Scotus Project 2024 Report

Fatima Mustafa, Alejandra Mercado, Yadira Acosta Tena

University of California - Davis

LIN 127 - Text Processing and Corpus Linguistics

Professor Kenji Sagae

December 6, 2024

The task and its purpose/motivation (Alejandra)

Our topic was to analyze the dynamics of Supreme Court justices by referring to the metadata and transcripts from the SCOTUS corpus. We sought to identify the difference in vocabulary used between conservative and liberal judges as well as gender and age based vocab differences. In addition, we wanted to show how a classifier is helpful but not always accurate when it comes to classifying human speech.

We retrieved metadata from a chosen transcript from 2007 and then tokenized the text with subsets for gender, age, political affiliation. From there, we created a summary of stats for word count. We made sure to focus on their most common words as well as the length of their individual utterances in total. To make the information more digestible, we formatted the data we collected thus far into charts to showcase our findings. Using the fasttext approach as shown in assignments such as HW 3, we created a training file with labels for sentences as either "conservative" or "liberal" learning. We derived these sentence labels using the .cha files and looking into the history of what both conservative and democratic advocates were most likely to support, their stances being used to compare them. With our chosen liberal member being Ruth and conservative member being Robert, we utilized their sentences to create the training file. After the training file was complete, we asked fasttext to predict the labels for the new phrases we had given it to test if the training file assisted in the classifier's accuracy. The classifier from then on was able to predict labels for the new phrases as either "liberal" or "conservative", showcasing its accuracy for each sentence we provided.

The insight we can gain from this topic is useful to identify any systemic bias and how that contributes to the political culture of the US. These differences and bias matter when considering the impact they have on the costly decision governmental institutions like the Supreme Court make which dictate the lives of Americans in future contexts.

Underrepresented populations would benefit from this work when participating in social justice movements such as rallies which may discuss controversial topics concerning factors such as gender and age. The work we found, especially concerning how ambiguous some utterances from the justices could be, would ultimately allow access to results which may foster a more equitable environment where the underrepresented may advocate for their rights where bodies such as the Supreme Court have a large part in what happens to their autonomy.

Background information (Yadira)

An academia journal has taken a similar approach to our project. An honors thesis titled "Gendered Speaking Patterns in Supreme Court Oral Arguments from 1981-2016", focused on the frequency in which a Supreme Justice Court official spoke during oral argumentation cases (Purser, 2018). This research primarily focused on the Supreme Justice Court official's gender and the number of times they spoke in a case to determine if there is gender bias and discrimination within the field (Purser, 2018). Moreover, the academia journal investigated oral

arguments around the same time period as the corpus we analyzed. Their topic relates back to our project in terms of the similarity of the corpus being used, as well as focusing on gender.

We needed to know both of the Supreme Court justices' background information in terms of their age, gender, and political affiliation in order to have extensive knowledge about the corpus we are analyzing. The corpus doesn't openly present their specific background information; however, we believe that this information is critical towards understanding our findings and analyzing our data more efficiently. The files we were analyzing, portray participation from Supreme Court justice Ruth Bader Ginsburg and Supreme Court justice John Roberts. Ruth Bader Ginburg was an elderly woman who was closely associated with liberal beliefs. John Robers was a middle aged man who was mostly associated with conservative beliefs. They had about 22 years of age difference.

We used the SCOTUS corpus, which was compiled of files from the year 1977 to 2014. We specifically chose files from 2007 because we wanted to analyze contrasting language usage between two people of different genders and from polarizing political affiliations. Throughout the entire corpus that contained files from 1977 to 2014, each folder had approximately 100 .cha files. We used metadata and transcripts from SCOTUS corpus, which consists of information relating to the dynamics of Supreme Court justices. We believe that analyzing language in these transcripts would allow us to identify the differences in vocabulary usage given their gender, age, and political party.

We used the 06–1005 case file for 2007. The case file is about 56 minutes long and contains a good amount of data. We only used the utterances spoken by Ruth Bader Ginsburg and John Roberts as a means to compare their use of language throughout the case for the first task. For task 2, using the fasttext model we looked outside the scope of Robert's and Ginsberg's utterances to incorporate a variety of labeled data from other justices. We selected roughly 50 sentences from both justices Robert and Ginsberg as well as justices Alito (conservative) and Breyer (liberal). We then labeled these as conservative or liberal according to the justices' political affiliation to then train the model on these labels and ask if to label new sentences we give it. For task 3, we selected 6 ambiguous sentences from the first three cases from 2007. These three cases contain information on what the main justices were saying, in addition to justices Alito and Breyer. Additionally, we did not assign the name of the judge to the sentence they said in order to test our classifier. The purpose of task 3, was to see if our classifier would accurately label the ambiguous sentences as liberal or conservative. We trained the fasttext model with a great range of sentences and their labels. We hoped that this would help determine our classifier accuracy and if we used enough data to train the model.

The tasks performed, including who worked on which task (Fatima, Alejandra, Yadira)

Fatima: I worked on task 1 to first download the transcripts and unzip the files. From here, I was a bit stuck on how to parse through all the .cha files and input them into a list. However, when I went to office hours I got some help on how to ask a LLM to parse through these files and only

retrieve the word utterances and put them into a list. Using all the 2007 .cha files, I used an LLM to parse and read John Robert's utterances and Ruth Ginsberg's and put them into separate lists. Then, using a similar approach to homework 1 and 2 I was able to get both of their most common 30 words to analyze it further in the report. Then, using matplotlib I created visuals for the top 30 utterances. After this I wanted to use NLTK as we did in the homework to get the part of speech tagging for each justices' common words to analyze how this is affected by their age, gender, and political affiliation. For task 2, I set up the fasttext model approach as done similarly in homework 3 but for this project focusing on labeling my training file liberal or conservative. Alejandra helped find the sentences utterances of the justices to test this approach and after testing it I tried to make it as accurate as possible. I increased the number of epochs to train the model longer so that it can learn better over that time. However, I still noticed the accuracy was not increasing above 50%. Using LLM I explained how I did my fast text model and asked how I can improve the accuracy for these sentences considering I had already increased the number of epochs. It suggested adding a higher learning rate which I implemented along with setting wordNgrams to be 3 that way the model can look at phrases along with words. This trigram provided more context to the model and helped increase the accuracy for the labels. For task 3, I helped Yadira implement the classifier and see how this differs from the fast text model. For the report, I worked on describing how the data was used and the approach done for the tasks. I also worked on inputting the analysis for the results and describing the challenges we ran into along with what went well.

Alejandra: I began by finding the length of the utterances for both Justice Ginsberg and Roberts separately. I utilized HW 3 in conjunction with fasttext to create a classifier that could dictate whether a sentence leaned more towards being "liberal" or "conservative". To create a training file for the classifier, I went through various files in the SCOTUS corpus, paying attention to utterances by Ginsberg, Roberts, Alito, and Breyer. After compiling 100 utterances, 50 being utterances from liberal leaning justices and 50 for conservative leaning justices, I added this to our data in order to successfully run the train_labels.txt file for the classifier. From there, I tested the model's accuracy and created new sentences for the classifier to utilize, such as "drugs are a huge issue in America", utilizing both the knowledge I gained about the topics each justice was speaking about in their cases (drugs, the death penalty, etc.) as well as researching which topics conservatives and liberals are for and against. With this, I added sentences which both tackled some of the topics the sentences I chose to put in the training set as well as some which had some relation to the training set but also contained new information which the training set didn't have a definite answer to, ultimately letting classifier decide whether the utterances were "liberal" or "conservative". From there, I had the classifier predict each sentence's political affiliation through a "liberal" or "conservative" label as well as its accuracy. I also provided the ideas to utilize part of speech tagging to compare how gender impacts politics, using files concerning justices other than Ginsberg and Breyer to create a more comprehensive view of each party's stances for the classifier, and the use of ambiguity in utterances to showcase how a classifier will

attempt to provide a response to what it is being asked, as discussed in lecture, even if it jeopardizes being incorrect in its answer.

Yadira: I contributed to the group project by working on task 3 of our Colab notebook, explained the background information part of our project, discussed the results for task 3, and wrote about ethical considerations. I worked on integrating information about the work that was done prior to our project. I provided background information about the judges' approximate age gap, gender, and political affiliation. This helped us attain more knowledge about the justices we were using in our project since the SCOTUS database didn't provide their age and political affiliation. I thought that knowing this information would allow us to determine the differences in their language use in cases we investigated and be able to determine their political labels through our classifier. Fatima provided her assistance and helped me implement the classifier in the third task. Before testing the classifier, I observed the SCOTUS database and solely focused my attention on what judges Ginsburg, Roberts, Breyer, and Alito were saying in the first three oral arguments. I went through the first three oral argument cases to study, read, and analyze the dynamics of language use throughout the cases. It was important to read their statements carefully, as it would help me choose ambiguous sentences that I would need to integrate in our classifier. Each case was centered around a particular topic, which would help our classifier distinguish the use of language from each court justice's statements in the cases. Case 06/1005 was concentrated on money laundering. Case 06/10119 focused on a murder. Case 06/1037 talked about retirement laws. Based on the background knowledge from each judge and the statements they made in each case, we would be able to see how accurate our classifier would label the statements. I then inputted the selected sentences into our classifier to determine our classifier's accuracy. The classifier then labeled the sentences either "conservative" or "liberal" based on its prior knowledge from task 2.

Analysis of the work and its results, including a discussion of limitations, unforeseen difficulties, etc. (Alejandra)

Due to our data originally being a .cha file, we have had trouble finding a way to read our file since we had not dealt with this format before. We ran into trouble in an attempt to read the information in the file due to our team having little to no previous experience with conversion utilizing the methods covered in class. With our limited knowledge, we attempted the following code after uploading our SCOTUS corpus directly to Colab, titled "OralArguments.zip":

```
Python
!ls
!unzip OralArguments.zip
!unzip OralArguments.zip -d /content/
```

```
!ls /content/
```

However, we found this method to be invalid since we could not read what the .zip stated as well as being unable to use the data in a way that we have learned in class, such as through .csv. It also did not function unless we reuploaded the file to Colab's Google Drive each time we ran our code after an extended period of time. Our confusion therefore led us to attempt to convert our list into a .csv file to begin with instead of working with the .cha file.

Analysis of the work and its results, including a discussion of limitations, unforeseen difficulties, etc. continued (Fatima)

After attending office hours, we learned to use an LLM to help us parse through these files for our tasks. Then we decided to look at all the 2007 files specifically focusing on utterances from Ginsberg and Roberts. What worked well here was finding their common utterances and plotting visuals for these. Likewise tagging these top 30 utterances based on their part of speech was fairly straightforward as we used the methods we learned from the homeworks to help us with this. Yet, once we approached task 2, which was the fasttext model approach we kept running into trouble. At first, we wanted to create an empty file for our training set and edit it directly in the file then proceed to training it. However, we kept running into runtime errors and such where it would not let us train without inputting something into the file at first. We then decided to give the file some labels at first and once we ran that section of code, we were able to open the file and edit in more of our labels without crowding the entire cell block. Once we did this, we went back and ran the block again to save these changes and then proceeded to train our model again using 100 epochs. We realized that when we were using less epochs, it wasn't training this well enough. This is because the data needs to be trained over a longer period of time to pick up on the patterns and identify how these labels (conservative vs liberal) are being assigned to the sentences. The next issue we then ran into was while it was labeling some sentences we asked it to label correctly, some were still incorrect and the accuracy rates were around the mid 50% when we wanted them to be near 100%. We asked an LLM how we can improve this training and it suggested adding a learning rate of 1.5 and once we did this we found that the accuracies went up higher than before. Another change the LLM advised us to do was add wordNgrams=3 which would look up to 3 words and this ended up working well for our accuracy rates. For task 3, we wanted to use openAI and ask the LLM to guide us through using this method to build the classifier. However, we kept trying and we realized that there were issues with the openAI packages not installing correctly. Hence, it would not let us use gpt4 which is why we switched to another method the LLM told us about. This method was to build a pipeline using transformers. Once we implemented this we were able to give it our ambiguous sentences and determine if the classifier was labeling them correctly compared to fast text.

Main Results (Alejandra)

We were able to showcase the output of the 2007 file containing the utterances of both Roberts and Ginsberg in order to further analyze it. From there, we were able to extract the utterances of both Roberts and Ginsberg from the file to print both their utterances, robe_utterance for Roberts and gins_utterance for Ginsberg. Since their utterances were separated, we found each justice's utterance length and top 30 most common words. For Ginsberg and Roberts, their utterance length was 23187 and 29307, respectively. For their most common words, they shared determiners such as "the", "you", and "that" but also showed high amounts of words that the other justice did not utilize. For example, Robert's utilized the words "mister" and "mr" frequently enough for the words to appear in his top 30 words compared to Ginsberg where these forms did not appear. From there, we displayed the data of each justice's 30 most frequent words to create a graph in order to depict the information better. Then on, we labeled the words for each justice for their part of speech. We created a training file for our classifier utilizing phrases from four justices, two from both the liberal and conservative party, to have enough data for an accurate classifier. We tested our classifier, gave it new sentences for it to classify, and tested the classifier to which it accurately labeled the sentences as either "liberal" or "conservative". In task 3, our classifier accurately determined the labels for two out of the six sentences. In regards to the sentences that were inaccurately labeled, it's important that we notice the predicted label score. The predicted label scores for the inaccurate labels were significantly lower than we expected. We also noticed how the scores could have skewed to either "conservative" or "liberal." We believe that the classifier was uncertain of how to label the ambiguous sentences. The classifier may have relied on our trained model to decide what to label each sentence. We were surprised by the sentence "They may have enormous gross revenue, but they may have they may have enormous expenses overseas" because it had a liberal score of 0.825 when Justice Alito is conservative. This suggests that our training model may have needed more utterances to improve the accuracy of our classifier.

Our results accomplished our main task by finding the differences in speech depending on factors such as gender, age, and political affiliation in government positions. Finding the length of both justice Ginsberg (liberal) and Roberts (conservative) utterances while finding that Ginsberg spoke less than Roberts speaks to the disadvantage women have in positions of power, with her being older than Roberts adding to the discrepancy between their total word count. Age may speak to how the value of women as well as what they say diminishes overtime, possibly contributing to less opportunities to exhibit their power in society. Our result found that factors such as gender and age could be a reason as to why Robert's word count was drastically more than 6,000 words higher than Ginsberg. After finding each justice's top 30 words, we concluded that Ginsberg's higher frequency of interrogative words such as "what" could reflect how women attempt to continue a conversation in positions of power where men are the majority (the remainder of the cabinet Ginsberg was apart of in 2007 was comprised of men) rather than commencing the conversation. In addition, Robert's high frequency of words such as "mr", a

noun which the part of speech tagger determined, did not appear in Ginsberg's top 30 words, further reflecting the traditional view of respect which conservative leaning members seem to utilize more often than liberal members. Our classifier then was able to classify each sentence accurately by either "conservative" or "liberal" based on the training data we gave it from the SCOTUS corpus, showcasing the differences in how each political party distinctly uses their vocabulary to better convey their message about their policies and values. Finally, utilizing sentences that are more ambiguous also showcases the classifier's need to classify something into either "liberal" or "conservative" when it may be difficult to know which political party representative stated the utterance. The classifier would rather give a response that is entirely incorrect than state that the sentence is difficult to classify.

Since we are using a classifier and not an LLM, a limitation for our project was that it was a lengthy process when training the classifier for the best accuracy. The accuracy was dependent on how much time we spent researching what training data to use for the NLTK tagger. The LLM would have provided a faster experience for finding our results. Though the LLM would have been less reliable, time was a limitation when attempting to create an accurate classifier.

Ethical considerations (Yadira, Alejandra, Fatima)

We believed that by integrating sentences that we believed were ambiguous, we would ultimately cause the classifier to mislabel the sentences. As a result, our classifier incorrectly labeled more than half of the sentences we chose to use. If we chose to use sentences pertaining to the political views of the justices like we did in the training model, we believe that our classifier would have been able to determine the correct labels. The sentences we gave the model led to unintentional biases with how we chose our sentences, for example, utilizing sentences that are more "liberal" or "conservative" depending on our own beliefs to find a result which best fits our perceptions of the party which may align more with our opinions. It is vital that the training data is reviewed thoroughly to ensure that the user who is creating the model is not biased towards the data inputted. Therefore, it's essential to have balanced data targeting both liberal and conservative labels to ensure there aren't skewed outcomes. Therefore, this can add to stereotypes relating to political affiliation, age, or gender that create "noise" as the model trains resulting in inaccurate labeling.

Conclusions, possibly including possible extensions of the work (Alejandra)

From our work, we can conclude that biases are present when considering gender and age in the workplace, specifically positions in government as demonstrated in our tasks. Due to a lack of time, we would have wanted to utilize the fasttext model by inputting a word and seeing what word output it would give in relation to their accuracy in relation to one another. In the future, we could consider expanding our data to other positions of power and job prospects, such as medicine, to make our argument about gender and age differences more comprehensive.

References

Purser, G. (2018, May). Gendered Speaking Patterns in Supreme Court Oral Arguments from 1981-2016. https://aquila.usm.edu/cgi/viewcontent.cgi?article=1586&context=honors_theses