simplest model within 10% of optimal RMSLE to decrease overfitting to training data.

4.5 Using the Model for Inference

Once a model can accurately estimate wealth, the crucial part is testing whether there are systematic differences between politicians that face an election and politicians that do not. We use the error (difference between predicted and actual wealth) for politicians up for election (holdout) and those who do not face election (test). The holdout consists of politicians who face election in 2018. We will further test to see if there is a statistically significant difference in error when comparing holdout group by party, experience (incumbent/entrant), and level of wealth.

Chapter 5

Results

We divide our results into two sections. Firstly we examine whether language and user characteristics such as age and gender can be used to accurately predict wealth. Secondly, we determine whether political actors engage in strategic deception, and if so, what types of political actors engage in this. To preface this section, we will focus on the random forest and boosted tree models, as they have the lowest RMSLE on unseen test data. We include the full breakdown of statistical tests in the Appendix.

5.1 Can Language Predict Wealth?

We find that language, in addition to demographic variables can predict wealth. Unlike income estimation, such as in Preoţiuc-Pietro et al. (2015), we believe that the task of wealth estimation is a more complex problem, given that the variation in wealth is generally greater than the variation in income. In particular, there are extreme differences in wealth between politicians, with wealth ranging from thousands, to hundreds of millions. We show results for the model type, and train/test split in table 5.1. Given that we evaluate multiple model types, we will focus only on the top performing models, as measured by RMSLE on the test data set. We will therefore focus on the random forest and boosted regression models, as they have better predictive power on unseen data. As show in Table 5.1, features used such as text embeddings combined with demographic information, can be used to accurately predict wealth of politicians.

Table 5.1: *RMSLE results by model type and train/test split.*

Model	Train RMSLE	Test RMSLE
Random Forest	0.3892	0.5953
Boosted Tree	0.5167	0.6481
Bagged Tree	0.5593	0.6770
OLS	0.6384	0.9264
Elastic Net	0.7667	0.6824
Support Vector Machine	0.7735	0.6984
Lasso	0.8044	0.7119
Decision Tree	0.7989	0.7137
Ridge	0.7347	0.7277
Single Layer Neural Network	0.7959	0.7381

We find that the best performing model on the train data is Random Forest, with an RMSLE of 0.3892. Furthermore, we find that the best performing model on the test data is Random Forest, with an RMSLE of 0.5953. Figure 5.1 demonstrates that the random forest model prediction against actual log wealth for the train / test splits. We note that cautious reading of figure 5.1 is required; although it appears that the test plot is skewed when compared to the train plot, an incorrect inference would be that this drives the results of the study. Any skew/bias would equally apply to the holdout data, meaning that we can then perform t-tests without bias.

5.1.1 Linking Language and Wealth

We believe that there is an underlying link between the language we use and the audience that legislators try and connect with. As an example, we will use document vector embeddings to demonstrate document similarity. Document similarity assesses vectors of documents to determine nearby vectors. This is akin to the colour vector *crimson* being more similar to *burgundy* than *turquoise*. We take a random politician, Elijah Cummings, and test to see whether similar documents have similar wealth. As we see in table 5.2, documents most similar to Elijah Cummings are all Democrats with similar values for wealth. This demonstrates there seems to be not only a link between language and political affiliation, but importantly also a link between language and wealth.

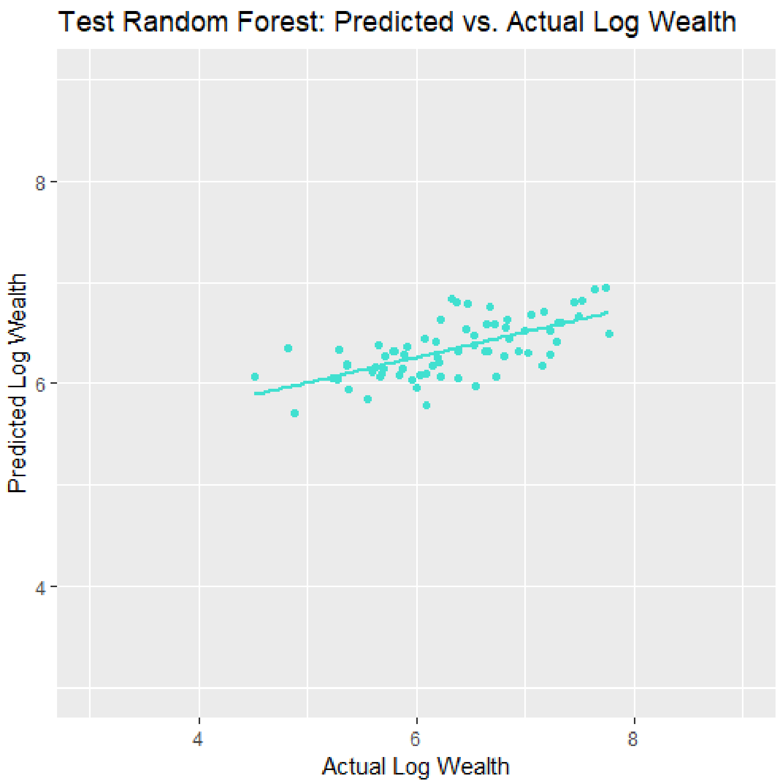

Figure 5.1: *Top: Random Forest Plot on Train Data; RMSLE: 0.39.*
Bottom: Random Forest Plot on Test Data; RMSLE: 0.60.

Table 5.2: *Documents most similar to Elijah Cummings in 2018.*

Name	Similarity	Party	Log Wealth
Elijah Cummings	1	Dem	6.46
Jerrold Nadler	0.9255	Dem	5.08
Jamie Raskin	0.9214	Dem	6.62
Jab Schakowsky	0.9174	Dem	5.87
Bennie Thompson	0.9062	Dem	5.95
Zoe Lofgren	0.8984	Dem	6.58

5.1.2 Variable Importance

As mentioned in the Methodology section, of key interest is the ability to accurately estimate wealth, rather than estimating *parameters* associated with particular variables. However, what we can do is measure variable importance instead. Variable importance in the case of the random forest and boosted regression models rely on a method called Gini Importance. Gini Importance is a measure that aims to maximise *information gain* from an additional split in a decision tree. In a nutshell, information gain selects variables that have the highest "benefit" in choosing accurate wealth values. In other words, variables that have little predictive power will not be chosen to split a branch on. We can see from figure 5.2, that user features, such as age, rank highly across both models. Importantly, both document embeddings and word embeddings have a high degree of variable importance, thus supporting the underlying hypothesis that text can assist in predicting wealth.

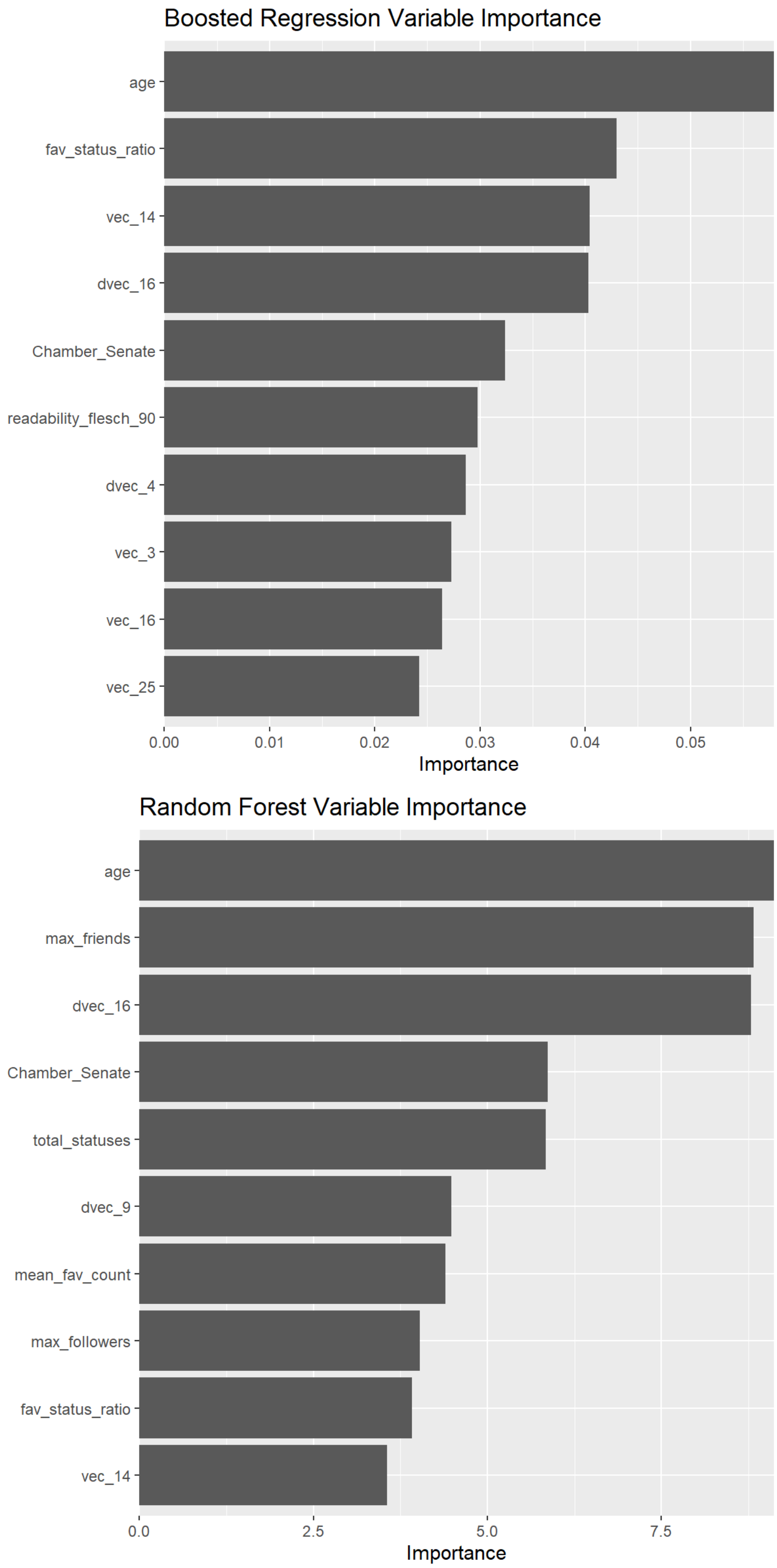

Figure 5.2: *Top: Variable importance of boosted regression tree.*
Bottom: Variable importance of random forest regression.
NB: A dictionary of variable names and meanings can be found in the Appendix.
NB: dvec and vec represent document vectors and word vectors respectively.

5.2 Evidence For and Against Strategic Deception

As a homogenous group, we do not find that evidence that those facing election appear wealthier to voters. Figure 5.3 visually demonstrates this difference between those that face election, and those that do not. We hypothesise that political actors change their communication style in response to key events, such as elections, in order to appeal to their audience on Twitter. Table 5.3 shows t-test results for random forest regression.

Table 5.3: *T-Test results for Random Forest Regression.*
NB: Hypothesis is that the test group (not up for election) and the holdout (up for election) have the same error distribution.

Diff. in Means	T-Statistic	P Value	Alternative	Hypothesis
0.1045497	1.2770730	0.2040009	Two Sided	Entrant
0.1650732	1.9266250	0.0563217	Two Sided	Incumbent
-0.3969961	-1.3064644	0.3012419	Two Sided	Entrant
0.2560718	3.2798252	0.0011440	One Sided: Greater	1st Tercile
-0.0018381	-0.0296692	0.4882501	One Sided: Less	2nd Tercile
-0.2287608	-2.8262949	0.0067703	Two Sided	3rd Tercile
0.1343800	0.9929075	0.3268382	Two Sided	Democrat
0.0792266	0.7581028	0.4505296	Two Sided	Republican

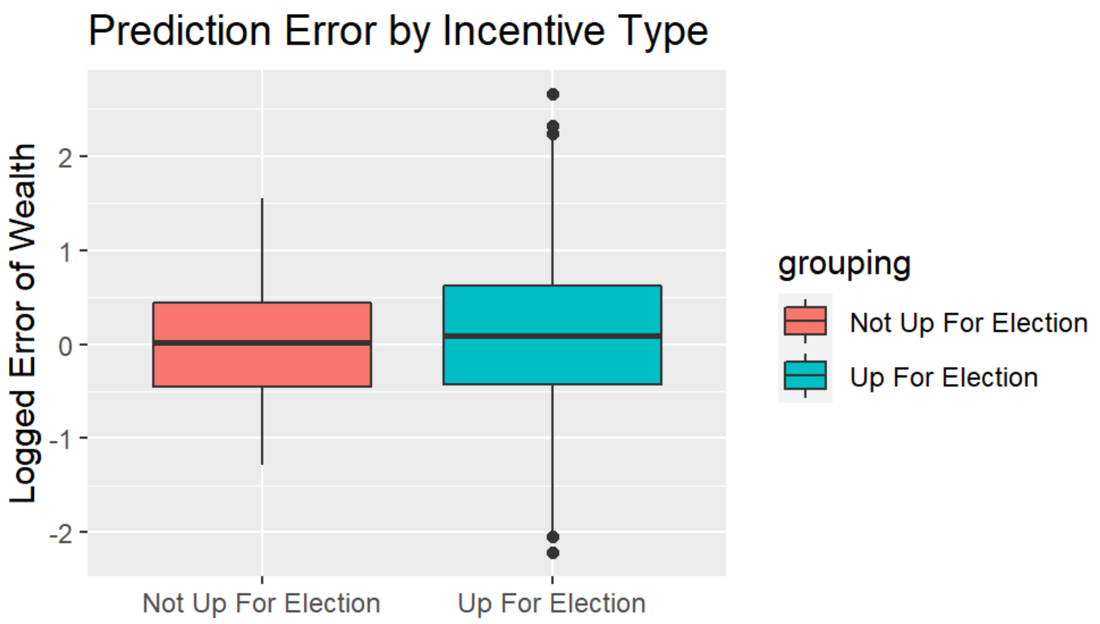

Figure 5.3: *Density plot of error by group.*

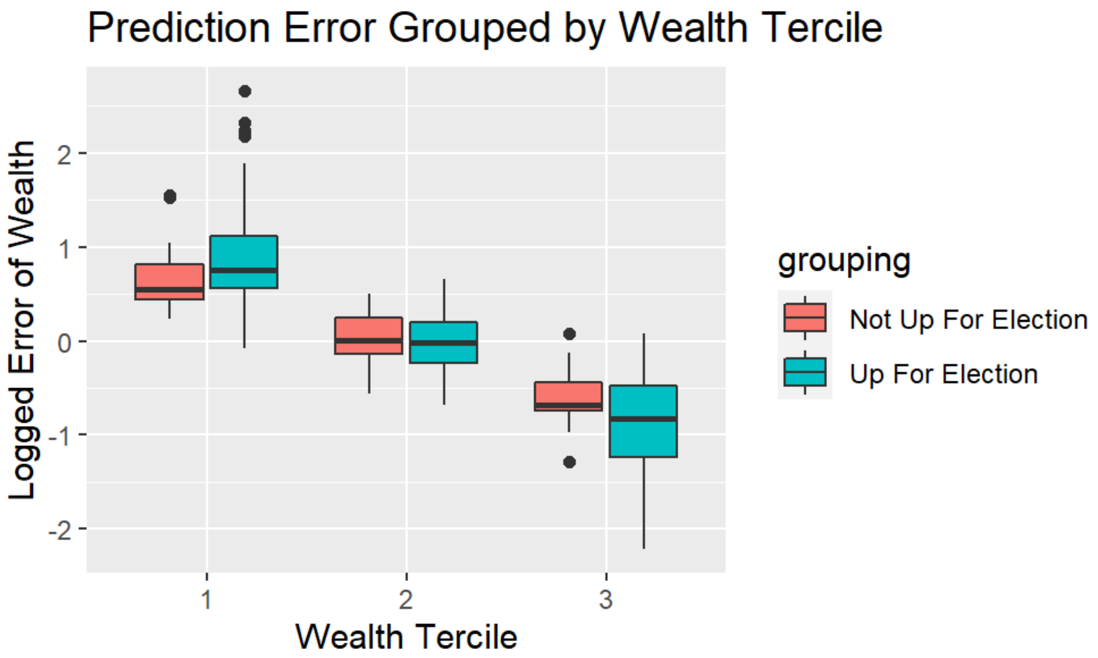

Figure 5.4: *Histogram of Error by Tercile of Wealth*

5.2.1 Tall Poppy Syndrome Two Ways

Tall poppy syndrome is the expectation that tall poppies should grow together, and that poppies deemed too tall should be "cut down to size". We find that, when facing election, those at the bottom tercile of wealth appear wealthier than they are ($p<0.05$) by a fifth of an order of magnitude (0.20). These same results are directly mirrored in the top tercile of wealth, where the top tercile of wealth is predicted to be *less wealthy* by a fifth of an order of magnitude (0.23) than they really are ($p<0.05$).

This indicates that political actors appear to use communication on social media as a tool to strategically deceive voters. It lends weight to the theory that politicians attempt to appear closer to the median wealth of politicians than they actually are. Figure 5.4 demonstrates estimated error by tercile group.

5.2.2 Does Experience and Party Influence Propensity to Exaggerate?

Contrary to our initial hypothesis, we do not find evidence that experience is a factor that drives candidates to engage in strategic deception. Furthermore, we do not find that exaggeration is a distinctly partisan issue, as expected error between groups is not significant. Table 5.3 shows a breakdown of different t-tests conducted. We include further t-tests for the boosted regression model in the Appendix A.3

5.2.3 How Robust are Results to Different Model Specifications?

Chapter 6

Discussion

6.1 Text as a Predictor

This paper adds to the growing body of research that demonstrates that, firstly, social media and associated public data can be used for accurate estimation of economic outcomes. Secondly, we demonstrate that text and language gives important clues as to social class and status. This reaffirms previous work by Sloan et al. (2015) and Preoţiuc-Pietro, Lampos, and Aletras (2015). Importantly, this extends previous work on using social media and machine learning by allowing preliminary predictions of *wealth*, in addition to previously established metrics such as social class and income.

An important caveat, similar to Preoţiuc-Pietro et al. (2015), is that the textual predictors appear to work best when used in conjunction with demographic features. We have constructed a t-SNE plot using the 25 features from document embeddings, and taken a random subsample of 25 data points from the 1st and 3rd terciles from the training data. A t-SNE plot allows us to represent high dimensional (25 dimensions) in a 2D plot. Plot 6.1 demonstrates that while we can group the data, by themselves, merely using one set of features does not perfectly capture the relationship between wealth and language.

However, core to this research question is the assumption that people with different social class, and wealth, behave differently on social media. We believe that this may be for a variety of reasons, but chiefly, we believe that politicians change communication

Figure 6.1: *t-SNE Plot using 25 features from document embeddings, grouped by 1st and 3rd wealth tercile.*
NB: x and y axes are merely representations of two dimensions from the dimension reduction algorithm (t-SNE). They do not directly correlate with a vector of features.

styles in order to maximise their personal gain and/or minimise chances of losing at an election. Clearly, given that the variable importance graphs (Figure 5.2) demonstrate that deep textual features such as document embeddings and word embeddings rank highly in feature importance, we believe that this reaffirms our hypothesis that language and perceived wealth are intrinsically linked.

6.1.1 The Role of Incentives in Explaining Strategic Deception

We argue that strategic deception by political actors is in response to the incentives that each politician faces. Under classic economic voting theory, voters tend to reward incumbents in good times, and punish the incumbent in bad times, and has empirical support (Duch and Stevenson, 2008; Lewis-Beck and Stegmaier, 2008). However, an important extension of economic voter theory goes beyond just economic outcomes; Gerber and Lewis (2004) find that legislators take policy positions that are close to the median voter when the vote share for that policy is large and homogenous. Similarly, we believe that this extension on legislator representation goes beyond policy position, but also translate to the *perceived* wealth of politicians. Evidence presented in this study lends credence to the argument that political actors engage in deception, chiefly by overexaggerating or minimising their wealth, depending on individual circumstances.

6.2 Policy Implications

To our knowledge, this is the first study that examines using textual data to estimate wealth using real wealth disclosures. This is an important first step in furthering the use of high-dimensional data as a predictive tool of socio-economic factors such as wealth and income. However, this study also demonstrates that it is possible to use *publicly available* information to estimate *private information*, such as personal wealth. We believe that there are tangible benefits in using alternative data to understand latent information about political actors, particularly in environments that promote user generated content without corresponding checks on information quality. However, we note that using machine learning techniques on user-generated content such as Twitter provides its own set of ethical challenges. Lack of ownership over personal data and profit incentives of peer-to-peer platforms mean that the underlying business models of these platforms rely on incentivising users to stay engaged on the platform, thus promoting distribution of hyperbolic content. This study is not intended to moralise on whether social platforms impact society in a positive or negative manner, rather, our aim is to ultimately highlight that current incentive structures facing content-producers are not adequate to ensure factual information distribution.

Our second point is on the use of peer-to-peer communication by political actors; the evidence suggests that political actors engage in strategic deception on social media. Given that Twitter use among Congress is nearly at 100% (Golbeck et al., 2018), a pertinent question is what impact exaggeration has on voter preferences. Political actors can easily publish information to a large audience without factual oversight that accompanies more traditional media outlets. This demonstrates the challenge that platforms such as Twitter face in combatting misinformation. By allowing political actors to create their own "versions" of the truth, in a sense, they can portray their version of the facts. Though we are not quite at George Orwell's *1984*, the quote *"Reality exists in the human mind, and nowhere else"* (Orwell, 1983) underlines that the seriousness of creating one's own narrative of the facts.

We have demonstrated that quantitative methods exist and can infer whether or not exaggeration is occuring. For a Twitter specific implication, we argue that current methods of hyperbole detection and outright falsehoods are not adequate in the Twittersphere. We recommend that these large corporations can and should do more to make the content-consumer more aware of whether information is likely to be exaggerated, hyperbolic or factually incorrect. Automatic labelling of tweets that are likely to be deceptive would allow content-consumers to better distinguish between fact and fiction online.

6.3 Limitations

6.3.1 Data Quality & Generalisation

As noted in the Method, the pool of data available is relatively small when compared to traditional machine learning applications. This means that extrapolating results to other countries, including English speaking democracies such as Australia, UK or Canada should be done with caution. Though careful attention to the sampling process is done, smaller data sets mean that there is higher variation depending on the train/test split. However, we have implemented three-fold cross validation and reduced model complexity by choosing the simplest model within 10% of the best performing (by RMSLE) model to help with prediction on unseen data.

Furthermore, as the wealth data relies on the accuracy of the Personal Financial Disclosures, including handwritten forms, it does present the opportunity for error, whether through by manual keying or using optical character recognition. Given that we rely on the underlying legislative requirements, such as asset and liability *ranges*, differences between the actual and reported wealth for politicians in the United States will persist.

6.4 Further Research

A key area for further research is furthering our understanding of how public information is increasingly used to uncover latent information about political actors. Given that much of this data is held privately, such as by Twitter, undoubtedly it helps researchers to better understand social science questions, however, we argue that encouraging discussion on data ownership and incentives is an important next step in allowing consumers to engage with platforms on their own terms.

Lastly, on the topic of strategic deception; this work lays the groundwork for further work in how politicians use language to deceive in online media, and lends evidence to the argument that they deceive in response to individual circumstances. We believe that this could be feasibly replicated in other countries that have wealth disclosure data. Furthermore, we demonstrate that the use of textual features has predictive power when coupled with regression tree ensemble methods and demographic features. Further research may be able to shed more light on whether information, such as images associated with social media posts, can assist in inferring deception from online sources.

Appendix A

Appendix

Table A.1: *List of features used in prediction.*

Category	Feature Implementation	Features	Notes
User Features			
User Profile	Number of Followers	1	Uses account with max followers
User Profile	Number of Friends	1	Uses account with max followers
User Profile	Total Favourites	1	Uses account with max followers
User Profile	Total Statuses	1	Uses account with max followers
User Profile	Average Followers	1	
User Profile	Average Friends	1	
User Profile	Average Favourites	1	Ave. favourites received
User Profile	Max Favourite Count	1	Max favourites received
User Profile	Follow/Friend Ratio	1	Calculated for 2017-2018
User Profile	Favourite/Status Ratio	1	Calculated for 2017-2018
User Profile	Ave. Favourites per tweet	1	Calculated for 2017-2018
User Profile	Ave. Retweets per tweet	1	Calculated for 2017-2018
Demographic	Political Party	1	
Demographic	State	1	
Demographic	Chamber	1	

Category	Feature Implementation	Features	Notes
Shallow Features			
Lexical Diversity	Type Token Ratio	3	10th, 50th, 90th Percentiles
Readability	Flesch	3	10th, 50th, 90th Percentiles
Readability	SMOG	3	10th, 50th, 90th Percentiles
Readability	Mean Sentence Length	3	10th, 50th, 90th Percentiles
Sentiment	Jocker's Dictionary	1	
Deep Features			
Word Embedding	Global Vectors (GloVe)	25	
Document Embedding	Gensim	25	
Textual Scaling	Wordfish	1	

Table A.2: *List of models used, and associated methods.*

Model	Engine	Hyperparameter Tuning Method
OLS	lm	-
Lasso	glmnet	Bayesian Tuning
Ridge	glmnet	Bayesian Tuning
Elastic Net	glmnet	Bayesian Tuning
Decision Tree	rpart	Bayesian Tuning
Bagged Tree	rpart	Bayesian Tuning
Boosted Tree	xgboost	Bayesian Tuning
Random Forest	ranger	Bayesian Tuning
Single Layer Neural Network	nnet	Grid Tuning
Support Vector Machine	kernlab	Bayesian Tuning

Table A.3: *T-Test results for boosted regression tree model. NB: Hypothesis is that the test group (not up for election) and the holdout (up for election) have the same error distribution.*

T-Statistic	P Value	Alternative	Hypothesis
0.8216	0.4130	Two Sided	Entrant
1.3931	0.1662	Two Sided	Incumbent
-1.2877	0.3006	Two Sided	Entrant
2.0408	0.0249	One Sided: Greater	1st Tercile
-0.6093	0.2731	One Sided: Less	2nd Tercile
-2.8229	0.0067	Two Sided	3rd Tercile
0.5727	0.5703	Two Sided	Democrat
0.5092	0.6120	Two Sided	Republican

Bibliography

Aizawa, A (2003). An information-theoretic perspective of tf–idf measures. *Inf. Process. Manag.* **39**(1), 45–65.

Akerlof, GA (1978). "The market for "lemons": Quality uncertainty and the market mechanism". In: *Uncertainty in economics*. Elsevier, pp.235–251.

Aletras, N and BP Chamberlain (2018). Predicting Twitter User Socioeconomic Attributes with Network and Language Information. In: *Proceedings of the 29th on Hypertext and Social Media*. HT '18. Baltimore, MD, USA: Association for Computing Machinery, pp.20–24.

Anspach, NM (2017). The New Personal Influence: How Our Facebook Friends Influence the News We Read. *Political Communication* **34**(4), 590–606.

Araque, O, I Corcuera-Platas, JF Sánchez-Rada, and CA Iglesias (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Syst. Appl.* **77**, 236–246.

Bail, CA, LP Argyle, TW Brown, JP Bumpus, H Chen, MBF Hunzaker, J Lee, M Mann, F Merhout, and A Volfovsky (2018). Exposure to opposing views on social media can increase political polarization. en. *Proc. Natl. Acad. Sci. U. S. A.* **115**(37), 9216–9221.

Ball, J (2017). *Post-Truth: How Bullshit Conquered the World*. en. Biteback Publishing.

Bennett, WL and S Livingston (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *Eur. J. Disord. Commun.* **33**(2), 122–139.

Benoit, K and M Laver (2006). *Party Policy in Modern Democracies*. en. Routledge.

Benoit, K, K Watanabe, H Wang, P Nulty, A Obeng, S Müller, and A Matsuo (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* **3**(30), 774.

Blumenstock, J, G Cadamuro, and R On (2015). Predicting poverty and wealth from mobile phone metadata. en. *Science* **350**(6264), 1073–1076.

Blumenstock, JE (2016). ECONOMICS. Fighting poverty with data. en. *Science* **353**(6301), 753–754.

Budge, I, of Government Ian Budge, Director of the Research Unit Institutions and Social Change Hans-Dieter Klingemann, HD Klingemann, A Volkens, Fellow at the Wissenschaftszentrum and Professor of Political Science Hans-Dieter Klingemann, J Bara, Principal Teaching Fellow Department of Government Eric Tanenbaum, and E Tanenbaum (2001). *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments, 1945-1998*. en. Oxford University Press.

Buller, DB and JK Burgoon (1996). Interpersonal Deception Theory. *Commun. Theory* **6**(3), 203–242.

Carlson, TN (2018). Modeling Political Information Transmission as a Game of Telephone. *J. Polit.* **80**(1), 348–352.

Casero-Ripollés, A (2018). Research on political information and social media: Key points and challenges for the future.

Castles, FG and P Mair (1984). Left–right political scales: Some 'expert'judgments. *Eur. J. Polit. Res.* **12**(1), 73–88.

Center for Responsible Politics (2020). *Personal Finances*. https://www.opensecrets.org/personal-finances. Accessed: 2020-5-20.

Currah, A (2009). *What's happening to our news*. Reuters Institute for the Study of Journalism, Department of Politics and ...

Dai, AM, C Olah, and QV Le (2015). Document Embedding with Paragraph Vectors. arXiv: 1507.07998 [cs.CL].

Dautrich, K and T Hartley (1999). *How the news media fail American voters: Causes, consequences, and remedies*. Columbia University Press.

Dijck, J van (2009). Users like you? Theorizing agency in user-generated content. *Media Cult. Soc.* **31**(1), 41–58.

Documenting the Now: Hydrator (2020).

Duch, RM and RT Stevenson (2008). *The Economic Vote: How Political and Economic Institutions Condition Election Results*. en. Cambridge University Press.

Dutton, WH and G Blank (2011). "Next Generation Users: The Internet in Britain".

Flesch, R (1948). A new readability yardstick. en. *J. Appl. Psychol.* **32**(3), 221–233.

Fraser, N (1990). Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. *Social Text* (25/26), 56–80.

Gainous, J and KM Wagner (2013). *Tweeting to Power: The Social Media Revolution in American Politics*. en. Oxford University Press.

Gerber, ER and JB Lewis (2004). Beyond the Median: Voter Preferences, District Heterogeneity, and Political Representation. *J. Polit. Econ.* **112**(6), 1364–1383.

Gerstl-Pepin, CI (2007). Introduction to the Special Issue on the Media, Democracy, and the Politics of Education. *Peabody Journal of Education* **82**(1), 1–9.

Giatsoglou, M, MG Vozalis, K Diamantaras, A Vakali, G Sarigiannidis, and KC Chatzisavvas (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Syst. Appl.* **69**, 214–224.

Glavaš, G, F Nanni, and SP Ponzetto (2017). Unsupervised Cross-Lingual Scaling of Political Texts. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pp.688–693.

Gneezy, U (2005). Deception: The role of consequences. *Am. Econ. Rev.* **95**(1), 384–394.

Golbeck, J, B Auxier, A Bickford, L Cabrera, M Conte McHugh, S Moore, J Hart, J Resti, A Rogers, and J Zimmerman (2018). Congressional twitter use revisited on the platform's 10-year anniversary. *Journal of the Association for Information Science and Technology* **69**(8), 1067–1070.

Habermas, J (1991). *The Structural Transformation of the Public Sphere: An Inquiry Into a Category of Bourgeois Society*. en. MIT Press.

Hemphill, L, J Otterbacher, and M Shapiro (2013). What's congress doing on twitter? In: *Proceedings of the 2013 conference on Computer supported cooperative work*. CSCW '13. San Antonio, Texas, USA: Association for Computing Machinery, pp.877–886.

Holmes, DI (1985). The analysis of literary style—a review. *J. R. Stat. Soc. Ser. A* **148**(4), 328–341.

Hong, S and SH Kim (2016). Political polarization on twitter: Implications for the use of social media in digital governments. *Gov. Inf. Q.* **33**(4), 777–782.

Huber, J and R Inglehart (1995). Expert Interpretations of Party Space and Party Locations in 42 Societies. *Party Politics* **1**(1), 73–111.

Indaco, A (2019). From Twitter to GDP: Estimating Economic Activity From Social Media. en.

Jarvis, J (2008). *Supermedia: Saving Journalism So It Can Save the World*.

Jean, N, M Burke, M Xie, WM Davis, DB Lobell, and S Ermon (2016). Combining satellite imagery and machine learning to predict poverty. en. *Science* **353**(6301), 790–794.

Jockers, M (2017). *Extracts Sentiment and Sentiment-Derived Plot Arcs from Text*. https://cran.r-project.org/web/packages/syuzhet/syuzhet.pdf. Accessed: 2020-12-10.

Keane, J (2013). *Democracy and Media Decadence*. en. Cambridge University Press.

Keen, A (2007). The cult of the amateur. *New York*.

Laver, M, K Benoit, and J Garry (2003). Extracting policy positions from political texts using words as data. *Am. Polit. Sci. Rev.*

Laver, M and J Garry (2000). Estimating Policy Positions from Political Texts. *Am. J. Pol. Sci.* **44**(3), 619–634.

Lewis-Beck, MS and M Stegmaier (2008). The Economic Vote in Transitional Democracies. *Journal of Elections, Public Opinion and Parties* **18**(3), 303–323.

Littman, J (2017). *115th U.S. Congress Tweet Ids*.

Mc Laughlin, GH (1969). SMOG Grading-a New Readability Formula. *Journal of Reading* **12**(8), 639–646.

Merz, N, S Regel, and J Lewandowski (2016). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics* **3**(2), 2053168016643346.

Mikolov, T, I Sutskever, K Chen, GS Corrado, and J Dean (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems 26*. Ed. by CJC Burges, L Bottou, M Welling, Z Ghahramani, and KQ Weinberger. Curran Associates, Inc., pp.3111–3119.

Mullainathan, S and J Spiess (2017). Machine Learning: An Applied Econometric Approach. *J. Econ. Perspect.* **31**(2), 87–106.

Mullen, LA, K Benoit, O Keyes, D Selivanov, and J Arnold (2018). Fast, consistent tokenization of natural language text. *Journal of Open Source Software* **3**(23), 655.

Orwell, G (1983). *1984*. en. Houghton Mifflin Harcourt.

Owen, DM and DO Hughes (1991). *Media Messages in American Presidential Elections*. en. Greenwood Publishing Group.

PDF Liberation Hackathon (2014).

Pennington, J, R Socher, and CD Manning (2014a). GloVe: Global Vectors for Word Representation. In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp.1532–1543.

Pennington, J, R Socher, and CD Manning (2014b). GloVe: Global Vectors for Word Representation. In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp.1532–1543.

Pew Research Center (2019). *Americans almost equally prefer to get local news online or on TV set*. Tech. rep.

Preoţiuc-Pietro, D, V Lampos, and N Aletras (2015). An analysis of the user occupational class through Twitter content. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp.1754–1764.

Preoţiuc-Pietro, D, S Volkova, V Lampos, Y Bachrach, and N Aletras (2015). Studying User Income through Language, Behaviour and Affect in Social Media. en. *PLoS One* **10**(9), e0138717.

Řehůřek, R and P Sojka (2010). Software Framework for Topic Modelling with Large Corpora. English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, pp.45–50.

Report Cards for 2018 - Ideology Score - All Senators (n.d.). https://www.govtrack.us/congress/members/report-cards/2018/senate/ideology. Accessed: 2020-10-7.

Slapin, JB and SO Proksch (2008). A scaling model for estimating time-series party positions from texts. *Am. J. Pol. Sci.* **52**(3), 705–722.

Sloan, L, J Morgan, P Burnap, and M Williams (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. en. *PLoS One* **10**(3), e0115545.

Smola, AJ and B Schölkopf (2004). A tutorial on support vector regression. *Stat. Comput.* **14**(3), 199–222.

Snoek, J, H Larochelle, and RP Adams (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems* **25**.

Specht, DF (1991). A general regression neural network. en. *IEEE Trans. Neural Netw.* **2**(6), 568–576.

Statista (2020). *Media use in the U.S. by format 2020*. Tech. rep.

Stier, S, A Bleier, H Lietz, and M Strohmaier (2018). Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter. *Political Communication* **35**(1), 50–74.

Taylor, A, M Marcus, and B Santorini (2003). "The Penn Treebank: An Overview". In: *Treebanks: Building and Using Parsed Corpora*. Ed. by A Abeillé. Dordrecht: Springer Netherlands, pp.5–22.

Tsikerdekis, M and S Zeadally (2014). Online deception in social media. *Commun. ACM* **57**(9), 72–80.

Wrubel, L, J Littman, and D Kerchner (2018). *2018 U.S. Congressional Election Tweet Ids*.