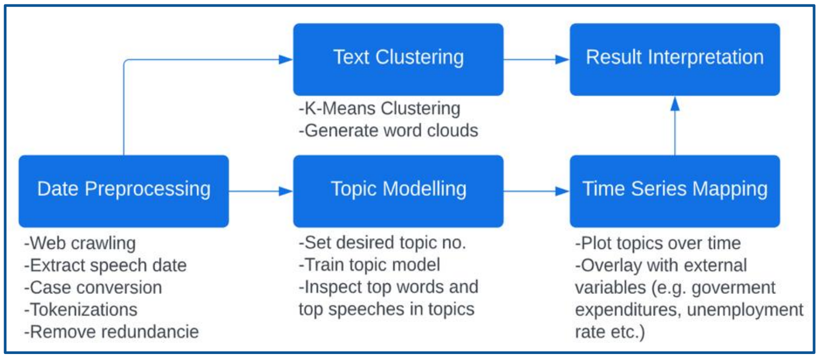
Overview

In this project, we will use various NLP techniques to analyze the speech of Barack Obama during his two consecutive presidential terms from 2009 to 2017. This project aims to gain insights into the major topics covered in the selected speeches and investigate how external factors may influence the frequencies of different topics in Obama's speech. The flowchart below demonstrates the processes involved in our analysis.



Data Processing

For the data collection process, we used the packages *urllib* and *beautifulSoup* in Python3 to extract each speech's web links and content in the given period. We used *urllib.request()* to fetch the URL, and apply *html.parser()* and *soup.find_all()* to extract all content from the "href" tags. Next, we obtained useful links which start with 'speeches' and end with 'htm'. We removed duplicates and added the appropriate prefix to obtain full website links.

We then looped over every URL with the *'lxml' parser*, extracted the title and content of each speech, and saved them into separate lists. Next, we removed the redundant strings such as ">American Rhetoric: Barack Obama:" and "[as prepared for delivery]" from them. In the last step, we find the date corresponding to Obama's speeches using *BeautifulSoup*. The title, content, and date of every speech are now available for further processing and analysis.

We applied the NLTK package to clean the speech content in data pre-processing. Using *WordNetLemmatizer*, we removed the punctuations and changed all contents into lower case. We then used the stopword provided from the package with some extra stop words extension to remove the non-useful content. Next, the *wordNetLemmatizer* changes the word into the original lemma form for nouns and verbs. Finally, we apply the *pos_tag* function to extract the common nouns from the cleaned texts, which will be used as training data in topic modeling.

Topic Modelling

Topic modeling is a machine learning technique that automatically analyzes text data to determine cluster-words for a set of documents. The goal is to abstract "topics" that occur in a collection of documents. We used K-means and LDA techniques to identify the main topics of Obama's speeches and the topics distribution.

In K-means, we tried to find which clusters keep occurring to have high silhouette scores in each K setting based on silhouette analysis. When K equals 15 to 20, the clusters with high silhouette scores can be summarized as economy, war and security, and health care. K-means does not give time tags to the topics. In other words, we cannot know when Obama gave the speeches about, for example, war.

We use the package *mallet* to do LDA. We defined an arbitrary number of topics as 30 and used *quick_train_topic_model()* to train our topic model. We print the top words of the 30 topics to summarize. The 30 topics could be merged into several speech themes, each assigned with clear, meaningful names: employment, health care, climate change, war and violence, and education. The function *load_topic_distributions()* was applied to load the topic distribution for each speech after training a topic model. Finally, we created the time series plot to visualize the topic distribution from 2009 to 2017 to discover the effect of extra variables on the speech topics.

Interpretation

As shown in Figure 1(Gallup, 2017), President Obama's approval rating rose relatively quickly from 41% in August 2011 to 53% in December 2012. During this period, we found that the primary topics of his speeches were war, work, and job, with the topic proportion of war reaching over 20% and the topic proportion of employment and work reaching above 13%. Based on the event timeline (Barack Obama Event Timeline | The American Presidency Project, n.d.), we found that after Al-Qaeda leader Osama bin Laden, who was responsible for the 9/11 attacks, was killed by US forces in Pakistan, a record number of Americans opposed the war. Under the pressure from lawmakers and the public, Obama decided to sizably reduce U.S. forces in Afghanistan, announcing on June 22, 2011, that 33,000 troops would be withdrawn from Afghanistan by the summer of 2012 (Laub, 2017). In October of that year, President Barack Obama announced another strategy to end combat missions in Iraq and remove all U.S. troops by the end of 2011. His decisions and speeches on the war may have been one of the factors that boosted his approval ratings.

In addition, Obama’s job-related activities, such as the introduction of the American Jobs Act in September 2011, may also help to enhance his approval ratings. According to a report by Pew Research Center (2020), Obama's job approval ratings improved considerably in December 2012, corresponding to greater optimism of the public toward the economy. While not all respondents considered the economy to be in good shape, the number of people describing economic conditions as poor dropped to 35%, the lowest since January 2008. 22% of respondents believed that plenty of jobs were available, which is 10% greater compared to early 2010 and the highest since 2008. The corresponding trend can also be seen in the unemployment graph (*Fig. 5*).

General conclusion

Based on the above analysis, we believe that approval ratings may be one of the factors influencing Obama's speech. When his approval rating was at its lowest point in August 2011, his speeches focused on areas of significant public concern (war and job), which was likely to raise his approval rating and establish the groundwork for a smooth re-election in his second term.

Appendix:

Figure 1

Obama Presidential Job Approval 2009-2017

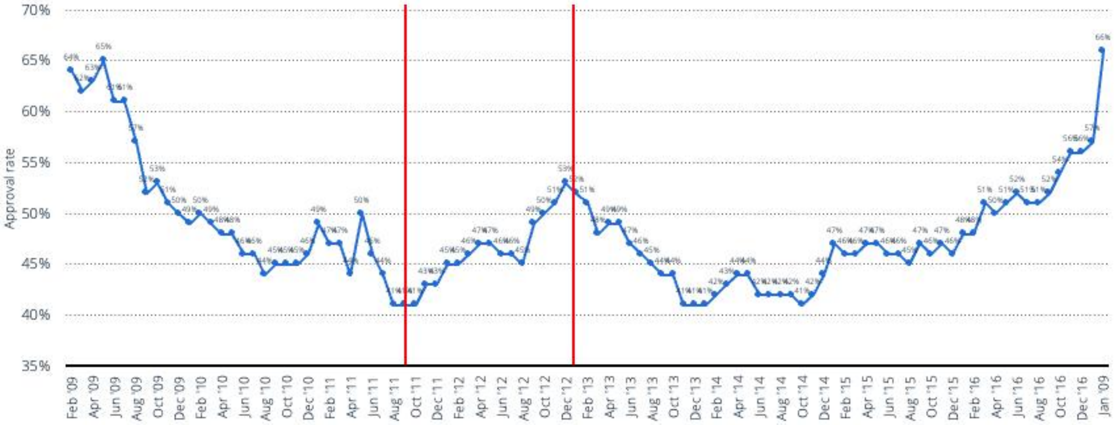


Figure 2

War Topic Proportion Time Series Plot

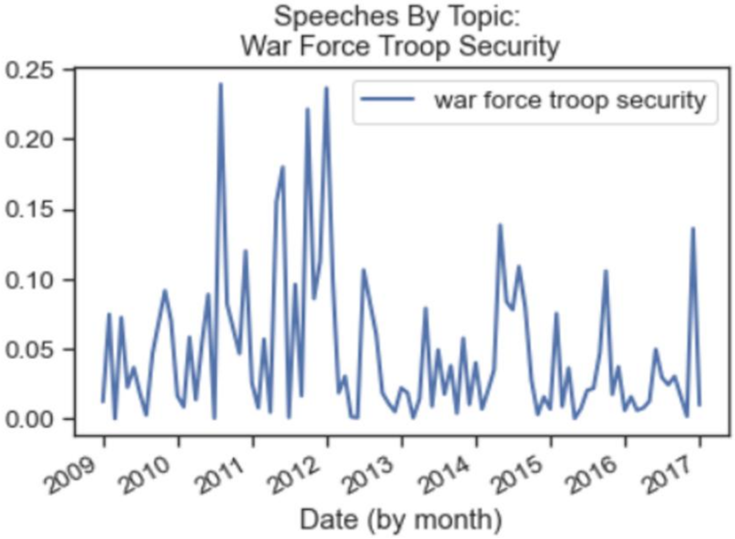


Figure 3

Job Topic Proportion Time Series Plot

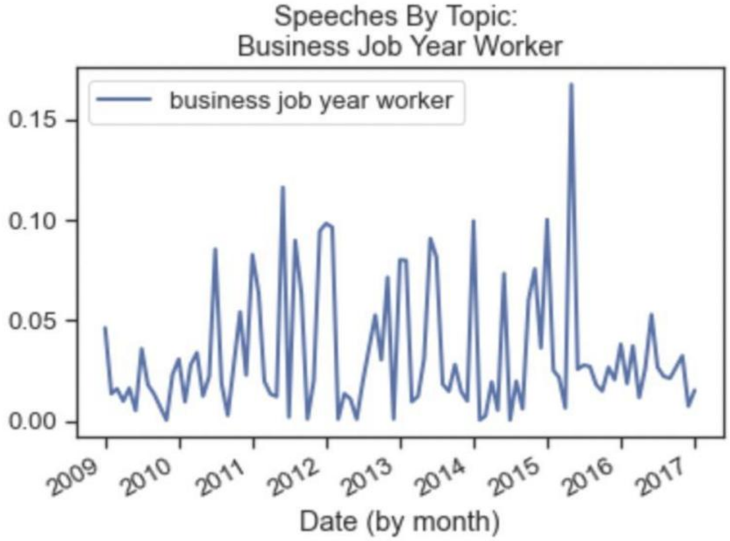


Figure 4

Work Topic Proportion Time Series Plot

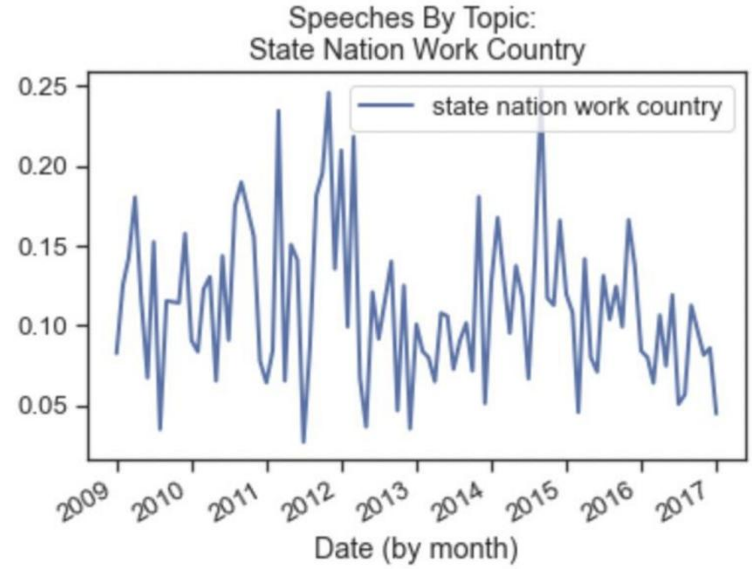
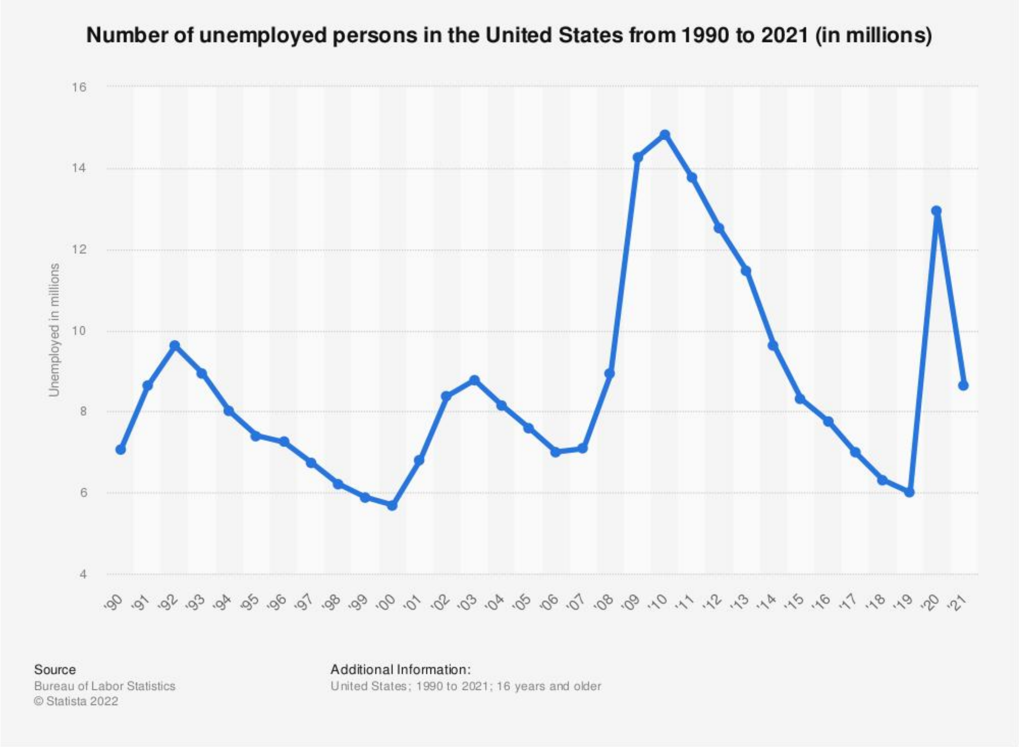


Figure 5

Unemployment Rate from 1990 to 2021



Source
Bureau of Labor Statistics
© Statista 2022

Additional Information:
United States; 1990 to 2021; 16 years and older

Reference:

Mallet: MAchine Learning for LanguagE Toolkit. (2022). Retrieved 7 June 2022, from https://mimno.github.io/Mallet/index

Gallup. (February 21, 2017). Do you approve or disapprove of the way Barack Obama is handling his job as president? [Graph]. In Statista. Retrieved June 07, 2022, from https://www.statista.com/statistics/205284/obama-job-approval-rate-by-the-american-public/.

Barack Obama Event Timeline | The American Presidency Project. (n.d.). The American Presidency Project. https://www.presidency.ucsb.edu/documents/barack-obama-event-timeline.

Laub, Z. (2017, May 1). The U.S. War in Afghanistan. Council on Foreign Relations. https://www.cfr.org/timeline/us-war-afghanistan.

Bureau of Labor Statistics. (February 10, 2022). Number of unemployed persons in the United States from 1990 to 2021 (in millions) [Graph]. In Statista. Retrieved June 07, 2022, from https://www.statista.com/statistics/193254/unemployment-level-in-the-us-since-1990/.

As Fiscal Cliff Nears, Democrats Have Public Opinion on Their Side. (2020, July 28). Pew Research Center - U.S. Politics & Policy. https://www.pewresearch.org/politics/2012/12/13/as-fiscal-cliff-nears-democrats-have-public-opinion-on-their-side/.