KU LEUVEN

FACULTY OF SCIENCE

Shaping a Leader's Words

An analysis of Barack Obama's speeches through natural language processing

Peter Day

Kylan Young

Supervisor:
Prof. dr. J. De Spiegeleer

Thesis presented in partial fulfillment of the requirements for the degree of Master of Statistics and Data Science

June 2023

Preface

First and foremost, we would like to thank our supervisor Jan De Spiegeleer for providing us with preliminary inspiration and data sources for this project. We would not have been able to complete our analysis in a timely manner without his continued guidance and availability.

Additionally, we would like to thank ex-presidents Barack Obama and George W. Bush for naturally delivering all of their great speeches for use in our research. We would also like to thank the staff at cafe De Blauwe Kater specifically for providing a great meeting atmosphere, as well as their many refreshments which aided us in making difficult decisions.

Peter: I'd like to thank my partner in rhyme, Kylan. A pleasure to work with! Also, my wife Sara for letting me drag us a quarter of the way around the world and putting up with my shenanigans. My parents, and Sara's parents for supporting us for the past 8ish years.

Kylan: I'd also like to thank my partner Peter, for teaching me how small amounts of regular effort can lead up to large amounts of programming work very quickly. Additionally, I'd like to thank my partner Marie for sticking with me and hearing me out about my worries whenever I needed some support, as well as my family and friends for being a source of endless motivation.

Abstract

We undertake a natural language processing analysis of text in speeches made by Barack Obama. Both descriptive and predictive analyses are made. We aim to discover what made Obama an admired speaker. We found few if any academic statistical analyses of Obama's speeches and look to fill that gap. Methods such as Doc2Vec embeddings, correspondence analysis, sentiment analysis, latent Dirichlet allocation and topic modeling are employed. At first, the linguistics of his speeches were investigated with additional comparisons being made. In comparing his speech to newspapers articles we found that he used shorter words and sentences. In a comparison with speeches delivered by George W. Bush, we found he used a more varied sentence structure and more unique words. Next, Doc2Vec embeddings were used to ascertain the reliability of used libraries. Furthermore, we tested to see if sentiment could be predicted based on these embeddings, but this endeavour was deemed infeasible. This collection of analysis techniques could be combined with audio analysis and gesture recognition to fill out a complete look at Barack Obama's speeches.

Beknopte samenvatting

In dit onderzoek voeren we een natural language processing analyse uit op de speeches gegeven door Barack Obama. Het doel van deze analyse is om te achterhalen waarom Obama werd beschouwd als een overtuigende spreker. Sterker nog vonden we weinig tot geen statistische analyses die eerder al werden uitgevoerd, wat betekent dat dit onderzoek mogelijks een uniek perspectief biedt. Om dit doel te bereiken worden zowel beschrijvende als voorspellende analyses gebruikt. Deze waaier aan technieken bevat onder andere Doc2Vec embeddings, analyse van correspondentie, sentiment analyse, latent dirichlet allocation modellen en topic modellering. Om te beginnen werd de structuur van de tekst zelf onderzocht, wat verder gekoppeld werd aan vergelijkingen met andere soortgelijke bronnen. In vergelijking met artikels die soortgelijke onderwerpen beschrijven, observeerden we hoe Obama zijn speeches kortere woorden en zinnen bevatten. Wanneer deze speeches dan verder vergelijkt werden met speeches gegeven door George W. Bush, observeerden we dat Barack Obama meer varieerde in woordkeuze en zinsbouw. De Doc2Vec library werd dan gebruikt om de betrouwbaardheid van gebruikte sentiment analyse packages na te gaan. Verder werd er ook getest of de resultaten van de Doc2Vec modellen gebruikt konden worden om sentiment score te voorspellen, maar dit bleek niet mogelijk te zijn. De technieken die in dit onderzoek uitgevoerd werden kunnen in combinatie met geluidsanalyse en houdingsonderzoek kunnen gebruikt worden om een volledig beeld te krijgen op Barack Obama en zijn speeches.

List of Abbreviations

ARI Automated Readability Index. 8

BERT Bi-directional Encoder Representations from Transformers. 18

CA correspondence analysis. 12, 35

LDA latent Dirichlet allocation. 17

PCA principle components analysis. 6

SMOG Simple Measure of Gobbledygook. 8

TF-IDF term frequency–inverse document frequency. 18

List of Symbols

t	Student's t-test statistic
U	Mann-Whitney U-test statistic
W	Levene W-test statistic

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Peter Day and Kylan Young

As the first African American president in US history, Barack Obama has left a lasting impact on the American populace. After being elected as the 44th president of the United States, he was inaugurated on January 20th, 2009.[43] In eight years over two terms, until early 2017, Obama's administration was notable for its stability and inclusive nature.[44] He positively impacted the lives of the American people, achieving some notable objectives. For example, during his presidency, a new reform law was passed to combat predatory pricing and financial exclusion within the American healthcare system. This law came to be known as the Affordable Care Act and is most likely one of Obama's best-known accomplishments.[22] Other achievements include his commitment to the Paris climate agreement,[17] as well as allowing young unauthorized immigrants to have a temporary lawful status through the Deferred Action for Childhood Arrivals presidential memorandum (DACA).[70]

It is perhaps due to this long term of stability that President Barack Obama is of interest in many fields of research. More specifically, it can be argued that his presidency was predominantly defined by his speeches and the manner in which he gave them.[16] Some sources would even go as far as to state that Barack Obama's presidential campaign would not have survived if not for his memorable speeches.[33] One of Barack Obama's most famous utterances came from his 2004 Democratic National Convention speech.[45] The same one that catapulted his political career to a national scale.

> Well, I say to them tonight, there is not a liberal America and a conservative America, there is the United States of America. There is not a Black America and a White America and Latino America and Asian America, there's the United States of America.

These two sentences contain everyday words and simple structure, yet sum up many of Obama's themes succinctly. First being inclusiveness, he wanted all Americans to enjoy the possibility of the American Dream. Another of his recurring themes comes from his first presidential campaign, and this is hope. This was summed up in his campaign slogan, "Yes, we can," a positive statement of ability.

Barack Obama belonged to the Democrat party. In the American two-party system, this party is identified as being more liberal and left-leaning than their Republican counterpart. Consequently, this also provides further context to the themes and subjects that Barack Obama's presidency covered most extensively. Besides the earlier mentioned inclusiveness and hope, another important and recurrent theme within his presidency is that of climate change. Both the increase in global temperature, as well as the resulting natural disasters hitting the North American continent had garnered the attention of the president. During his presidency, Barack Obama defined an action plan to mitigate the effects of climate change.[5] This stance on climate change was once again underlined through the president's commitment to the Paris Climate Agreement, which was subsequently undone by the following Republican president, Donald Trump, during his campaign of aggressively rolling back Obama's policies.[14] On the topic of past and future presidents, another possibility for analysis is comparing Barack Obama to his predecessor George W. Bush and his successor Donald Trump. Both of these presidents belong to the opposing party, therefore the argument could be made that their adversarial aspect might result in more contrast during comparison. Before this can be done, one must first assess the terms of each president, as well as their general surrounding conditions. This must be done to evaluate whether either president is suited for meaningful comparison with Barack Obama.

We will start by considering the more recent president, Donald Trump, who succeeded Obama and proceeded to roll back many of his policies as mentioned, and whose presidential career was surrounded by much controversy.[29] From making outrageous statements during his speeches to inciting rebellious activities, much doubt has been cast about his behavior. While some say that these actions are the result of some strategy, others would argue that it is merely that the president himself was inadequate at his job.[35] In a study published in September 2017, the simplicity of Donald Trump's use of language was analyzed and compared to other presidential candidates based on a Flesch-

Kincaid readability scale. This scale evaluates the simplicity of text based on the difficulty of reading it. The study observed that Trump reached a score of just under four, meaning that his language use was on the level of a fourth grader. For reference, the scores of the other presidential candidates hovered around eight or nine, which is comparable to the level at which Barack Obama speaks.[79] Adding this information to the fact that Trump only held the presidency for four years, with many calls for impeachment during the last year, it is decided to exclude his speeches from comparison with those of Obama. Since the main interest of this project is the use of language within presidential speeches. This controversy combined with the argument of poor language use might yield significant results, however, that does not mean these results will be interesting or unexpected, given the large contrast between the two presidents.

However, a more interesting comparison might be made with Obama's predecessor. With a presidential term that lasted eight years, as well as experiencing far less controversy, President George W. Bush makes a strong candidate for use in comparing speeches with Obama. During his presidency, he started the war in Afghanistan following the September 11 terrorist attacks in 2001.[67] The centerpiece of his campaign was the promise of large tax cuts using a surplus built up during the Clinton Presidency.[4] Both Bush and Obama presided over a destructive economic recession as well as the continuing war against terrorism in The Middle East. Given these conditions, we decided that George W. Bush is an excellent candidate to be used for comparison with Barack Obama.

In terms of previously performed research on Barack Obama's speeches, analysis is mostly performed from a different perspective than that of natural language processing, or data science. One study that is similar in its methods is from Khudoliy A. in 2014.[37] In this study, the author attempts to indicate the representation of Ukraine as a concept in the speeches of Obama. This was done to show that Ukraine as a country did indeed garner interest from the American government due to its move towards democracy during the war with Russia, which was sparked in 2014 by the Ukrainian crisis.[8] The methods deployed in this particular study involve obtaining groups of words from 15 speeches of Obama, which are then used to explore the existence of concepts within them. However, the analysis is performed from a cognitive-semantic standpoint and bears little to no similarity with this project besides the analogy of collecting information from speech structure to use in further analysis. Further reading ensures that the analysis leans closer to the field of linguistics than data science.

Another study from Schumacher and Eskanzi looks at campaign speeches from the 2016 election.[78] In this unpublished paper, they study two readability metrics for five presidential candidates. Readability indices were created to assist primary and secondary school teachers in selecting readings for their

students. As such, they do not aim to label text past secondary school age. This is a limited study, only looking at five to seven speeches from each of the candidates and they limit themselves to two metrics, one vocabulary based and one lexical. The Poynter Institute for media studies published a grammatical analysis of Obama's speeches.[12] This article looks at the March 2018 Obama speech on race in America and studies three grammatical constructs: allusion, parallelism, and two-ness.

Our goal is to conduct an extensive analysis of the speeches of Barack Obama using natural language processing techniques. As mentioned, the text of his speeches will be compared with speeches delivered by his predecessor George W. Bush. Additionally, comparisons will be made between Barack Obama's speeches about specific events and government programs with newspaper articles from The New York Times and The Wall Street Journal about those same topics. This study will look at a larger range of speeches than those from the studies previously mentioned, as well as using a larger range of methods to study those speeches.

1.1 Data Collection and Preprocessing

All programming was done in Python.[23] The main speech data set came from the American Rhetoric website.[20] It was scraped using the requests and beautifulsoup4 packages.[76][77] This was not straightforward, as the American Rhetoric site does not use consistent HTML code across its web pages. Initial cleaning was done with regular expressions and base Python code. The speeches were saved as both text and comma-separated values files.

For comparing Obama's speeches to newspaper articles, roughly one event per month over his eight-year presidency was selected as a comparison point. Speeches on these events were identified and collected. Articles from The New York Times and The Wall Street Journal were manually selected and collected.[2][1] These were accessed via the KU Leuven Libraries website. Often there were multiple articles about events, the initial or main article from the front page was chosen when possible. Data from these sources were saved as text files.

Additionally, The University of Santa Barbara's The American Presidency Project was used as a source of speeches for both Barack Obama and George W. Bush.[86][85] Speeches were selected from each President's event timeline page.

Further cleaning was done using the Python module unicodedata, to clean Unicode artifacts found in the text files. Regular expressions were used to

remove extra whitespace and comments regarding audience applause or laughter. Some methods required further cleaning such as removing punctuation, numbers and stop words. A combination of Scikit-Learn's English stop words and a list from Kaggle were used.[69][80] Finally, if lemmatization was required, Spacy's *en_core_web_md* model was used.[30] Extensive use was made of Pandas for data manipulation.[68][83] Plotly was used for creating plots.[32] When using deep learning methods, such as transformers for embeddings, the text was left intact to maintain as much context as possible. Deep learning methods do not require extensive cleaning such as stop word removal and lemmatization, as this could be removing valuable context information.

Chapter 2

Linguistics And Syntax

Peter Day

2.1 Introduction

The field of linguistics is generally broken down into five parts: phonology, morphology, syntax, semantics, and pragmatics. Phonology studies the sounds of a language, morphology considers the smallest parts of a language with meaning and how words are formed. Syntax examines the rules and structures of sentences, while semantics studies the meaning of the various parts of a language and pragmatics studies the context and meaning derived from that context.

Syntax of the Obama speeches will be considered including parts of speech, dependencies, or relations, between words, and various counts of syllables and words and the variance of these over the evolution of a text. Associations between words and speeches are investigated, as well as a grammatical construct called parallelism.

2.2 Principal Components Analysis

To begin this exploration, a Principal Components Analysis (PCA) comparing Obama's speeches with the articles from The New York Times and Wall Street Journal was conducted.[2][1] Spacy encodings were used for the input data to

the PCA.[30] We wanted to establish a baseline comparison to ascertain whether the speeches and articles could be distinguished from one another.

Using Spacy, we generated 300-dimension vectors using the *en_core_web_md* model, on the document level for use as input for a simple PCA analysis using Scikit Learn's algorithm.[69] The 300 dimensions were reduced to two via PCA for visualization on a biplot. Spacy uses a tok2vec architecture to create the encoding. Spacy segments the text into tokens and returns a numpy array. The biplot shows a clear separation between the speeches and associated articles, however, with the variables being the 300 encoding float values, the results are not interpretable.

Knowing there is a difference between the speeches and articles, we generated interpretable features on which to run PCA. The words in the various documents were tagged with their part of speech using Stanza's[73][41] neural architecture. Stanza takes a document and splits it into sentences, each of which then gets cast into lists of tokens. Stanza then tags each token with a part of speech, in this case using the universal pos convention, which has 17 tags.[18] The Spacy architecture uses the sum of the output of a CNN and BiLSTM for the hidden layer.[72] The weights at this hidden layer are then used as the encodings. We then calculated the proportion of each part of speech in each document to act as a comparison.

Next, we used the National Research Council Canada's NRCLex to assign emotion scores to each document.[42] NRCLex uses TextBlob as a basis with crowd-sourced emotion scores for approximately 27,000 words. The researchers behind NRCLex applied quality control measures to the crowd-sourced responses to ensure accurate labeling. There are eight emotion tags plus one each for positive and negative. For each of the 10 classes, every word is classed with a 0 or 1 as belonging to that emotion or not. NRCLex then returns a list of classes for which a given word was assigned a value of 1. For a sentence, the total scores for each class are summed. For a whole document, we looked at the relative frequency of each class or emotion.

A dependency parse tree takes a sentence, a flat list of words, and parses it into a tree structure, with the edges representing the directed dependency, or relation, of one word on another. The main verb is at the root of the tree, with, for example, an edge marked as nsubj pointing to the noun subject of the verb. The child nodes often, but not always, represent a complete phrase.[34] We calculated the depth of the tree for each sentence, using Spacy, in our speech and article corpora, and then saved the mean depth for each document, to use as a measure of sentence complexity.

TextBlob was used to calculate polarity and subjectivity of each document.[40]

TextBlob's sentiment function returns two scores: polarity, in the range from -1.0 to +1.0, and subjectivity, which has a score from 0, objective, to +1.0, subjective. Scores were calculated on a document basis.

Various sentence-level statistics were calculated as well. These include metrics such as characters per word and syllables and words per sentence. The number of polysyllabic words and difficult words per sentence were counted. These metrics were then used to calculate some of the well-known readability scores. These readability scores are designed to measure the ease with which one can read a text, or a grade level for which the text is appropriate. Some scores we calculated were the Flesch-Kincaid grade level index, the Dale-Chall readability score, and the Automated Readability Index (ARI). Difficult words are defined as those not in the 3000-word Dale-Chall easy word list.[84] We were not interested in whether the readability metrics accurately describe readability or grade level appropriateness, but rather used them for comparison.

2.3 BiPlot Analysis

A sub-sample of 100 Obama speeches covering his eight-year presidency was collected encompassing topics from Obama accepting the Nobel Peace Prize[50] to a statement on an Ebola outbreak in Africa.[57] Newspaper articles from The New York Times and The Wall Street Journal covering these same 100 events were also collected to use as a comparison and to highlight interesting facets of Obama's speeches.[2][1] The previously mentioned metrics were then calculated for each of these 300 texts and compared. These features were also used as variables for PCA and plotting on a biplot, to identify which had the greatest impact in discriminating between Obama's speeches and the articles.

The greatest separation between the Obama speeches and the newspaper articles came from component one, which explained 33.80% of the variance observed in the data. (figure 2.1) Among the variables with the largest loadings were Coleman-Liau, Flesch-Kincaid, SMOG, and ARI, all readability scores, along with syllables per word and dependency parse tree depth. All of these had loadings with positive values between 0.206 and 0.227, meaning that as a text was further to the right on the biplot the more complex it was with longer words and longer sentences. Most of Obama's speeches fell to the left side of the plot, suggesting shorter words and shorter sentences, although number of unique words had the second lowest loading value of principle component one at -0.132, meaning Obama used more unique words, giving his speeches a greater variety of vocabulary. This variety found in Obama's speeches will become a theme in this chapter.

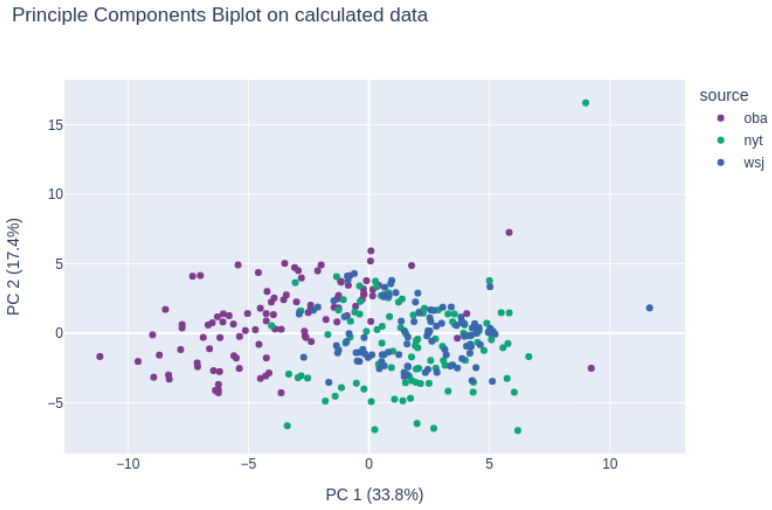

Figure 2.1: PCA on calculated data comparing Obama, New York Times and Wall Street Journal.

The Obama speech which is the furthest to the right on the biplot is a relatively short statement on the Brussels Airport and metro bombings,[64] with a word count of 228 versus his median of 2106 for this dataset of 100 speeches. It has the minimum TextBlob subjectivity of only 0.146 against a median of 0.453, meaning it is more objective than all his other speeches. Another speech in a similar location on the biplot is his 2015-01-08 statement on the Charlie Hebdo shooting in Paris,[63] which has similar attributes and reads similarly. Perhaps, not surprisingly these statements have NRLex anger values of 0.097 and 0.164, both higher than the median of 0.056. The Brussels statement has a Coleman-Liau score of 11.02, which is about 2 points higher than that for all Obama speeches, but quite close to the median values for The New York Times and Wall Street Journal, with median values of 11.33 and 11.45 respectively. The Coleman-Liau index is based on the number of letters per word and the number of words per sentence. Thus the larger-value texts have longer words and longer sentences.

The newspaper article farthest to the left, landing in the Obama speech area, on the biplot is a New York Times article about, paradoxically, a shooting at the Umpqua Community College in Roseburg, Oregon.[13] Perhaps owing to the

many quotes in the story, the readability scores are lower than the median values for New York Times articles. For instance, the Gunning-Fog index has a value of 11.170 versus the median value of 16.615 for all New York Times articles. It also has 789 unique words versus a median of 619. By comparison, the Obama speeches have a median Gunning-Fog value of 12.78, thus this particular article is much closer to the Obama median than that for The New York Times.

When looking at median values, The New York Times and Wall Street Journal articles have higher readability scores and higher sentence tree depth than the Obama speeches, meaning they employ longer words and longer sentences and structurally more complex sentences. (figure 2.2) From an emotion and sentiment standpoint, the Obama speeches have about 40% more joy and their TextBlob polarity is nearly 53% higher, thus more positive. The syntactic difference could be attributed to the difference between text prepared to be spoken and that to be read. Obama's speeches are meant to be understood and easily comprehended by the entire United States population, while the newspaper articles are aimed at a subset of the population. For instance, the Wall Street Journal has a focus on economic and financial news, and would thus target people who have an education in those fields. Text which is meant to be read can be re-read for clarity, while spoken speeches are transient.

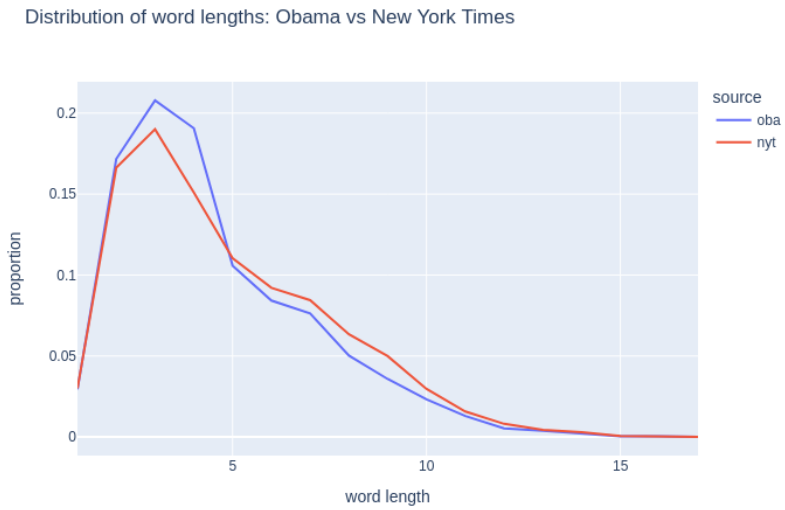

Figure 2.2: Distribution of word length, comparing Obama and The New York Times.

With that difference in mind, we compared the Obama speeches to George W. Bush's speeches collected from the University of California Santa Barbara American Presidency Project site.[85] This set includes over 80 remarks, addresses, and statements from Bush's eight-year presidency. We generated the same set of statistics for the Bush speeches.

PCA was, once again, conducted on these metrics. Once metrics that relate to overall count and which could simply imply the length of a document are removed, there is little to distinguish the text of speeches of Obama and Bush from these variables. Principle component one accounts for 31.49% and component two accounts for 16.38% of the observed variance, so from the two components used for a biplot, there is still just over 50% of the variance unexplained. That being said, Bush's speeches tend toward higher readability scores with a median Flesch-Kincaid score of 10.82 versus 9.98 for Obama ($U = 2823.0$; $p = 0.0002$), while Obama tends to use more unique words, with a median of 774.5 per speech against 397 for Bush ($U = 6099.0$; $p < 0.0001$).

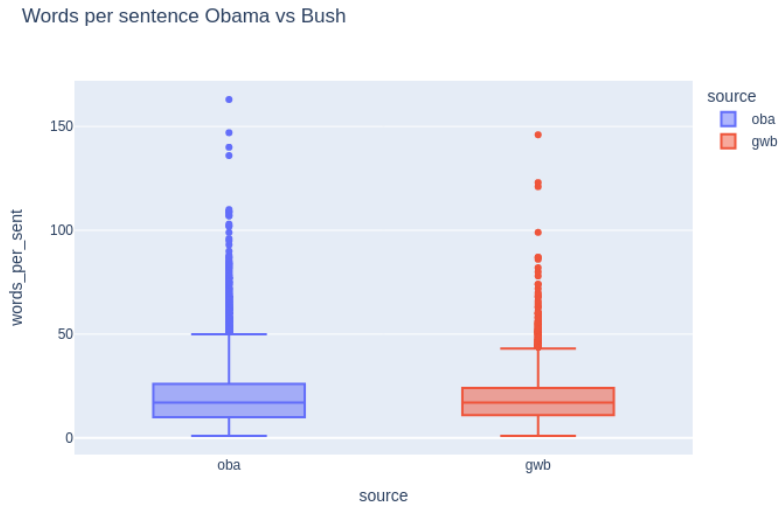

Figure 2.3: Words per sentence, comparing Obama and Bush.

Where there is a difference between these two speakers, however, is in the variance from one sentence to the next within a particular speech. Bush's standard deviation of number of words per sentence is 21.36% smaller than Obama's, with a standard deviation of 10.46 versus 13.31 for Obama ($W =$

233.561; $p < 0.0001$). (figure 2.3) Though both have a median of 17 words per sentence. For dependency parse tree depth, both presidents have a median value of 5, however, Bush's standard deviation for that metric is 11.32% smaller than Obama's at 2.38 versus 2.67 for Obama ($W = 128.505$; $p < 0.0001$). Likewise, the standard deviation of Bush's Dale-Chall readability index is 17.94% smaller than Obama's ($W = 172.433$; $p < 0.0001$).

To compare two speeches given under similar circumstances, we can compare both presidents' inaugural speeches, given on 20 January, eight years apart, in 2001 and 2009.[49][9] Obama's had a median of 19 words per sentence with a standard deviation of 15.51, while Bush had a median of 15 with a standard deviation of only 7.78 ($W = 233.561$; $p < 0.0001$), half that of Obama.

Essentially, Obama with more unique words and greater differences in number of syllables per word and words per sentence varies his language more than Bush. This variation in speech could help hold a listener's attention. With static sentence structure, a listener might start to tune out the speaker with repetitive speech. The novelty of varying sentence structures and new words being worked into a speech could make a listener focus on Obama's speech more than one being given by Bush.

2.4 Correspondence Analysis

Correspondence Analysis (CA) is a popular multivariate technique for comparing two categorical variables. It is used in the field of corpus linguistics to examine a corpus of documents and the words in those documents. CA starts with a contingency table comprised of a document in each row, and a column for each word. The cells are the counts of the word in a particular document. CA explores whether the observed counts in the table are different than would be expected if the variables, speeches, and words, were independent. A measurement of independence for the data is the Pearson Chi-square test. However, one condition for this test is that no more than 20% of the expected table values can be less than 5. With data, such as this, many observed counts are very low, as few of the words will have high counts in any particular speech. Thus, the results of the Chi-square test may not be accurate.[25]

In CA, total inertia is a measurement of the variation observed in the contingency table. CA is computed using singular value decomposition on the contingency table, which returns eigenvalues and an eigenvector. The eigenvalues can be interpreted as the correlation between the two variables, in this case, speeches and words. The sum of the eigenvalues is the total inertia of the data. Thus, each dimension in the decomposition contributes to the total amount of inertia.

When reducing the data to two dimensions for visualization, the more inertia that is explained by these two dimensions, the better the total inertia of the data is explained.

A biplot is a plotting of two of the reduced dimensions as a scatter plot, just as in PCA. Most often, the first two dimensions, explaining the most inertia are used, as to best represent the variation observed in the data. This section investigates speeches and words found in those speeches. The closer two speeches are to each other on the biplot, the more similar they are in terms of word usage. The closer two words are to each other on the biplot, the more speeches they have in common. And lastly, the closer a speech is to a word, the more often that word occurs in that speech than would be expected under independence. CA was employed via the Prince package.[27]

The combined CA biplot, with speeches from both presidents and the 200 most used words, after stop words had been removed, shows little separation between the two sources. Bush's speeches are in red and Obama's speeches are in blue. (figure 2.4) Bush has a higher density of speeches lower and to the right, while the upper left has more Obama speeches. Words occurring furthest down and to the right, in the area more dense with Bush speeches, include "growth", "nuclear", "market", "intelligence", "tax", "international", and "regime". Words occurring in the upper left, an area dominated by Obama speeches, include: "think", "care", "school", "sure", "student", "education", "college", "start", and "idea".

One Bush speech appears in the upper-left, from 2008-01-08, is a remark upon signing the No Child Left Behind Act which reauthorized the Elementary and Secondary Education Act of 1965.[10] This act aimed to improve education at the elementary level by holding schools more accountable for the educational outcomes of the children based on standardized testing.

One of the Obama speeches which falls in the lower right is from 2012-03-27 and was an address at the Hankuk University of Foreign Students in South Korea.[54] In this speech, he largely focuses on security issues facing The United States and South Korea, including threats from North Korea and Russia. He takes time to talk about The Nuclear Nonproliferation Treaty and START (Strategic Arms Reduction Treaty). Another Obama speech in this area of the biplot focuses on a treaty keeping Iran from developing nuclear weapons.

In the center of the plot, with much overlap between the two presidents are words that appear in both of their speeches more often, including “government”, “forward”, “progress”, “leader”, “people”, and “country”.

In making this CA biplot, we can locate speeches associated with certain words more than expected, and see speeches grouped by words, and to some degree,

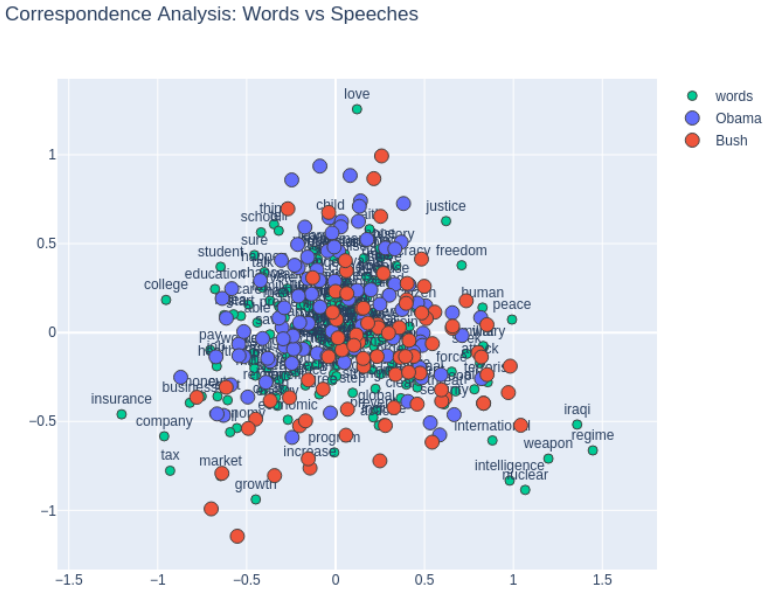

Figure 2.4: Correspondence analysis, comparing Obama and Bush.

by subject.

2.5 Parallelism

Parallelism is a construct in grammar, in which a particular phrase structure is repeated.[12] It can be as simple as a list, where three nouns are repeated as in, “She likes dogs, cats, and fish.” A more complex example comes from Obama’s speech at the 50th anniversary of the civil rights marches in Selma, Alabama.[62]

> For everywhere in this country, there are first steps to be taken, there’s new ground to cover, there are more bridges to be crossed.

The repeating pattern is, using the Penn Treebank tags, NNS-TO-VB-VBN or plural-noun – to – verb – verb-past participle.

We wrote a simple algorithm to scan sentences for repeating structures of length four to seven words long, to avoid overly simple repetitions being detected. It makes use of the NLTK Stanza CoreNLP parsing capabilities.[6][41] We detected that a median of 10.34% of Obama's sentences have parallelism versus 8.82% of Bush's sentences, essentially the same ($U = 4998$; $p = 0.0981$). These lists and counts of parallelisms in each speech are not meant to be exhaustive, but to enable comparison between speakers and speeches.

Obama's usage of parallelism varies by speech. In a 2011 speech about ending the Iraq war, only 4% of the sentences employ parallelism. This is a fairly dry speech with many facts about the number of troops and length of deployment. He is not trying to appeal to people's emotions.

Another speech from 2011 has one of his highest incidences of parallelism, occurring in 22% of the sentences. This speech was given at a dedication of a memorial to Martin Luther King Jr. on Washington D.C.'s National Mall.[52] This is a topic that Obama is quite passionate about. It contains the following sentence:

> It led him to see his charge not only as freeing black America from the shackles of discrimination, but also freeing many Americans from their own prejudices, and freeing Americans of every color from the depredations of poverty.

This contains parallelism in the form IN-DT-NNS-IN-NN, or preposition – determiner – plural noun – preposition – noun. The two phrases are, "from the shackles of discrimination," and "from the depredations of poverty." This repeating structure, besides sounding pleasant, much like a rhyme, can bring attention to an idea as the ear clues in on the repeated grammatical construct. Some studies have found that parallelism helps listeners process the second phrase more easily.[11]

2.6 Conclusion

Syntactical and grammatical constructs of Obama's speeches were considered. Barack Obama employs common grammatical tools such as parallelism, to keep the listener's attention. His speeches can be grouped with words common to a particular speech, that is, speeches in which some words occur more than

expected. His speeches tend more towards words such as, "college," "education," and "student" than Bush's. He keeps his sentences more simple than text meant to be read in a newspaper but not different than speeches by George W. Bush. One difference from Bush, however, is that he changes his sentence structure, within one speech, more, varying word length and sentence length.

Chapter 3

Topic Modeling and Related Techniques

Peter Day and Kylan Young

3.1 Topic Modeling

Besides basic linguistic analysis, there are many more angles available to investigate Barack Obama's speeches. One such angle is that of topic modeling. This is a classic natural language processing technique used to summarize text. It assigns a distribution of topics to each text, each of which is itself, a distribution of words from that text. One of the older and still most popular methods for topic modeling is latent Dirichlet allocation (LDA).[7] LDA applies unobserved, latent, topics to the observed data, in this case, text. A specific number of topics is chosen to represent the whole corpus, then each document is assigned a proportion of each of the topics. Through the words used to represent each topic, we can learn about the meaning of each document.

One drawback to LDA is that it does not find or calculate the appropriate number of topics for a corpus. This must be done by the user. Another is that it assumes a Dirichlet distribution for the prior in the calculation of topics, which may not be guaranteed.

There have been newer topic modeling algorithms introduced since LDA. One such topic called BERTopic uses deep-learning generated embeddings as the basis

for topic generation.[26] Another method, introduced in 2018, is hierarchical stochastic blockmodels (hSBM). It applies a network approach via community detection.[24][31] One advantage to hSBM is that it automatically detects the number of topics, unlike LDA. HSBM also does not require the assumption of Dirichlet distribution for the topics as it uses a completely different approach.

Due to their current popularity, we have utilized both LDA and BERTopic to perform topic analysis on all obtained speeches. BERTopic uses transformer generated encodings, hence the BERT (Bi-directional Encoder Representations from Transformers) in BERTopic. These encodings go through dimensionality reduction and are then clustered. These results get weighted with term frequency–inverse document frequency (TF-IDF) values to generate the topic distributions. Each of these steps can be personalized with a choice of algorithms. We found BERTopic to yield results that were not useful. It very often returned only four or five topics, even when the number of topics parameter was set to, say, nine. Unfortunately, the number of topics parameter seems to set a maximum number of topics for BERTopic to reduce to. The use of only four or five topics does not coincide with manual investigation of Obama's speeches. This is due to the fact that these speeches appear to cover many more topics.

Continuing on from the disappointing BERTopic results, it was decided to employ LDA via the Gensim implementation, using Spacy for tokenization and lemmatization.[75][30] As previously stated, one drawback to LDA is the lack of an automated or mathematical choice of the number of topics. These must be chosen by the user. A common method used to select the number of topics is that of maximizing the coherence score. This method involves the use of a so-called coherence measure calculated for different numbers of topics. These scores are then plotted, from which the user manually decides the optimal number of topics based on the maxima within the graph. In practice, however, the coherence score curve may not be convex, and thus the choice of maximum becomes subjective, as a local maximum may have to be chosen. A possible solution to this issue is to use multiple methods to calculate coherence, and then use the number of topics that attains a local maximum in a large number of methods. Another option is to manually investigate the corpus and choose the number which covers the various subjects that were manually observed. For instance, when we learn topics from all collected speeches and articles, the four coherence methods come to a slight agreement that eight topics would be sufficient. However, if we manually choose nine topics, the added topic could be labeled as education, but the words in the education distribution are mostly covered by other topics in the eight-topic model. When we rank the four coherence methods by score over 11 topics, then sum them, eight is the highest ranked.

The chosen eight topics were labeled: economy, democracy, international

relations, security, health care, terrorism, the Middle East, and civil rights. In table 3.1, each topic is listed together with the highest weighted words, to give context to the reasoning behind each topic label.

Topic 0 economy	Topic 1 democracy	Topic 2 intl relations	Topic 3 security
energy	peace	progress	gun
business	freedom	africa	intelligence
company	democracy	partner	protect
oil	free	region	national
economic	human	human	court
financial	citizen	global	enforcement
crisis	europe	partnership	attack
clean	generation	asia	public

Topic 4 health care	Topic 5 terrorism	Topic 6 middle east	Topic 7 civil rights
health	iraq	nuclear	love
care	military	iran	god
tax	terrorist	israel	school
pay	afghanistan	weapon	white
insurance	troop	international	story
business	attack	deal	talk
reform	iraqi	sanction	black
cut	serve	program	faith

Table 3.1: Topic modeling: eight topics - economy, democracy, international relations, security, health care, terrorism, Middle East, civil rights

The speech with the smallest standard deviation of topic values among the eight topics, that is the speech that was most equally spread across all eight topics was the 2012 victory speech given on the night Obama was elected president to his second term.[56] The topic proportions range from 0.061 to 0.223. This is an upbeat speech, mentioning many of the facets Obama believes make America a great country. He touches briefly on many themes, hence the even spread of values across the eight topics. He speaks about the "American Family" and that for America to succeed it needs to be done together, hence the relatively large loading on the civil rights topic, which includes much of this type of language. The words hope and hopeful are spoken 11 times in this speech. This is not surprising as he made the idea of hope key in both of his campaigns.

On the other end of the spectrum, the speech with the largest standard deviation across the eight topics, also, the speech most loaded on one topic, is a short

statement from 2011 announcing that he is running for a second term in office.[51] The topic with the highest proportion is democracy at 0.564. This is also the highest proportion to be found in the whole data set. This is a short announcement acknowledging that while his administration has completed some of what it set out to accomplish, there is still much work to be done. It is a rally the troops pep-talk speech meant to get his supporters fired up and ready to go.

Not surprisingly a speech such as the 2012 State of the Union address has no zero-valued topics.[55] The topic values range between 0.057 for health care and 0.273 for civil rights. The civil rights topic includes words such as "black," "white," "school," "love," "God," and "bless." The State of the Union address is a yearly speech given by the sitting President in front of a joint session of Congress. It summarizes the previous year and lays out the president's priorities for the upcoming year, as such it touches on all facets of politics and government and would cover all topics.

We plotted a time series of each topic over Obama's eight-year presidency.3.1 In one we fit an ordinary least squares trendline to the topic values. The topics stay fairly steady over the eight years. The speeches' civil rights topic value decreases slightly over this time range ($t = -3.793$; $p < 0.001$), largely due to a relatively elevated proportion of this topic at the beginning of his first term. The most dominant topic, by far, is civil rights, being the highest proportion topic in a speech more often than the other seven topics together for a total of 328 times. Though, in only 155 of these, does civil rights have a proportion greater than 0.33, and thus many speeches touch on a wide variety of topics. We tried a PCA biplot of Spacy encodings of the 400+ American Rhetoric speeches, colored by topic, however, no pattern was found.

3.2 Traveling Words

Following these topic modeling results, we also performed an additional variation on the technique meant as a descriptive analysis. The idea of this technique is based on the concept of tracking words which travel through speeches together, meaning that sets of words occur together in the same speeches often. Of particular interest was the word “hope.” The idea of hope was key throughout Obama's presidential campaigns, as evidenced by the popular red and blue colored poster.[21] Obama had also already published a book titled The Audacity of Hope.[46] Hope suggests concepts such as belief, aspiration, confidence, a desire for better things in the future.

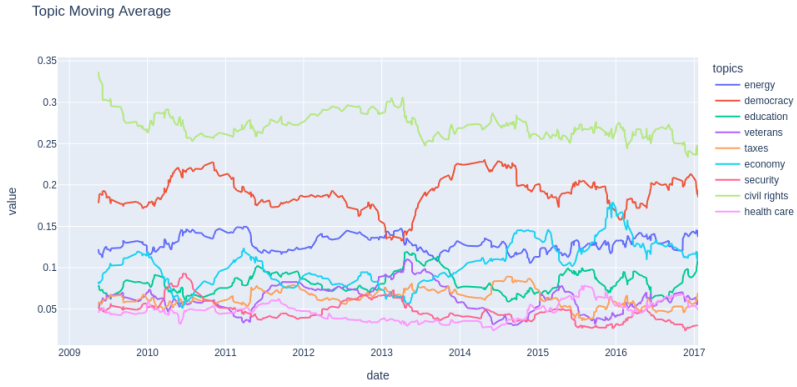

Figure 3.1: Topic modeling moving average.

For this analysis, we first calculated the tf-idf values for all lemmatized words after stop words had been removed. This indicates documents with higher observed occurrence of a word than would be expected. Two speeches with high tf-idf values for "hope" are a speech at Dr. Martin Luther King Jr.'s church in Atlanta, Georgia in 2008 where hope is mentioned 31 times, and the 2004 Democratic National Convention speech where hope is mentioned 14 times.[47][45]

K-means clustering was then applied to the tf-idf matrix to cluster words together, that is words which have high tf-idf values in the same speeches and low values in the same speeches. In this sense, these words 'travel' through the speeches together. We looked at only the 100 most frequent words, settling on this value through a scree plot. Tslearn's k-means algorithm was investigated,[81] but deciding there was little reason to believe there was any periodicity to the observed values, we instead used sklearn's algorithm.[69] The 100 words were placed into nine groups of varying sizes. One group has only "war" and "peace." A three-word group includes, "care," "health," and "insurance." The largest group includes various words related to government, politics and the economy. "Hope" appears in a group with, among other words, "opportunity," "community," "believe," and "progress." In the speech at Dr. King Jr.'s church some version of "believe" occurs 11 times, while "community" is spoken eight times, "fight" is also in this speech eight times as is "change" 11 times. (table 3.2)

Of the ten speeches with the highest mean tf-idf value for the hope related words, nine of ten of these have the highest LDA topic proportion on civil rights,

0	1	2	3	4
hope	american	world	isil	family
opportunity	government	america	iraq	day
believe	economy	new		life
fight	business	nation		love
community	congress	today		god

5	6	7	8
war	care	united	nuclear
peace	health	security	iran
	insurance	states	
		continue	
		support	

Table 3.2: Words which travel together, nine groups

with a mean value of 0.283 and a max of 0.388 from a January 2009 address given at the Lincoln Memorial in Washington D.C. two days before his first inauguration.[48] The one speech with the dominant topic not civil rights, was a speech in Jamaica with economy as the most dominant topic. In it, Obama touches on various topics of concern for Latin America and the Caribbean.[59] "Hope" is only mentioned three times.

Through this method we are able to identify speeches with similar themes, we could label some speeches "hope" speeches. We can also see what words occur together. These words then give the speeches some meaning. We were able to identify words which commonly appear alongside "hope." The "hope" speeches embrace the ideas of community, belief, change, and fighting, or perhaps better said as struggling, as to not be confused with war type fighting. This set of words is close to the dictionary definition of hope, "the feeling that what is wanted can be had or that events will turn out for the best."[19]

3.3 Emotion and Sentiment Analysis

Sentiment analysis attempts to assign a value to a document indicating whether that document is positive or negative in sentiment. Likewise, an alternative called emotion analysis attempts to assign, to individual words, values for eight different emotions. Sentiment analysis is a basic technique in understanding a text and is often applied to reviews of products or services to understand whether

or not the customer likes these and where improvements can be made.[39] We applied it to Obama speeches to understand the overall tone of any given speech.

For performing sentiment analysis, two different approaches were used. Firstly there is the simple approach, using the TextBlob library.[40] TextBlob not only analyzes sentiment but also subjectivity, that is whether the text is factual or consists of judgements based on the writer's ideas and beliefs. TextBlob sentiment analysis is built upon a regular expression based library called pattern. Looking at TextBlob's sentiment polarity values over Obama's timeline, there does not appear to be much of a trend. One consideration to note is that the polarity returned by TextBlob for a whole document is not the same as the mean for that document when polarity is calculated on a sentence-by-sentence basis.

Secondly, more complex libraries were used. The first library is NRCLex, which calculates scores on the eight emotions as mentioned earlier. The second complex library is called pysentimiento.[71] This library is based on the BERTopic framework, but was subsequently trained on a large data set of tweets and adjusted for sentiment analysis. The result of the pysentimiento library for a document is either 'POS' for positive, 'NEU' for neutral and 'NEG' for negative, as well as an individual score for each.

When looking from a document perspective, the highest polarity, that is, the most positive speech is a short statement from Obama after meeting, then president-elect Donald Trump.[65] This is a short, to the point, unemotional, businesslike statement discussing the transition from one administration to another. It does not read as being particularly positive. When calculated on a per-sentence basis then averaged, the most positive speech is an address to the African Union in Ethiopia.[60] (table 3.3) It starts out quite positive and thankful. The most negative sentence in this speech with a score of –0.80 is, "Dignity was seen as a virtue reserved to those of rank and privilege, kings and elders." He is pointing out that dignity was not available to all citizens, this is subjective, and thus has some sentiment involved. The most positively scored sentence, at +1.0, is, "History shows that the nations that do best are the ones that invest in the education of their people." He's pointing out that Africa's youth are ready to be economically competitive on a world scale, but there is still much work to be done by the nations of Africa to stay on level ground with more technologically advanced countries, though it would not seem to be a particularly positive sentence.

NRCLex is a package from the National Resource Council Canada, which used crowd sourced labels to assign values for eight emotions and two sentiments, positive and negative, to a corpus of words.[42] We compared the NRCLex positive and negative scores to the values from TextBlob's polarity. When

Polarity	Date	Speech
0.127	2015-07-28	Address to African Continent Reps
0.125	2013-07-24	Knox College Speech on the Economy
0.106	2015-01-21	State of the Union Address 2015
0.105	2011-12-06	Speech on Economy at Osawatomie HS
0.097	2014-01-29	State of the Union Address 2014

Table 3.3: TextBlob Polarity: most postitive speeches

creating a correlation table of those values from the 400+ American Rhetoric corpus, we find that NRCLex's positive and TextBlob's polarity have a correlation of only 0.436, while NRCLex's negative and polarity have a correlation of -0.575, only a little improvement. If these were reliable numbers for sentiment, we would expect very high correlation. Perhaps they are not good measures of sentiment, but they might still be useful in comparing one document with another.

The eight emotions in NRCLex's model are fear, anger, trust, surprise, sadness, disgust, joy, anticipation. To continue the investigation of hope, we looked at the speeches with the highest sum of trust, joy, and anticipation, which would seem to be a good representation of hope. (table 3.4) The three speeches with the highest sum of these are: a statement on the passing of Supreme Court Justice Antonin Scalia,[66] a statement on a new education bill being sent to Congress,[61] and the previously mentioned brief statement after meeting with Donald Trump. In this last document, it is striking that what reads as unemotional gets scored positively by two different algorithms.

'hope'	Date	Speech
0.571	2016-11-10	First Meeting with President-Elect Donald Trump
0.523	2010-03-13	Blueprint for Reforming No Child Left Behind Act
0.522	2016-02-13	Address on the Passing of Justice Antonin Scalia
0.504	2015-12-10	Every Child Succeeds Act Signing
0.495	2016-02-20	Weekly Address: A New Chapter with Cuba

Table 3.4: Most 'hopeful' speeches - sum of NRCLex trust, joy and anticipation emotion values

Finally, we calculated sentiment using the pysentimiento library. The calculated sentiment scores for each speech have been plotted in Figure 3.2.

As mentioned earlier, this library produces a score for positive, neutral and negative sentiment. This score is then realized as a percentage of the document

that contains that sentiment. As can be seen in the figure, most speeches are labeled as being mostly positive. This coincides with original expectations, as a president delivering a speech generally wants to spread a message of positivity and hopefulness. An observation of note is that around index 150, there are a few speeches clearly marked as containing predominantly negative sentiment. This coincides with manual observations, which concluded that these speeches are generally concerned with serious topics. Examples of subjects labeled with a lot of negativity are: Announcing the death of Osama bin Laden,[53] discussing economic reforms and debates between election candidates.

3.4 Conclusion

We were able to describe Barack Obama's speeches and assign some meaning using topic modeling through latent Dirichlet allocation. We also looked from the word point of view and found groups of words which appear in speeches together through clustering of term frequency - inverse document frequency values. We found that eight topics was a satisfactory number of topics for describing his speeches and a majority included some component of civil rights. We also found a group of words representing 'hope' which travel together.

Figure 3.2: Plotted pysentimiento sentiment scores for Barack Obama's speeches. The speeches are ordered according to index, which coincides with the order of their delivery dates.

Chapter 4

Vector Embeddings and Prediction

Kylan Young

4.1 Introduction

A very important aspect within analysis of Obama's speeches is that of distinction between documents. For example, when performing a sentiment analysis using NRCLex or the VADER rule-based library, it is of great importance to make sure that any speech is categorized correctly. To ensure that this is indeed the case, vector embeddings may be used to identify differences between speeches in a large dimensional vector space based on usage of words and tokens. In this chapter, this differentiation of speeches is attempted through the use of a Doc2Vec model from the GenSim library. Additionally, the validity of prior performed sentiment analysis is then confirmed through comparing observations and results of speeches that are located far apart from each other in the vector space. Finally, the obtained Doc2Vec vectors are used to construct a model for sentiment prediction using obtained vector coordinates as the predictor. This model is then evaluated in its usefulness.

4.2 The Doc2Vec model

As stated by Le and Mikilov in 2014: "Text classification and clustering play an important role in many applications. At the heart of these applications is machine learning algorithms which require the input to be represented as a fixed-length feature vector." [38]In the case of this study, the machine learning algorithm does not immediately come into play until predictive modeling has started. However, representing the different speeches of Obama as a fixed-length feature vector is of applicable at an earlier stage. Previous techniques of document classification include the bag-of-words and/or bag-of-n-grams algorithms. These algorithms simply concern themselves with converting frequencies of occurrence to different numerical features, either in terms of words themselves or different n-grams. [28] However, as per Le and Mikilov, these algorithms come with some major drawbacks. These drawbacks occur in many forms. When using bag-of-words the word order is lost, causing different sentences to be represented in exactly the same way regardless of sentence structure as long as the same words are used. Bag-of-n-grams does manage to hold on to short context of words, but this algorithm, as well as bag-of-words, suffer from data sparsity and high dimensionality, which counteracts the initial simplicity and efficiency of the techniques. Another drawback of the techniques is the loss of semantics, where the meaning of words is lost and therefore words with similar meanings are not necessarily identified to be much closer to each other than they are to other words with completely different meanings.

Another algorithm that can be considered is the latent dirichlet allocation (LDA) model. This model is often used in the context of topic analysis, which can also be seen as an alternative method of distinguishing documents from one another. However, as sentiment analysis concerns itself with the words and sentences contained within the documents themselves and not the topics, it is perhaps more fitting to use a different technique. This brings us to the Doc2Vec model, which is contained within the GenSim python library. As Le and Mikilov describe, the Doc2Vec model is based on the Paragraph Vector, which is an unsupervised framework that learns continuous distributed vector representations for pieces of text.[38] This paragraph vector is then concatenated with several word vectors obtained from words within the document. Subsequently, all vectors are trained using stochastic gradient descent and backpropagation [15].

Given that this model is able to categorize documents based on word use while maintaining semantics and sentence structure, it has been decided that the Doc2Vec model will be used to obtain vector embeddings for each of Barack Obama's speeches. However, before vector embeddings can be obtained, it first has to be decided which modeling algorithm will be used within the Doc2Vec model itself.

4.3 Distributed memory and Distributed bag-of-words

Since the Doc2Vec model is based on the word2vec technique, the two algorithms that are considered for use are loosely based on two algorithms used within word2vec, but with some extensions to account for the higher classification level of the documents themselves.

The first algorithm that is considered is the distributed memory version of the paragraph vector. This algorithm is largely synonymous with the continuous bag-of-words algorithm used in word2vec.[38] The CBOW model creates a sliding window around the current word that is being classified, and attempts to predict the word based on the context within this sliding window. This context is represented through feature vectors, which are then concatenated to form the word vector. When looking back at the distributed memory model then, the structure is largely the same. However, to incorporate the document level within the algorithm, the additional paragraph vector is added. This paragraph vector is then also concatenated to create the final word prediction vector, creating a document-unique vector for this word. Through use of this additional vector, the concept of the document is also recorded within the context of the predicted word, leaving a numerical representation of the document for further use.

Moving on to the alternative algorithm, we have the distributed bag-of-words model. This algorithm is also based on the simpler word2vec model, however this algorithm is based on the skip-gram model instead of the earlier described continuous bag-of-words. The skip-gram algorithm reverses the workings of the continuous bag-of-words, as instead of predicting the word based on the context, the context is now predicted based on the single word. This results in slowing down calculations significantly, but ultimately this algorithm is considered to be much more accurate when it comes to infrequent words. When looking at the distributed bag-of-words model, the technique is much the same, however in this case the algorithm is actually much faster than its distributed memory counterpart, as the word vectors do not need to be saved, only the document vector. In general, Le and Mikilov recommend to use both model types in conjunction to obtain the best results. However, they also state that simply using the distributed memory model on its own will also be enough to obtain state of the art results.[38] This is due to the retention of word context and word vectors being superior to the workings of the distributed bag-of-words model. What this eventually means is that once the word and document vectors are obtained for the speeches, the distributed memory model and the combined model will both be much more accurate in retaining word context and semantics.

To obtain a complete overview of the vector embedding results, the decision

was made to construct vectors using both the distributed bag-of-words model and the distributed memory model separately, both to compare performance as well as to not overcomplicate the analysis and vector interpretation through the use of a combination of both models.

4.4 Vector embedding results

To prepare the data for use in the Doc2Vec model framework, all speeches obtained from the UC Santa Barbara website were tokenized using the SpaCy library, while cleaning out any punctuation and stop word tokens. The tokens were not lemmatized for this analysis, as this would mean that the semantic meaning of the words would be lost, which is one of the most important factors in distinguishing speeches. For each speech, the full list of tokens obtained from the transcript was supplied and tagged with the title of the speech itself for obtaining the document-unique vectors. The dataset was then shuffled and split into a train and test set with a 70-30 train-test split. Both a distributed memory and distributed bag-of-words model were then initialized, with a vectors having a dimensionality of 300 dimensions to allow a moderate amount of semantic complexity. Both models were trained for 50 epochs with a learning rate of 0.02 to obtain the final models.

A first test that was performed on both models was finding word vectors that were most similar to a given word. This was done to test the retention of semantics within each model. To this end, both models were tasked with finding the top 10 most similar words of ‘Obama’, ‘America’, ‘education’, ‘war’ and ‘president’. The results of this test can be found in Table 4.1

Subsequently, the dimensionality of each document vector was reduced using T-SNE to get a clear visualization of where each speech is located within the vector space. Every document was also colored according to the highest sentiment score, calculated with pysentimiento in a previous chapter. This was done to identify speeches which are located far apart from one another. The visualization can be seen in Figure 4.1

From the similar word vector results, it can be seen that the distributed memory model is quite a bit more accurate at producing similar words than the distributed bag-of-words model. This is to be expected, as that was described by Le and Mikilov. The similar words produced by the distributed memory algorithm are not always on the mark, but they are significantly better than the random gibberish that distributed bag-of-words comes up with.

In terms of the visualization, an interesting pattern can be observed, although

Prompt	DBOW model	DM model
Obama	Lilly, compels, fairly, Assistant, medicine, robotics, stone, Ravenstahl, infant, wrangling	Chief, straight, visited, Hillary, Vietnamese, Hawaii, mass, Congress, FEMA, applause
America	Abused, fusion, Put, lessen, jailer, Protestants, picture, relative, locate, Microsoft	President, decades, the, sees, lying, changed, this, ASEAN, pursued, eliminated
Education	Brostrom, Party, paint, hop, anchor, McFaul, structured, starting, violated, unequal	Justice, invest, faith, systems, free, Laos, students, protest, doctors, training
War	Taiwan, Syrians, stems, inherited, overlap, perpetuated, remiss, completes, Shanghai, convene	Leader, term, space, Ambassador, senseless, Israeli, financial, force, threats, U.S.
President	Region, tone, avert, Northern, filled, goodwill, seventh, employ, brothers, invalidated	Most, side, Cubs, head, Navy, Hill, sides, above, State, negotiation

Table 4.1: Top 10 similar words produced by either the distributed bag-of-words model or the distributed memory model, based on 5 different prompts

Figure 4.1: T-SNE visualization of embeddings in 300-dimensional vector space, coloured by sentiment analysis result.

it is not clear what is causing the pattern. This is due to the fact that it is not possible to discern what both T-SNE dimensions represent, since these were constructed by reducing the dimensionality of the 300-dimensional vectors. However, this visualization does allow us to identify speeches that are located quite far apart from each other in both dimensions.

4.5 Sentiment analysis validation

For the first dimension, positively labeled speeches exist both at high and low values. Speeches at low values include the First presidential inaugural address, which can be seen as a speech that would usually be filled with much hopefulness and promises. Meanwhile, positive speeches on the high values include the white house correspondents dinner speech, which is a speech given to commemorate specific individual parties on their merits, obviously implying the use of many a positive word. When looking at negatively marked speeches, low values include 'On foreign and domestic strategies on terrorism', while high values include a speech concerning itself with the Sandy Hook elementary school shooting. From the titles alone, it is noticeable how the pysentimiento library is accurately ascertaining the sentiment contained within these documents, even though they are located far apart on the vector embedding space. Identical observations can be made for the second dimension of the T-SNE visualization, further cementing the fact that the sentiment analysis results obtained from the pysentimiento library can be deemed trustworthy. These speeches will be considered during the validation of previous sentiment analysis, to ascertain the performance of used libraries.

4.6 Predictive model construction

Moving on, another objective in this chapter is to ascertain if it is possible to construct a model that is able to predict the sentiment score of a given speech when supplied with its vector embedding coordinates. To this end, three different Doc2Vec models were used. These models are the distributed memory model, the distributed bag-of-words model and finally the combination of both models, which was considered to be the most accurate according to Mikilov and Le.

For constructing a model capable of predicting sentiment, 200-dimensional training and test vectors were constructed for each document within all 3 Doc2Vec algorithms. The dependent variable, being the sentiment score, was

obtained through the use of the pysentimiento library, after which the highest sentiment score was encoded into a numerical format. This format encoded a positive sentiment to 1, a neutral sentiment to 0.5 and a negative sentiment to 0. These values were then used to train and test the model results. The final models consisted of a neural network with 4 dense layers of which the input and 2 subsequent layers used the ReLu activation function. The final layer then produced the output using a sigmoid activation function. Each neural network was then optimized in terms of training accuracy before testing it against the test set of sentiment scores. The models were trained for 100 epochs each. Finally, the accuracy score of each neural network model was calculated for comparison purposes.

4.7 Model evaluation

In the following table, each neural network is visualized together with the Doc2Vec algorithm used to produce the training vectors. The final column illustrates the ultimately obtained accuracy score for each model.

Model	Doc2Vec Algorithm	Accuracy score
A	Distributed bag-of-words	0.750
B	Distributed memory	0.750
C	Combination	0.7578

Table 4.2: Accuracy score for each neural network model, based on the Doc2Vec algorithm that was used to produce training vectors.

Given that the accuracy score of each model on the test set reaches 75% at best, it can be said that this method of predicting sentiment score of a document is not advisable. Care should be taken in boldly attempting to translate vector embeddings for further use, especially since the encodings themselves are not inherently understood due to the unsupervised nature of the algorithms. Of particular interest is the small increase in accuracy when using the combination of both algorithms for constructing training and test vectors. This result might coincide with earlier statements made by Le and Mikilov pertaining to a very small increase in retention of semantics when using both algorithms in conjunction with each other.[38]

4.8 Conclusion

To conclude this chapter, we were able to successfully investigate the retention of semantics in different Doc2Vec algorithms. Subsequently, we managed to produce a visualization through which each speech of Obama could be clearly differentiated from each other, after which the validity of previous sentiment analysis could be verified. From the results, it became apparent that both speeches containing large amounts of positive and negative words were labeled correctly by the pysentimiento library. This further cements the trustworthiness of these results.

Additionally, neural network predictive models were constructed based on the obtained vector embeddings. However, both due to the difficulty of interpreting these encodings, as well as sub-par accuracy scores on the test set, it can be concluded that this method of operating is not at all desirable if one wants to correctly categorize speeches based on sentiment.

Chapter 5

Selma 50th Anniversary Speech

Peter Day

In March 1965 in Selma, Alabama a series of civil rights marches took place to raise awareness about the systematic disenfranchisement of millions of African Americans by various Southern states. Even though the Civil Rights Act of 1964 had been enacted the previous year, many Southern states kept in place numerous rules and regulations which made it difficult for African Americans to vote. Inspired by the death of a Baptist Deacon during a peaceful demonstration at the hand of an Alabama State Trooper, a long march, designed to inspire, was organized in Selma. At the first attempt, the unarmed marchers were beaten and tear gassed as they crossed the Edmund Pettus Bridge into the neighboring county. After several more attempts, over a couple of weeks, and with protection from the Alabama National Guard, under orders from the Federal Government, the marchers were able to make the three-day 87 km trek to the Alabama state capital in Montgomery. With national sentiment supporting the protesters, the U.S. Congress passed the Voting Rights Act which went into effect in August 1965.[82][3]

On 7 March 2015, Barack Obama spoke at the 50th Anniversary commemorating the 1965 protests.[62] He then led, with President George W. Bush, 40,000 people across the Edmund Pettus Bridge. This speech is widely regarded by many as one of Obama's best speeches. The speech was born from a discussion with his, then-head speechwriter, Cody Keenen.[74] It went through five drafts,

getting passed back and forth between the speech writing team and Obama.[36] It embraces all sides of America and celebrates its various people, from various backgrounds which make America.

Obama's 2004 Democratic National Convention keynote speech was seen as his entry onto the national political scene. In hindsight, as good as that speech is, it sounds rushed and relatively static. In his 2015 Selma speech, Obama takes long pauses to allow the listener to ponder an insightful idea on what makes America, or a list of people involved in the march in 1965. His voice rises and falls to make a point or force the listener to pay attention. He draws parallels between the current day and important people and events in America's history. He looks back and forth at the crowd, making eye contact with those in attendance and emphasizing his points with hand and facial gestures.[58]

This speech is almost twice the length of a median Obama speech in the American Rhetoric data set, at 196 sentences versus 104. However, as far as characters per word, syllables per word, and words per sentence, it falls very close to the median. The complex word indices also are very close to the median, with 0.348 versus a median of 0.337 for the Dale-Chall difficult word percentage, 0.112 versus 0.117 for the Gunning-Fog complex word percentage and a SMOG polysyllabic word percentage of 0.121 versus a median of 0.131. It is, thus, not surprising that the readability indices are all close, but just a little lower, to the median values, making this, from a count perspective, an average speech.

As previously mentioned, Obama has greater variance in his sentence structure than George W. Bush, and this speech is no different. Though, again, the values in this speech are very close to the mean values overall. By the numbers, an average Obama speech.

From the correspondence analysis, this speech lands fairly centrally, among many other speeches, suggesting from the perspective of word frequency it does not differ much from other speeches. The closest words on the CA biplot are 'young', 'day', 'woman', 'hope,' and 'old.' The most similar speech on the plot is a 2010 speech at the 58th National Prayer Breakfast. (figure 5.1) This is not surprising as he quotes the Bible numerous times in the Selma speech and his delivery tone takes on that of an inspired preacher. In addition, both of these speeches embrace diversity from mentioning the many backgrounds of influential Americans to embracing the many faiths in a large country.

When plotting all the American Rhetoric speeches on a PCA biplot, the Selma speech lands quite close to the center of the speech cloud. Where it does differ is in parallelism, it has 25.0% more parallel constructions on a proportion basis than an average Obama speech. (figure 5.2) These parallelisms include, "a clash of armies" and "a clash of wills" in "It was not a clash of armies, but a clash

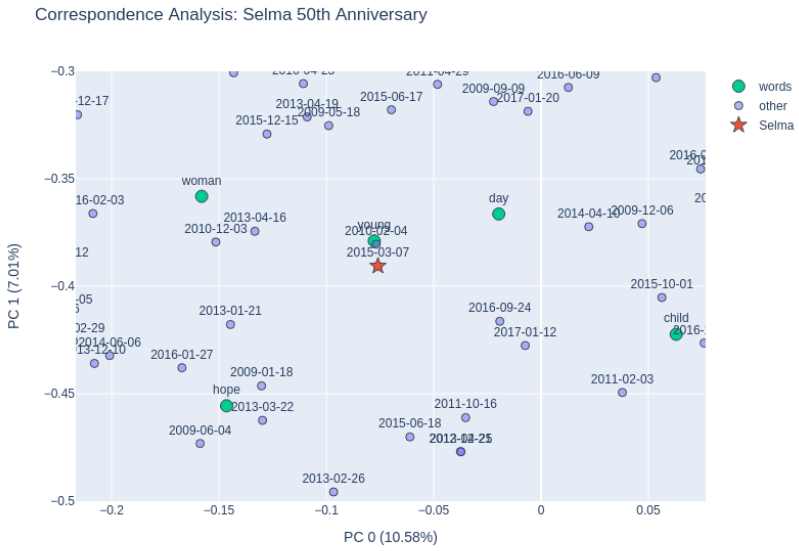

Figure 5.1: Correspondence Analysis on Obama speeches, highlighting the Selma speech.

of wills; a contest to determine the true meaning of America." Later, in the sentence, "Meanwhile, the Voting Rights Act, the culmination of so much blood, so much sweat and tears, the product of so much sacrifice in the face of wanton violence, the Voting Rights Act stands weakened, its future subject to political rancor," there are two six-word parallels, "the culmination of so much blood," and "the product of so much sacrifice."

The proportion of polysyllabic words, those having more than two syllables, is about 10% lower than in a typical Obama speech. Looking at the word length histogram, we can see that the Selma speech has a greater proportion of three, five, and seven-letter words, but fewer words with greater than eight letters. While "Americans" is the nine-letter word with the highest proportion in both the Selma speech and overall, "democracy" and "character" appear more often in the Selma speech. "Democracy" makes up 3.85% of the nine-letters words in the Selma speech, while only 1.46% overall. Other words which appear more often in the Selma speech than in the full data set are "history," "young," and not surprisingly, "Selma."

Despite the percentage of words which are pronouns being 8.65% lower in the

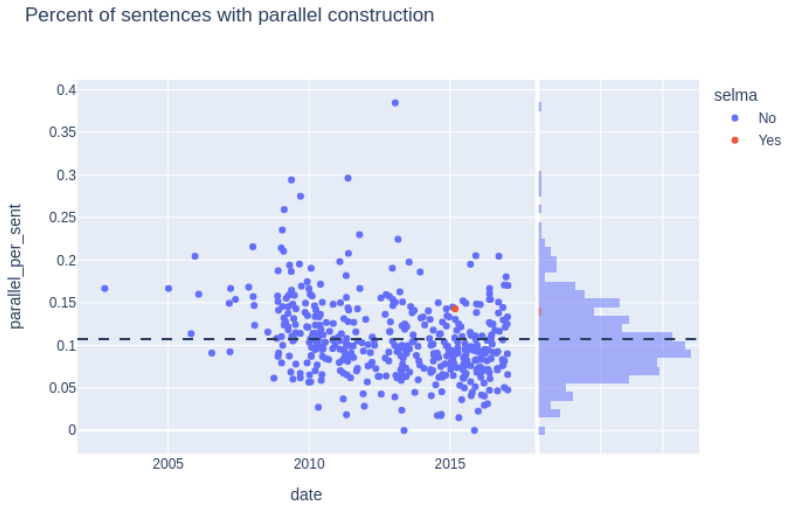

Figure 5.2: Percent of sentences in Selma speech with parallel construction.

Selma speech than the others, a key pronoun, "we" is central to this speech. Sentence 149 has the highest proportion of pronouns at 0.5 in the sentence, which reads, "That's who we are." In fact "we" appears more than 80 times in this speech. Of the 400 plus speeches in the American Rhetoric database, the Selma speech has the 16th highest difference between observed and expected counts of "we," with 35 more occurrences of "we" than expected. Obama often mentions the shared American experience of the people from various backgrounds who make up America. He attempts to draw people together by using the inclusive, "we," and states this explicitly near the climax of the speech, his voice rising with every sentence:

> Because the single-most powerful word in our democracy is the word "We." "We The People." "We Shall Overcome." "Yes, We Can." That word is owned by no one. It belongs to everyone.

By many measures, this speech is quite average. Yet it is among the favorite of his speeches for many people. Obama is able to bring many people together with the ideas in the speech. Even though it is celebrating the march at Selma, or rather, celebrating the spirit and outcome of the march, he draws in people from across the spectrum who have gone through similar struggles. He invokes other

historical struggles and the people involved. If a listener was not aware of what occurred at Selma in 1965, they might understand one of the other examples mentioned and thus be drawn into his argument. His speech is inclusive and allows all who listen to share in the emotion. Is all this predictable from simple language analysis? It may require understanding some external information such as American history. Certainly, his delivery and gestures play a part in the emotion of the speech, which cannot be experienced through text alone. Perhaps, part of what makes Obama so respected as a speaker, is that even a great speech, such as this Selma speech, is quite average.

Chapter 6

Conclusion

Peter Day and Kylan Young

Both descriptive and predictive natural language processing techniques methods were used in our study of Barack Obama's speeches. We found that Obama uses shorter words and shorter sentences than those found in newspaper articles from The New York Times and The Wall Street Journal. When comparing his speeches to those from George W. Bush, while we found the median values of many of our metrics similar, the sentence to sentence variance in Obama's speeches was greater.

We were able to categorize his speeches into topics and found that a majority of them included some element of civil rights. We found that the words 'believe,' 'opportunity,' and 'community' often appear in speeches with 'hope.' More of his speeches lean positive rather than negative. Those that are negative often usually have terrorism and economic reforms as subjects. His most liked speech, by our calculated metrics, is in fact in many ways an average speech.

When investigating vector embeddings, we found that the results of our sentiment analysis could be deemed trustworthy. This was ensured through checking that speeches located far apart from one another were marked with the correct sentiment. Additionally, we attempted to use these vector embeddings to predict the sentiment of a speech. However, this was deemed not to be a viable method for use in prediction models.

Furthermore, we confirmed our previous findings and methods of analysis by performing them on a single speech. The speech chosen was very monumental in

Barack Obama's career, and is twice the length of his average speech. We found that for multiple chosen indices the speech lands close to the median of his other speeches. The results of the single speech again underlined the greater variance in Obama's speech structure when compared to Bush. These observations are a bit confounding however, as the speech is highly favored among many people while showing average results across the board in analysis. Therefore further studies could be performed to investigate other means through which Obama's manner of speaking was made to stand out.

Further studies could include audio analysis of Obama's speeches. Perhaps, categorizing when he takes pauses. Is there a commonality to the topics just mentioned or does sentiment value of the sentence relate to the length of pause? Another area of research could include gesture recognition. His hand movements could be clustered and categorized, then associations with the coinciding text could be searched for. The newer hierarchical stochastic blockmodel method for topic modeling should be investigated. Its network approach is quite different from previous methods and might yield interesting results.

We have managed to shed some light on the key to the success of Barack Obama as a public speaker. His use of language is easy to understand and inclusive leading to many a memorable speech. Through his variance of wording and sentence structure, he was able to hold the people's attention and inspire the nation.

Bibliography

[1] The wall street journal., 1959. Accessed via KU Leuven Libraries and ProQuest Central.

[2] The new york times., 1996. Accessed via KU Leuven Libraries and Nexis Uni.

[3] Archive, C. R. M. 1965: Selma and the march to montgomery. https://www.crmvet.org/images/imgmont.htm, 2023. Accessed: 03-05-2023.

[4] Archives, W. H. President bush helped americans through tax relief. https://georgewbush-whitehouse.archives.gov/infocus/bushrecord/factsheets/taxrelief.html, 2008. Accessed: 2022-11-20.

[5] Archives, W. H. Climate change and president obama's action plan. https://obamawhitehouse.archives.gov/president-obama-climate-action-plan, 2015. Accessed: 2022-11-19.

[6] Bird, Steven, E. L., and Klein, E. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[7] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine learning research 3*, 4-5 (2003), 993–1022.

[8] Brittanica, E. Ukraine crisis. https://www.britannica.com/topic/Ukraine-crisis, 2014. Accessed: 2023-05-01.

[9] Bush, G. W. Inaugural address. https://www.presidency.ucsb.edu/documents/inaugural-address-52, 2001. Accessed: 03-05-2023.

[10] Bush, G. W. Remarks on signing the no child left behind act of 2001 in hamilton, ohio. https://www.presidency.ucsb.edu/documents/remarks-signing-the-no-child-left-behind-act-2001-hamilton-ohio, 2002. Accessed: 03-05-2023.

[11] Carlson, K. *Parallelism and prosody in the processing of ellipsis sentences*. Routledge, 2013.

[12] Clark, R. Why it worked: A rhetorical analysis of obama's speech on race. https://www.poynter.org/reporting-editing/2017/why-it-worked-a-rhetorical-analysis-of-obamas-speech-on-race-2/, 2017. Accessed: 09-05-2023.

[13] D. Vanderhart, K. Johnson, J. T. Gunman attacks oregon college; 10 reported dead. *The New York Times* (October 2015), Section A, Column 0, Pg 1. Accessed: 03-05-2023.

[14] Daley, J. U.s. exits paris climate accord after trump stalls global warming action for four years. https://www.scientificamerican.com/article/u-s-exits-paris-climate-accord-after-trump-stalls-global-warming-action-for-four-years/, 2020. Accessed: 2022-11-17.

[15] David E. Rumelhart, Geoffrey E. Hinton, R. J. W. Learning representations by back-propagating errors. *Nature 323* (1986), 533–536.

[16] Degani, M. *Introduction*. Palgrave Macmillan UK, London, 2015, pp. 1–6.

[17] Denchak, M. Paris climate agreement: Everything you need to know. https://www.nrdc.org/stories/paris-climate-agreement-everything-you-need-know, 2021. Accessed: 2022-11-19.

[18] Dependencies, U. Universal pos tags. https://universaldependencies.org/u/pos/, 2022. Accessed: 2023-02-05.

[19] Dictionary.com, L. hope. https://www.dictionary.com/browse/hope, 2023. Accessed: 03-05-2023.

[20] Eidenmuller, M. E. American rhetoric online speech bank. https://www.americanrhetoric.com/barackobamaspeeches.htm, 2023. Accessed: 2023-02-05.

[21] Fairey, S. Barack obama "hope" poster. https://www.artic.edu/artworks/229396/barack-obama-hope-poster, 2008. Accessed: 03-05-2023.

[22] for Public Affairs, A. S. About the affordable care act. https://www.hhs.gov/healthcare/about-the-aca/index.html, 2022. Accessed: 2022-11-19.

[23] Foundation, T. P. S. Python 3.10.6. https://www.python.org/, 2023. Accessed: 05-05-2023.

[24] Gerlach, M., Peixoto, T. P., and Altmann, E. G. A network approach to topic models. *Science Advances 4*, 7 (2018), eaaq1360.

[25] Greenacre, M. J. Correspondence analysis. *WIREs Computational Statistics 2*, 5 (2010), 613–619.

[26] Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794* (2022).

[27] Halford, M. Prince. `https://github.com/MaxHalford/prince`, 2023.

[28] Harris, Z. S. Distributional structure. *WORD 10*, 2-3 (1954), 146–162.

[29] History. President donald trump impeached. `https://www.history.com/this-day-in-history/president-trump-impeached-house-of-representatives`, 2019. Accessed: 2022-11-20.

[30] Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python.

[31] Hyland, C. C., Tao, Y., Azizi, L., Gerlach, M., Peixoto, T. P., and Altmann, E. G. Multilayer networks for text analysis with multiple data types. *EPJ Data Science 10*, 1 (jun 2021).

[32] Inc., P. T. Collaborative data science. `https://plot.ly`, 2015.

[33] Jaffe, G. Which barack obama speech is the one for the history books? `https://www.washingtonpost.com/posteverything/wp/2016/07/22/which-barack-obama-speech-is-the-one-for-the-history-books/`, 2016. Accessed: 2022-11-17.

[34] Jurafsky, D., and Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. Prentice Hall PTR, USA, 2023.

[35] Kayam, O. The readability and simplicity of donald trump's language. *Political Studies Review 16*, 1 (2018), 73–88.

[36] Keenan, C. *Grace: President Obama and Ten Days in the Battle for America*. Mariner Books, Boston, 2022.

[37] Khudoliy, A. A semantic-cognitive analysis of the concept of ukraine in the speeches of b. obama (2014). *Cognitive studies (Warsaw) 16*, 16 (2016), 153–163.

[38] Le, Q., and Mikolov, T. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning* (Bejing, China, 22–24 Jun 2014), E. P. Xing and T. Jebara, Eds., vol. 32 of *Proceedings of Machine Learning Research*, PMLR, pp. 1188–1196.

[39] Liu, B. *Introduction*, 2 ed. Studies in Natural Language Processing. Cambridge University Press, 2020, p. 1–17.

[40] Loria, S. Textblob: Simplified text processing, 2020. https://pypi.org/project/textblob/.

[41] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations* (2014), pp. 55–60.

[42] Mohammad, S. M., and Turney, P. D. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence 29*, 3 (2013), 436–465.

[43] Nelson, M. Barack obama. `https://millercenter.org/president/obama`. Accessed: 2022-10-25.

[44] Nelson, M. Barack obama: Impact and legacy. `https://millercenter.org/president/obama/impact-and-legacy`, 2022. Accessed: 2022-10-25.

[45] Obama, B. 2004 democratic national convention keynote address. `https://www.americanrhetoric.com/speeches/convention2004/barackobama2004dnc.htm`, 2004. Accessed: 03-05-2023.

[46] Obama, B. *The Audacity of Hope: Thoughts on Reclaiming the American Dream*. Crown, 2006.

[47] Obama, B. Ebenezer baptist church address. `https://www.americanrhetoric.com/speeches/barackobama/barackobamaebenezerbaptist.htm`, 2008. Accessed: 03-05-2023.

[48] Obama, B. Pre-inauguration address at the lincoln memorial. `https://www.americanrhetoric.com/speeches/barackobama/barackobamapreinaugurallincolnmemorial.htm`, 2009. Accessed: 03-05-2023.

[49] Obama, B. President barack obama's inaugural address. `https://obamawhitehouse.archives.gov/blog/2009/01/21/president-barack-obamas-inaugural-address`, 2009. Accessed: 03-05-2023.

[50] Obama, B. Remarks by the president at the acceptance of the nobel peace prize. `https://obamawhitehouse.archives.gov/the-press-office/remarks-president-acceptance-nobel-peace-prize`, 2009. Accessed: 03-05-2023.

[51] Obama, B. Announces candidacy for 2012 presidency. `https://www.americanrhetoric.com/speeches/barackobama/barackobama2012prescandidacy.htm`, 2011. Accessed: 03-05-2023.

[52] Obama, B. Remarks by the president at the martin luther king, jr. memorial dedication. https://obamawhitehouse.archives.gov/the-press-office/2011/10/16/remarks-president-martin-luther-king-jr-memorial-dedication, 2011. Accessed: 03-05-2023.

[53] Obama, B. Remarks by the president on osama bin laden. https://obamawhitehouse.archives.gov/the-press-office/2011/05/02/remarks-president-osama-bin-laden, 2011. Accessed: 03-05-2023.

[54] Obama, B. Remarks by president obama at hankuk university. https://obamawhitehouse.archives.gov/the-press-office/2012/03/26/remarks-president-obama-hankuk-university, 2012. Accessed: 03-05-2023.

[55] Obama, B. Remarks by the president in state of the union address. https://obamawhitehouse.archives.gov/the-press-office/2012/01/24/remarks-president-state-union-address, 2012. Accessed: 03-05-2023.

[56] Obama, B. Remarks by the president on election night. https://obamawhitehouse.archives.gov/the-press-office/2012/11/07/remarks-president-election-night, 2012. Accessed: 03-05-2023.

[57] Obama, B. Remarks on the united states response to the ebola epidemic in west africa. https://www.presidency.ucsb.edu/documents/remarks-the-united-states-response-the-ebola-epidemic-west-africa-0, 2014. Accessed: 03-05-2023.

[58] Obama, B. President obama delivers remarks on the 50th anniversary of the selma marches. https://en.wikisource.org/wiki/President_Obama_Delivers_Remarks_on_the_50th_Anniversary_of_the_Selma_Marches, 2015. Accessed: 2023-02-05.

[59] Obama, B. Remarks by president obama in town hall with young leaders of the americas. https://obamawhitehouse.archives.gov/the-press-office/2015/04/09/remarks-president-obama-town-hall-young-leaders-americas, 2015. Accessed: 03-05-2023.

[60] Obama, B. Remarks by president obama to the people of africa. https://obamawhitehouse.archives.gov/the-press-office/2015/07/28/remarks-president-obama-people-africa, 2015. Accessed: 03-05-2023.

[61] Obama, B. Remarks by the president announcing student aid bill of rights. https://obamawhitehouse.archives.gov/the-press-office/2015/03/10/remarks-president-announcing-student-aid-bill-rights, 2015. Accessed: 03-05-2023.

[62] Obama, B. Remarks by the president at the 50th anniversary of the selma to montgomery marches, march 7, 2015. https://www.presidency.ucsb.edu/documents/remarks-commemorating-the-50th-anniversary-the-selma-montgomery-marches-for-voting-rights, 2015. Accessed: 03-05-2023.

[63] Obama, B. Statement by the president on the attack in france. https://obamawhitehouse.archives.gov/the-press-office/2015/01/07/statement-president-attack-france, 2015. Accessed: 03-05-2023.

[64] Obama, B. Honoring the victims of the attack in brussels, belgium. https://obamawhitehouse.archives.gov/the-press-office/2016/03/22/presidential-proclamation-honoring-victims-attack-brussels-belgium, 2016. Accessed: 03-05-2023.

[65] Obama, B. Remarks by president obama and president-elect trump after meeting. https://obamawhitehouse.archives.gov/the-press-office/2016/11/10/remarks-president-obama-and-president-elect-trump-after-meeting, 2016. Accessed: 03-05-2023.

[66] Obama, B. Remarks by the president on the passing of the u.s. supreme court justice antonin scalia. https://obamawhitehouse.archives.gov/the-press-office/2016/02/13/remarks-president-passing-us-supreme-court-justice-antonin-scalia, 2016. Accessed: 03-05-2023.

[67] on Foreign Relations, C. Timeline: The u.s. war in afghanistan. https://www.cfr.org/timeline/us-war-afghanistan, 2023. Accessed: 2022-11-09.

[68] pandas development team, T. pandas-dev/pandas: Pandas, Feb. 2020.

[69] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[70] Pimienta, M., Morse, A., and Walsh, S. Deferred action for childhood arrivals: Federal policy and examples of state actions. https:

//www.ncsl.org/research/immigration/deferred-action.aspx, 2020. Accessed: 2022-11-19.

[71] Pérez, J. M., Giudici, J. C., and Luque, F. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks, 2021.

[72] Qi, P., Dozat, T., Zhang, Y., and Manning, C. D. Universal dependency parsing from scratch, 2019.

[73] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (2020).

[74] Rappeport, A. Democratic speechwriters see obama's selma address as 'among his very best'. https://archive.nytimes.com/www.nytimes.com/politics/first-draft/2015/03/09/obamas-selma-speech-considered-among-his-very-best/, 2015. Accessed: 03-05-2015.

[75] Řehůřek, R., and Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta, Malta, May 2010), ELRA, pp. 45–50. http://is.muni.cz/publication/884893/en.

[76] Reitz, K. requests. https://requests.readthedocs.io/en/latest/, 2023. Accessed: 05-05-2023.

[77] Richardson, L. beautifulsoup4. https://www.crummy.com/software/BeautifulSoup/, 2023. Accessed: 05-05-2023.

[78] Schumacher, E., and Eskenazi, M. A readability analysis of campaign speeches from the 2016 us presidential campaign, 2016. Accessed: 09-05-2023.

[79] Strauss, V. Readability formulas and obama's speech, gingrich's phd. https://www.washingtonpost.com/blogs/answer-sheet/post/can-u-read-this-stick-to-words-not-formulas/2012/01/25/gIQAjM8YRQ_blog.html, 2012. Accessed: 2022-11-20.

[80] Swami, R. All english stopwords (700+). https://www.kaggle.com/ds/1003424, 2020.

[81] Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., Payne, M., Yurchak, R., Russwurm, M., Kolar, K., and Woods, E. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research 21*, 118 (2020), 1–6.

[82] Thornton, J. M. Selma to montgomery march. https://encyclopediaofalabama.org/article/selma-to-montgomery-march/, 2023. Accessed: 03-05-2023.

[83] Wes McKinney. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference* (2010), Stéfan van der Walt and Jarrod Millman, Eds., pp. 56 – 61.

[84] Wikipedia. Readability. https://en.wikipedia.org/wiki/Readability, 2023. Accessed: 03-05-2023, used for readbility equations.

[85] Woolley, J., and Peters, G. The american presidency project: George w. bush event timeline. https://www.presidency.ucsb.edu/node/346007, 2020. Accessed: 2023-02-05.

[86] Woolley, J., and Peters, G. The american presidency project: Barack obama event timeline. https://www.presidency.ucsb.edu/node/348249, 2021. Accessed: 2023-02-05.

FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICS
B-3001 Leuven