Stephanie Cruz

Natural Language Processing

Table of Contents

[Project Background and Description 2](#_Toc58421069)

[Data Description 2](#_Toc58421070)

[Research Questions 2](#_Toc58421071)

[Initial Assessment and Unknowns 2](#_Toc58421072)

[Cleaning the Data 4](#_Toc58421073)

[Data Analysis Methods 4](#_Toc58421074)

[Analysis 5](#_Toc58421075)

[Predicting Party with Naïve Bayes 5](#_Toc58421076)

[Predicting Party with SVM: Linear Kernal 5](#_Toc58421077)

[Predicting Party with SVM: Polynomial Kernal 7](#_Toc58421078)

[LDA Topic Modeling 7](#_Toc58421079)

[Republican Statements 7](#_Toc58421080)

[Democratic Statements 10](#_Toc58421081)

[Summary of Results 13](#_Toc58421082)

[Next Steps and Outstanding Questions 14](#_Toc58421083)

# Project Background and Description

An analysis will be performed on public statements of United States Senators between the years 2014-2016. This data will be used to analyze and determine whether a senator’s party can be predicted based on their statement as well as which topics republican and democratic senators talk about the most in public statements.

# Data Description

The following

* **Democratic Statements**
  + 33 Senators
  + 20,604 Statements
* **Republican Statements**
  + 32 Senators
  + 14,080 Statements
* **Independent Statements**
  + 1 Senator
  + 420 Statements

All the data was scraped from the public websites of the individual senators.

# Research Questions

The data analysis will answer the following research questions:

1. Can the party affiliation be determined from a public statement?
2. What are the top three topics that republican senators discussed in their public statements?
3. What are the top three topics that democratic senators discussed in their public statements?

# Initial Assessment and Unknowns

After an initial assessment of the data, there are statements from the following 32 republican senators. Some senators make far more statements than others, with Chuck Grassley releasing over a thousand public statements during that time. Alternatively, Richard Shelby released only 90 public statements during the period that the data covers.



There are also statements from 33 democratic senators. Some senators make far more statements than others, with Sherrod Brown releasing over two thousand statements. Two other democratic senators released over one thousand statements. Conversely, a couple democratic senators only released a couple hundred statements.



In addition to statements from both republican and democratic senators, there were 420 statements from the Independent senator, Bernie Sanders.

On average, democratic senators released more statements than republican senators, with democratic senators releasing, on average, 624 statements, compared to only 440 statements on average for republican senators.

# Cleaning the Data

The data obtained was fairly complete and clean. The following steps were taken to ensure appropriate data analysis:

1. Remove the data from Bernie Sanders, the only Independent senator
2. Remove statements from Michael Crapo – for some reason the data that was scraped from this senator’s site was not formatted correctly and could not be parsed.

# Data Analysis Methods

The following methods were used to predict whether statements were made by republican or democratic senators:

1. Naïve Bayes
2. Support Vector Machine with a linear kernel
3. Support Vector Machine with a polynomial kernel

The following method was used to perform topic modeling:

1. Latent Dirichlet Allocation

# Analysis

An in-depth analysis was done to determine if party could be predicted from the statement itself. For each of the senators that was included in the analysis, each statement was read and labeled as being either ‘R’ for Republican or ‘D’ for democrat. A CountVectorizer was created, with 100 as the number of max features, and standard English stop words were used. Further, words less than three characters long were removed. Numbers were removed as well. The data was split into training and testing data, where seventy percent of the data was used to train the machine learning models, and thirty percent of the data was used to test the accuracy of the machine learning model.

## Predicting Party with Naïve Bayes

The first machine learning model that was utilized was a Multinomial Naïve Bayes model. The training data was fit into the Naïve Bayes learning model, and then the model was used to predict whether statements in the test set were Republican or Democratic statements.

Chart

Description automatically generated

The above confusion matrix shows how accurate the Naïve Bayes model was at predicting whether statements that were made were by Republican or Democratic senators. The model accurately predicted the party of the representative about 65% of the time. The model incorrectly predicted Republican where it should have predicted Democrat over 20% of the time, which is problematic.

## Predicting Party with SVM: Linear Kernal

The next machine learning model that was used was Liner SVM model with a C-value of 10 – with the C-value indicating how much misclassifying a training example should be avoided. The Linear SVM model was trained, and then predicted the party for the group of test statements.

A picture containing timeline

Description automatically generated

The above confusion matrix shows the accuracy of the Linear SVM model. The model is only 64.73% accurate. This model performed about the same as the Naïve Bayes model above. Errors, however, are more equally distributed, with about 15% of statements being incorrectly classified as Republican, while around 20% of statements being incorrectly classified as Democratic statements.

Next, a Linear SVM model was used with a c-value of 100 to see if this would increase the accuracy of the model. The confusion matrix for this model is shown below:

A picture containing chart

Description automatically generated

This model was less accurate than our previous models, only predicting about 63% of the statements correctly.

## Predicting Party with SVM: Polynomial Kernal

Lastly, an SVM model with a polynomial kernel, with a c-value of ‘50’ and a degree of ‘3’ was used to predict the party of a senator making a public statement. Once again, seventy percent of the data was used to train this model, and thirty percent of the data was used to test the model.

A picture containing chart

Description automatically generated

The above confusion matrix shows that this model is our most accurate, predicting the party of a representative about 74% of the time. Incorrectly classified statements were also fairly evenly distributed between misclassifying as republican or democratic – with around 13% of statements falling into each category.

# LDA Topic Modeling

Latent Dirichlet Allocation was used to determine what the top three topics are in both republican and democratic statements. The goal here is to determine what the main talking points are for each party and see how those differ between the two groups.

## Republican Statements

Republican statements were initially used to determine which topics were the most likely to be used in public statements. A breakdown about the number one topic for Republican senators is shown below.

Chart, bubble chart

Description automatically generated

For this topic, senator, health, veterans and care were the top four words that were used. The ‘senator’ is probably a reference to themselves or other senators in statements. However, looking at the top words, we can infer that healthcare, veterans and legislation are some things that senators are making statements about for this topic. The second most likely topic is shown below:

Chart, bubble chart

Description automatically generated

For this topic, federal, department and committee are the top words. Other words include, government, agency along with law and energy. So, this topic likely includes discussion of government agencies, along with legal references and discussion about energy. Lastly, the third most likely topic is shown below:

Chart

Description automatically generated

The top word for this topic is ‘president’, along with other common words. So, the third topic most likely to be talked about in public statements was the president.

## Democratic Statements

Next, democratic statements were analyzed to determine the topics most likely to be discussed in public statements by democratic senators. The first topic is shown below:

Chart

Description automatically generated

The first topic revolves around discussing the senate. The word ‘today’ is used to indicate that these statements are talking about current events. Information about the second most likely topic is shown below:

Chart, bubble chart

Description automatically generated

The second most likely topic has the words ‘new’ and ‘schumer’ as the top two words. Senator Chuck Schumer was one senator who had released over 1000 statements over the time period we are analyzing. ‘Federal’ and ‘funding’ are also words towards the top of the list for the second topic, which could indicate that government funding and spending are a topic of these statements. Information about the third statement is shown below:

Chart, bubble chart

Description automatically generated

This last topic from democratic senators seems to involve health care and veterans, with the top three words being ‘health’, ‘care’, and ‘veterans’.

# Summary of Results

Our overall analysis found the following with regards to our Republican and Democratic statements:

* Democratic senators made about 180 more statements per senator than republican senators between 2014-2016.
* Three democratic senators made over a thousand public statements during the time period studied, compared to only one republican senator
* We can accurately predict which party made a public statement about ¾ of the time.
* Both parties talked quite a bit about the following topics:
  + The president
  + The Senate and senators
  + Law enforcement
  + Energy
  + Veterans
  + Healthcare
* The Republican Party talked more about the following topics:
  + Iron
  + Time
  + Military
  + Nuclear
  + EPA
  + Security
  + Senators Rubio and Grassley
* The Democratic Party talked more about the following topics:
  + People
  + Communities
  + Funding
  + Safety
  + New
  + Women
  + Senator Schumer

Overall, democratic senators make more press release statements and tended to talk more about people and communities. Republican senators made fewer press release statements and tended to talk about the military, nuclear energy, the Environmental Protection Agency and Iran. It is generally not surprising that democrats tended to talk more about domestic policy and domestic issues, while republicans tended to speak more to topics of national security. It is also not surprising that certain topics, like health care came up with both parties, given the timeframe - this was shortly after the Affordable Care Act was initially passed, so certainly not surprising that health care was a common topic. It is also not surprising that Veterans came up as a common topic – Veterans are commonly talked about by both parties.

# Next Steps and Outstanding Questions

The remainder of press release statements for the time period being studied should be found and scraped for a more complete analysis. Some senators are simply not represented in this data. Additionally, further analysis should be done to answer the following questions:

* How positive are typical statements be republican and democratic senators?
* Do long serving senators tend to release more negative statements than newly elected senators or vice versa?
* Do democratic senators from red states tend to have statements that look more like statements from republican senators and vice versa?
* Are certain topics seen in senators that are voted out of office in the next election?