DSAN-5400 Natural Language Processing Project Proposal

Erin Brzusek, Nikhil Patla, Vinny Turtora, Mohammad Yassin and Sebastian Villalobos Alva

The Linguistic Chameleon: How Politicians' Adapt Their Speech in the Political Arena

GitHub Repository: https://github.com/np767/5400_Final_Project

Problem:

"All men having power ought to be mistrusted" - James Madison

Public perception of politicians has often been shaped by the belief that public officials adjust their rhetoric to fit the audience they address. Variations across campaign speeches, congressional addresses, and bipartisan events frequently reveal different versions of the same individual, prompting the question of whether political rhetoric reflects genuine convictions or a calculated manipulation of constituents.

This project seeks to quantify this "chameleon effect", that is to say how politicians reshape their language to align with their audience. By tracking five politicians across multiple contexts (partisan rallies, formal congressional floor speeches, and bipartisan events) our analysis will test whether linguistic patterns shift depending on audience and setting. Our aim is to illuminate how politicians tailor their language to maintain popularity and to explore how data science techniques can make political inconsistency measurable and visible.

Related Work:

In *Toeing the Party Line: Indexicality and Regional Andalusian Phonetic Features in Political Speech [1]*, the researchers analyzed the speech of 32 Spanish politicians to examine how linguistic style-shifting ("hyper-vernacular" use of regional features) reflects social positioning and identity construction in political contexts. Focusing on four Andalusian phonetic features associated with working-class speech, the study employed mixed-effects logistic regression and found that the city of origin was the strongest predictor of variation (ie. aspiration and elision are characteristic of Granada, Linares, and Malaga).

Similarly, *Speaking two "Languages" in America: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently [2]* presents a computational semantic analysis of political speeches from Republicans and Democrats in the USA. The study modeled semantic spaces as a function of party, candidate, and time of election. They found that both Republicans and Democrats showed unique and systematic word association patterns across similar topics.

Finally, in *A data science approach to 138 years of congressional speeches [3]*, the study revealed a few key insights: the readability index of speeches increased until the 96th congress then began to decline, speeches have become more positive over time, and there are statistically significant differences between Democratic and Republican speeches.

Proposal:
We will construct a novel dataset composed of speeches delivered by five current and prominent U.S. politicians, collected from the Library of Congress and other public archives. Each politician's speeches will be categorized by context (partisan rallies, formal congressional floor speeches, and bipartisan events) to enable comparison across settings. Using techniques from computational linguistics and natural language processing, we will analyze changes in vocabulary, sentence complexity, tone, and rhetorical framing across contexts. This approach will allow us to quantitatively assess variations in political rhetoric.

Methods to Achieve Objective:
Potential methods include:

- Cleaning the data by removing stop words, removing inflections from nouns, verbs, and adverbs, and lemmatizing words using a POS tagger (http://nlp.stanford.edu/software)
- Linguistic feature analysis (ie. average sentence length, lexical diversity, pronoun use to capture shifts in formality or inclusivity, words used to conclude sentences)
- Artificial neural network as implemented in the *word2vec* mode, including skip-gram and continuous bag-of-word mechanisms (Python's Genism package)
- Computational semantic space modeling to identify whether recurring themes differ depending on audience
- TF-IDF analysis for each setting of speech by politician
- Embedding-based analysis (similarity, distance based, vector spaces, plane spaces) to measure a politician's semantic speech space
- Clustering (Hierarchical clustering, GMM) or classification models (SVMs, Mixed-Effects Logistic Regression, Softmax Regression) trained on one setting (ie. partisan rallies) and tested on another (ie. formal congressional floor speeches) to evaluate speech style

Data Sources:

- Congressional API: https://api.congress.gov/#/
- GovInfo: https://www.govinfo.gov/app/collection/crec
- Transcript Library: https://www.rev.com/transcripts
- Rev Transcripts: Search Results | Rev

Evaluation:
Potential metrics include:

- Statistical Validation
    - Use a paired t-test to determine if the observed differences in linguistic features (ie. average sentence length across settings) are statistically significant (p-value < 0.05)

- Authorship Identification Model
    - Deploy a trained model to predict authorship across settings to determine whether there is a drop in the F1-score or accuracy
- Human Input
    - Have a small panel blindly decide whether paired speeches (A vs B) “sound like the same person” or whether the tone/position changes. Compare human judgments to model metrics.

Bibliography:

1. Toeing the Party Line: Indexicality and Regional Andalusian Phonetic Features in Political Speech
2. Speaking two "Languages" in America: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently
3. A data science approach to 138 years of congressional speeches - PMC