Applying Neural Networks and Topic Extraction to Study Speech by Gender and Ethnicity in the United States Congress

Faria Mardhani and Yulia Zamriy | W266 @ UC Berkeley | Project GitHub

Abstract

As opposed to previous research on Congressional speech, we focus our research on studying demographic differences in speech regardless of Party affiliation. We use the NLP toolkit for speech classification and topic extraction. We build an ensemble of Multi-Layer Perceptron and CNN models to classify Congressional speeches by gender and ethnicity, achieving roughly a 76.5% accuracy on test sets. To deepen our understanding of the differences between these groups in terms of speech, we apply Latent Dirichlet Allocation models across all Congressional speeches. The topics and themes addressed between genders and between races are distinct, confirming that Congressional diversity is critical to equal representation of all salient issues.

Introduction

The 116th Congress has smashed records for the number of women and minorities in Congress, but the numbers are still incredibly skewed. As of this year, women make up roughly 24% of Congress, while being 51% of the population, and whites make up 78% of Congress, while only being 61% of the population.

With representation of minorities so low, it follows that issues that affect minorities more might not get enough weight or attention in Congress. Moreover, when the issues do come up, there may just not be enough representation from affected groups to provide their perspective. Pearson et al. [6] found that women in Congress spend significantly more time in their speeches discussing women than men do.

In this paper, we develop a set of neural network classification models and topic extraction models to conceptualize how distinct the speech of racial minority and female Congress members are from their white male counterparts. The ability of the classification model to classify speech by gender and race will tell us how significantly speech differs between these groups, while the topic model will give us more insight into how, specifically, their speech differs.

Background

The primary inspiration for this research was Gentzkow et al. [1] where authors studied changes in partisanship (ideological polarization) in Congressional speech using bigrams and multinomial model of speech. To aid the analysis, the researchers overlaid topics to get a better understanding of the relationships in the data, but those topics were manually encoded. In general, extracting polarizing ngrams is the predominant technique of feature engineering in political speech classification studies that used SVMs [4], [5] Naive Bayes [5], and correlations [2]. In terms of evaluation metric, they mostly focused on accuracies for the classification tasks. The success has been mixed: Diermeier at el. [4] reported 92% accuracy in party classification, while Yu et al. [5] measured accuracy by year with best accuracy of around 88% for one of the years.

Analyzing differences in speech by gender has also received significant attention. Pearson et al. [6] used hypothesis testing to determine if Congresswomen speak more often, especially during important debates. Pearson et al. [7] focused on rhetoric and thematic aspects of language used by Congresswomen. They used multiple regression to conclude that sex differences have stronger influence on the likelihood of using gendered rhetoric than partisan differences.

Understanding topics is one of the key elements of political speech analysis. Quinn et al. [3] built a statistical learning topic model for legislative speech and identified 42 topics, but also emphasized that it should only be used for supplemental exploratory analysis, and not as a main tool in researching political themes. Kaufmann [8] took a different approach, and used multiple regression to study the relationship between partisan identification and gender-specific cultural issues. They found that topics such as reproductive rights, female equality, and legal protection for homosexuals are primary determinants of partisanship, but the relationship is not as direct for men and their stance on culturally divisive issues.

In contrast to previous research, we are taking the study of congressional speech classification away from party affiliations and into understanding the relationship between language and demographic diversity as represented by binary[1] class memberships in gender and ethnicity (white vs. non-white). Neither of these categories received proper quantitative attention in congressional speech research (at least to our best knowledge). Moreover, we will step away from statistical tools such as regressions and SVMs, as well as extensive feature engineering to extract polarizing words and phrases. Instead, we will focus our attention on using the natural language processing toolkit (word embeddings, multi-layer perceptron and convolutional neural networks) for classification.

After the classification, we will use topic modeling with Latent Dirichlet Allocation [12] to try and interpret the results as well as identify important topics that get disproportionately less coverage in Congress due to lack of diversity. We hope that if these topics are indeed raised more often by marginalized population categories, then as the Congress becomes more diverse, it will start paying more attention to the issues of those most in need.

Data Overview

Datasets used

1. Congressional Record for the 97th-114th Congresses: Parsed Speeches[2]
2. Wikipedia for the information on African-American, Hispanic and Latino, and Asian Americans and Pacific Islands Americans in the United States Congress
3. Congress.gov for the remaining information on Congresspeople (for example, birth year)

Data Overview

In total, we used 2.8MM speeches made by 1,790 Congresspeople across 18 congresses (1981-2016). Within this framework, women delivered 8.1% of speeches, while racial minorities delivered 8.6%. The representation of our target groups (women and people of color) have increased from single digits to almost 20% in 35 years. Speech statistics have also changed across all segments: on average, Congresspeople make fewer speeches that are longer.

The majority of our target population was members of the Democratic party. However, based on their speech statistics (speech count per person and word count per speech), both groups (women and people of color) significantly differ from an average Democratic representative. Hence, party membership cannot be the strongest determinant of used speech patterns. See Appendix 1 for more details on data used.

Data Preparation

Sampling

While building modeling samples for each target group, we used the following rules[3]:

1. Exclude speeches with fewer than 30 words (arbitrary choice), which should take about 15-20 seconds to voice (see Appendix 2 for an example of a short speech). This increased median word count from around 40 to 230-250 words.
2. Use 50%/50% split between target and non-target groups (and use all available target group speeches).
3. 60%-20%-20% split for training-validation-testing with random splits on speech level, not speaker level.

Preprocessing

We used multiple types of speech preprocessing to test in our models:

1. In Multi-Layer Perceptron Models:
    a. Vectorized unigrams and bigrams of full speeches using sklearn TfidfVectorizer and selected top 10,000 for modeling based on f1 score with sklearn SelectKBest
    b. Split speeches onto 30-word chunks[4] and encoded them with Universal Sentence Encoder[5] [15]
2. Convolutional Neural Networks:

---

[1] We understand that gender and definitely ethnicity are not binary. But currently there is not enough Congressional representation to measure these classes in higher dimensionality,

[2] Data was available since 43rd Congress, but we decided to limit ourselves to a period where there was a significant number of our target group representatives (women and racial minorities)

[3] The detailed statistics on sampling can be found here (tabs “Descr file” and “Models”)

[4] We tried different pre-processing methods with 30 and 50-word chunks and made our decisions based on processing times for each method.

[5] We kept only 100 out of 512 elements of the embedding due to memory issues.

a. Tokenized speeches that were truncated at 750 words[6] with Keras Tokenizer
b. Split speeches onto 50-word chunks and tokenized them with Keras Tokenizer
c. Split speeches onto 30-word chunks and encoded them with Universal Sentence Encoder

3. Topic Modeling with Latent Dirichlet Allocation:
   a. Removed stop words and lemmatized remaining words (using nltk library)
   b. Created bigrams and trigrams (using gensim phrasers)

In the cases where we split speeches into chunks, the final prediction for a speech was done based on simple average probability across all chunks belonging to the speech.

Modeling Approach

Classification

When we began our work on classification models, we ironically were not hoping to achieve 100% accuracy. It would have been disheartening to learn that speech across genders and race is so distinct. However, we were hoping to beat 50% accuracy with a sizeable margin (our modeling samples were designed to have 50% of target population). Another reason not to expect very high accuracy was that a large volume of speeches are procedural statements and comments that do not allow much room for self expression or opinions (see Appendix 2 for examples).

Our first modeling approach was to build a simple Multi-Layer Perceptron using ngrams as features. Without much tuning, this model achieved 75.6% accuracy for Gender and 75.5% for Ethnicity models on validation samples. As we tested multiple pre-processing techniques and more model types (details in the Appendix 3), the accuracy on the training sample increased to around 80-81%, but the accuracy on the validation stayed about at about the same 73%-75% levels. Moreover, the models would converge after only 5-10 epochs depending on the chosen parameters.

However, after comparing prediction distributions across models (see Appendix 4 for selected models), it became apparent that they might be capturing different drivers of class separation. We tested all possible combinations of implemented models, and the best performing ensemble was consistent for Gender and Ethnicity models: MLP-ngrams + CNN + CNN-chunks to deliver 76.7% accuracy on the validation set with almost no drop on the test set: 76.5% (see Appendix 5).

Topic Modeling

In addition to predicting race and gender of Congress members based on speeches, we wanted to understand how exactly Congress members' speeches differed in terms of topics they addressed. Our topic modeling analysis was a quest to answer a few questions that we put forward in our proposal: 1) Given that there is gender and racial imbalance in Congress, are there some salient topics that are not getting proportionate coverage in Congressional speeches? 2) Does the category that comprises most Congressional members (white men) adequately address topics that are often brought up by women and minorities; e.g., are the topics raised by most Congress members an accurate portrayal of the issues that are important across the country?

We conducted our topic modeling using Latent Dirichlet Allocation (LDA) [12], a generative probabilistic model which uses probability to predict the topic(s) underlying a certain document. We built two separate models for gender and ethnicity using the gensim package on our validation sets. In each of these sets, our target group comprised 50% of the dataset to override any natural underrepresentation of topics.

In order to find the optimal model, we trained several models on variations of a few parameters. Firstly, while preprocessing the speech, we only retained unigrams to trigrams that appeared at least 30 times across all of the data and had a normalized pointwise mutual information score[7] of at least 0.5 (with scores ranging from -1 to +1 [13]). We varied the number of iterations and passes on the data, with each model incrementing the previous by 5, up until 20 passes/iterations. The number of topics in each model ranged from 1 to 36.

We found that coherence[8] hit 0.50 at around 11 topics, slowly increased to 0.55-0.56, and maximized at 36 topics. Though the models producing 36 topics had higher coherence scores, we found that certain topics were related to each other in these models, and probably could have been combined into just one topic. We also found that by reducing the number of topics (say to 21 or 26) we lost some salient topics that we felt were important to include. While the slight

[6] That was around 90% percentile of word counts per speech.
[7] We used this metric as it was easier to interpret results.
[8] We used C_v coherence score for model evaluation as according to [14] it produces the best results (though at high computational costs)

increase in coherence wouldn't necessitate us to use the 36-topic models, we agreed on including 36 topics to ensure that relevant topics remained in the dataset.

Our final models for both the gender and ethnicity validation data were built on the maximum number of iterations (20), passes (20), and topics (36). Ultimately, our topic models yielded a list of topics ranging from "War and Defense" to "Healthcare", the "Judicial System", and "Natural Disasters" (Appendix 6). There were also some topics composed of generic verbs, common Congressional phrases, and positive words. They were all retained in further analysis.

Discussion of Results

The first step of our analysis was to determine if there is significant difference in speeches made by women and by racial minorities. As was demonstrated in the Modeling Approach section, the developed models can detect the gender or race of the speaker with much higher than chance probability. Now, we try to isolate the factors contributing to that distinction. Within the context of this language-based project we focus on two perspectives: speaker's choice of words and choice of topics.

Top differentiating words

Our final classification models are combinations of three different types (MLP, CNN and CNN-chunk). Hence, it is hard to extract a concise list of feature importance. Our solution is to break all speeches in the validation files into bins based on the predicted probability that the speech was made by a woman or a person of color. Then we use f-1 score to inspect top differentiating words (see Table 1).

The results suggest that the models assign higher probabilities to speeches that contain identifiers for corresponding target groups: "woman" and "black" (in contrast, "men" and "white" do not make an appearance as top words in low probability speeches). These words are most likely used in the context of broader topics (similarly to "communities", "access", "minorities", "civil" etc.). On the other hand, it is hard to hypothesise why "ensure" made the top of the list in the Gender model.

An interesting insight is that the top 10 words in the Ethnicity model drive a lot more distinction between speeches in low vs. high probability bins. It is possible that people of color bring up topics that are more underrepresented in Congress compared to women.

Table 1. Top 10 most differentiating words within each model

Gender				Ethnicity			
	Speech count (validation sample)				Speech count (validation sample)		
Word	prob < 0.4	prob > 0.6	F1 score	Word	prob < 0.4	prob > 0.6	F1 score
ensure	**2,025**	**3,809**	**98**	**black**	**411**	**5,525**	**716**
california	**2,394**	**4,923**	**83**	**african**	**165**	**3,246**	**524**
woman	**822**	**2,109**	**67**	**caucus**	**335**	**2,287**	**496**
proud	**1,795**	**3,071**	**65**	**consume**	**465**	**1,462**	**469**
unanimous	2,502	1,997	62	senators	4,338	1,069	343
communities	**2,145**	**3,945**	**58**	**civil**	**1,284**	**3,860**	**326**
ought	1,700	963	57	**minorities**	**119**	**1,249**	**321**
distinguished	2,727	2,090	56	**gentlewoman**	**402**	**1,586**	**320**
access	**1,876**	**3,493**	**53**	**congressional**	**2,665**	**4,543**	**317**
womens	**623**	**1,964**	**53**	printed	1,711	307	312

Topic Modelling Results

Our topic models revealed a variety of mostly non-overlapping topics covered by Congress members of different genders and different races. In the gender model, the top topics predicted to be brought up by women were "Healthcare and Women's Health", "Poverty and Welfare", "Children's Diseases", "Economy", and "Natural Disasters." Meanwhile, the top topics related to speeches with lowest probability of being spoken by women were "Costs", "Common Phrases", "Legislation", "Agriculture", "Diplomacy and Security".

In the ethnicity model, the top topics predicted to be covered by people of color were “Art and Culture”, “Great Nation”, “Common Congressional Phrases 5”, “Children’s and Family Programs”, and “Economy”. On the other end of the spectrum, we have “Foreign Trade”, “Energy”, “Common Congressional Phrases 4”, “Random Verbs 2”, “The Navy.”

Some interesting insights can be gleaned from this analysis. Firstly, we notice that both women and minorities tend to touch more upon minority issues relating to poverty/welfare, education, women’s rights, protections and programs for children, and healthcare. Since women and minorities have small numbers in Congress, it’s very possible that these topics are grossly underrepresented. Secondly, we notice that some of the topics more likely to be covered by men and white Congress members are “Common Phrases” or “Common Congressional Phrases.” From our research through the speeches, we believe that this is because generally Congress members who use these phrases are the ones presiding over a session (generally ones with a higher post/position of power). Therefore, in this seemingly innocuous topic of “Common Phrases”, we might actually be seeing a power imbalance between genders and races.

In Figures 1 and 2, we examine the top topics for women and members of color over time, overlaid with the actual presence of women and members of color in Congress. We notice that most of the attention these topics receive is cyclical over time. On the other hand, such topics as “Art and Culture”, and “Poverty and Welfare” sadly have been receiving very little attention even in the most recent Congresses in the dataset despite more diverse members.

Figure 1

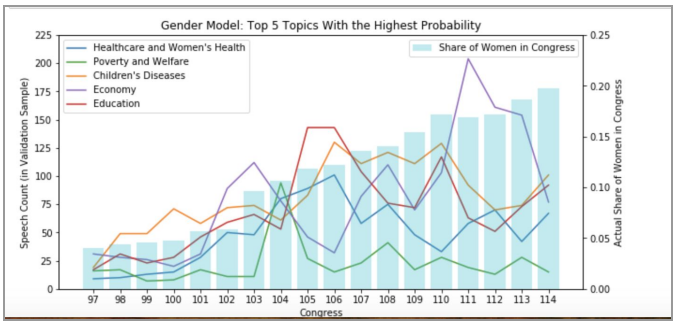
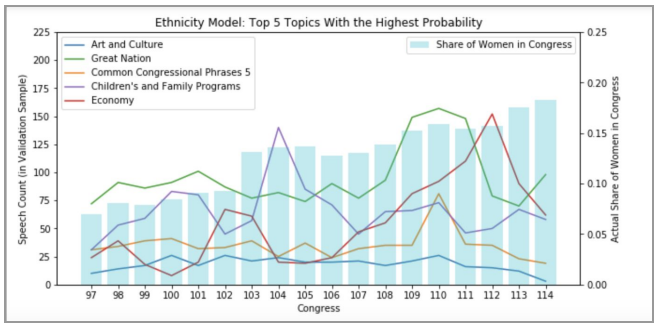


Figure 2



Speech Misclassification Analysis

Before we discuss misclassification, it’s worth noting that misclassification is not necessarily a bad thing in our project. False positives, especially with high probabilities of being in our target groups, potentially mean that people outside target groups bring up topics that are more common in our target groups: “Education”, “Economy”, “Children's and Family Programs”, etc. False negatives, on the other hand, can be the speeches that are procedural and do not contain any opinions or agenda (unless the agenda is to fill time). This is supported by the large number of speeches with such primary topics as “Common Congressional Phrases”, “Random Verbs” etc.

Figure 3. Speech Misclassification Distribution

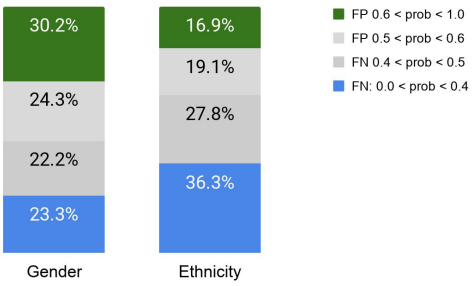


Figure 3 contains information about distribution of errors. Based on this chart, most misclassifications fall into mid-probability area: around 47% of erroneous class assignments have probabilities between 0.4 and 0.6. There is also significant confusion among ensemble models within this segment as they reach consensus only for 20% of speeches in both Gender and Ethnicity models. In this segment of misclassification the most frequent topics are “Random Verbs”, “Positive Words”, and “Common Congressional Phrases” in both models (more details in Appendix 7).

Despite similar overall shares of misclassifications and rates of confusion in Gender and Ethnicity models, the distribution of misclassifications is quite different. The Ethnicity model has a much higher share of false negatives (blue) compared to false positives (green). We were not able to explain this through topics.

Speaker analysis

Most of the work in this project was intentionally focused on speeches with very little attention to speakers. However, the results of the speaker-level analysis are worth additional discussion.

Our Gender validation sample consisted of around 62,716 speeches (50% female) and 1,702 Congresspeople (13% female). When speech-level probability is aggregated to speaker level by taking simple average, correct gender classification rate is 94%. The model makes only one error at classifying congress women (99.5% correct), and classifies 93% of congressmen correctly (all with low number of speeches per person).

On the Ethnicity model side, the validation sample included 62,840 speeches (50% by people of color) corresponding to 1,702 Congresspeople (13% people of color), and speaker-level correct ethnicity classification rate is 87% (with 85% accuracy within target class and 87% within the non-target class).

Such level of accuracy on the speaker level potentially tells us that, despite significant variation in topics and words across speeches, Congresspeople with similar backgrounds share similar language patterns. However, this aspect of the study requires more in-depth analysis to be deemed reliable.

Conclusions

We found that our classification accuracy was maximized by creating an ensemble of Multi-Layer Perceptron and CNN models. This model performed consistently across Gender and Ethnicity, and delivered roughly a ~76.5% accuracy on the validation and test sets. We also explored a speaker-level analysis by taking the average probability score of all the speeches by a certain speaker, and classification of speakers improved dramatically. In our Gender model, the models classify women correctly 99.5% of the time, and men correctly 93% of the time in validation sample. In the Ethnicity model, the target group (members of color) is classified correctly 85% of the time, and the non-target group is classified correctly 87% of the time (also on validation sample).

Through LDA topic modelling, we were able to distinguish the set of topics most often addressed across our target and non-target groups. We found that there were distinct difference in themes covered by women vs. men, as well as themes covered by racial minorities vs. white members. This indicates that certain themes are likely marginalized in Congressional speeches, because the presence of women and people of color in Congress is much lower than the overall nation. This also confirms the idea that a majority of Congress members (white men) may not be holistically addressing issues that face the nation.

Future Work

While working on the project, especially, towards the end, we have accumulated a long list of items to try next. For example, we would build topic models before classification to separate formulaic speeches (with common congressional phrases and random verbs) and those that relate to actual issues (economy, health care etc.) and use them as two separate samples for modeling gender and ethnicity. We could also augment classification models with additional features representing topics and their contributions to speech. Enriching the list of features could potentially help us amplify different aspects of speeches and reduce misclassifications.

Another limitation of this project is that we did not capture sentiment of speeches: is a speaker raising the topic they support or object? If we were to add that along with voting records, it would make for a revealing study of how well speeches are aligned with voting actions.

Lastly, it is still not clear how exactly the ensemble models are able to classify gender and ethnicity so accurately at the speaker level. With additional time and resources, we would study whether the models' accuracy is related to the topics each group addresses or simply language patterns/rhetoric that may be different across groups. From our analysis, it does not seem that the ensemble is properly separating speeches by topics as there is still significant overlap in most problematic mid-probability areas; therefore, it would be compelling to study what exactly is driving this prediction accuracy.

References

[1] M. Gentzkow, J.M. Shapiro, M.Taddy. *Measuring polarization in high-dimensional data: Method and application to congressional speech*. National Bureau of Economic Research (2019). Link.

[2] J. Jensen, E.Kaplan, S. Naidu, L. Wilse-Samson. *Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech.* Brookings Papers on Economic Activity, Fall 2012 Link.

[3] K.M Quinn, B.L. Monroe, M. Colaresi, M.H. Crespin, D.R. Radev. *How to Analyze Political Attention with Minimal Assumptions and Costs*. Midwest Political Science Association (2010), pp. 209-228. Link.

[4] D. Diermeier, J-F Godbout, B. Yu, S. Kaufmann. *Language and Ideology in Congress*. B.J.Pol.S. 42, pp. 31–55 Copyright r Cambridge University Press, 2011. Link.
[5] B. Yu, S. Kaufmann, D. Diermeier. *Classifying Party Affiliation from Political Speech*. Journal of Information Technology & Politics, 5:1 (2008), pp. 33-48. Link.
[6] K.Pearson, L. Dancey. *Elevating Women's Voices in Congress: Speech Participation in the House of Representatives.* Political Research Quarterly. Vol 64, Issue 4, 2011. Link.
[7] K.Pearson, L. Dancey. *Speaking for the Underrepresented in the House of Representatives: Voicing Women's Interests in a Partisan Era*. Politics & Gender, 7 (2011), pp. 493 –519. Link.
[8] K.Kaufmann. *Culture wars, secular realignment, and the gender gap in party identification*. Political Behavior, Vol. 24, No. 3, September 2002. Link.
[9] J.M. Box-Steffensmeier, S. De Boef, T. Lin. *The Dynamics of the Partisan Gender Gap*. The American Political Science Review Vol. 98, No. 3 (Aug., 2004), pp. 515-528. Link.
[10] J. Grimmer. *A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases*. Political Analysis (2010) 18: pp. 1-35. Link.
[11] P. Ban, J. Grimmer, J. Kaslovsky, E.West. *A Woman's Voice in the House: Gender Composition and its Consequences in Committee Hearings*. Harvard University. December 10, 2018. Link.
[12] D.M. Blei, A.Y. Ng, M.I. Jordan. *Latent Dirichlet Allocation*. Journal of Machine Learning Research 3 (2003) 993-1022. Link.
[13] G. Bouma. *Normalized (Pointwise) Mutual Extraction in Collocation Extraction*. Link.
[14] M. Roder, A.Both, A. Hinneburg. Exploring the Space of Topic Coherence Measures. WSDM'15, February 2–6, 2015. Link.
[15] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. St.John, N. Constant, M Guajardo-Cespedes, S. Yuan, C. Tar, Y-H. Sung, B. Strope, R. Kurzweil. *Universal Sentence Encoder*. Link.

Appendix

Appendix 1. Data Overview

Target variable overview		
Speech count	**Total**	3.7MM
Speech count with matched speaker information	**Total**	2.8MM
Share of speeches by target group	**Women**	8.1%
	People of color	8.6%
	Democrats	51.4%
Share of speeches by target group who are also Democrats	**Women**	72.6%
	Racial minorities	88.5%
Speech count per person by target group: median	**Women / Men**	118 / 134
	Racial minorities / White	98 / 139
	Dem / Rep	125 / 139
Word count per speech by target group: median	**Women / Men**	160 / 55
	Racial minorities / White	119 / 56
	Dem / Rep	65 / 53

Congressional trends from 1981 to 2016			
		97th Congress (1981-1982)	**114th Congress (2014-2016)**
Word count per speech	**Median**	36	137
	Mean	175	267
Speech count per person	**Median**	182	80
	Mean	371	125
Share of target group in congress	**Women**	4.0%	19.7%
	Racial minorities	7.0%	18.3%
Age	**Median**	51	60

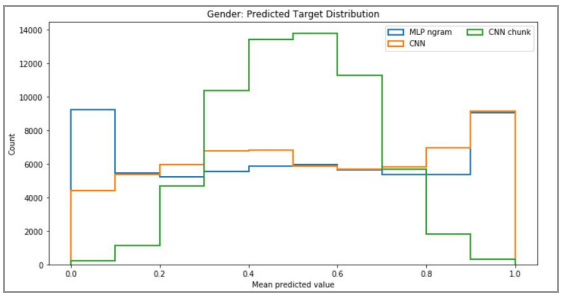
Appendix 2. Speech Examples (unedited)

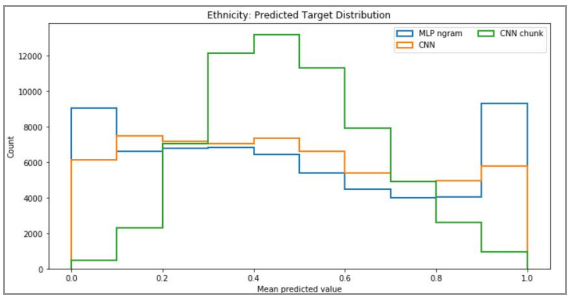
Example of a speech less than 30 words:
Speaker: HASTINGS, ALCEE (Democrat from FL, made in 2006).
Mr. Speaker. on that I demand the yeas and nays.
Example of speech in procedural format
Speaker: MITCHELL, GEORGE (Democrat from ME, made in 1994).
'I announce that the Senator from Delaware . the Senator from California . the Senator from Colorado . the Senator from Connecticut . the Senator from Kentucky . the Senator from Hawaii . the Senator from Louisiana . the Senator from Massachusetts . the Senator from Nebraska . the Senator from Massachusetts . the Senator from Connecticut . the Senator from Ohio . the Senator from Maryland . the Senator from Rhode Island . the Senator from Alabama . the Senator from Illinois . and the Senator from Minnesota are necessarily absent.'

Appendix 3. Model parameters

Gender Model Iterations		
	Pre-processing	**Final Parameters**
MLP-ngram	1. Create unigrams and bigrams using TfidfVectorizer of full speeches 2. Select top 10,000 based on f1 score	- layers: 2 - units: 64 - dropout rate: 0.2
MLP-USE	1. Speech split into 30-word chunks 2. Mapped to Universal Sentence Encodings	- layers: 2 - units: 128 - dropout rate: 0.2
CNN	1. Tokenize speeches truncated at 750 words 2. Mapped to Glove embeddings	
CNN-chunks	1. Speeches split into 50-word chunks 2. Mapped to Glove embeddings	- layers: 2 - filters: 64
CNN-USE	1. Speech split into 30-word chunks 2. Mapped to Universal Sentence Encodings 3. Re-aggregate the data to speech level	- kernel size: 5 - pool size: 1 - dropout rate: 0.2
sepCNN[9]	1. Tokenize speeches truncated at 750 words 2. Initialized embeddings randomly	- layers: 2 - filters: 64 - kernel size: 5 - pool size: 3 - dropout rate: 0.2 - block size: 2

Appendix 4. Predicted Target Distributions for Gender and Ethnicity Models





[9] Based on F.Chollet. *Xception: Deep Learning with Depthwise Separable Convolutions*. Adapted code from here.

Appendix 5. **Final Model Performance for Gender and Ethnicity**

Final Model Performance (Accuracy)				
Gender model				
	Validation - Overall	**Validation - Male**	**Validation - Female**	**Test - Overall**
MLP-ngram	75.6%	76.0%	76.0%	75.2%
CNN	73.5%	70.0%	77.0%	73.2%
CNN-chunks	74.3%	72.0%	77.0%	67.1%
Ensemble	76.7%	74.0%	79.0%	76.3%
Ethnicity model				
	Validation - Overall	**Validation - White**	**Validation - Racial Minority**	**Test - Overall**
MLP-ngram	75.5%	82.0%	69.0%	75.2%
CNN	73.6%	80.0%	68.0%	73.7%
CNN-chunks	74.5%	80.0%	69.0%	67.4%
Ensemble	76.7%	83.0%	70.0%	76.5%

****Appendix Continued Below*****

Appendix 6. Final Topics generated by LDA Topic Model, alongside top 10 words for each.

Gender Model:

	0	1	2	3	4	5	6	7	8	9
Environment	water	land	forest	environmental	act	state	national	project	legislation	protect
Energy	energy	oil	price	fuel	use	cost	production	power	increase	plant
Air Travel	border	air	aircraft	aviation	fly	airport	travel	haiti	flight	security
Random Verbs	life	know	many	live	family	man	us	great	come	love
Great Nation	history	nation	american	americans	day	today	black	america	great	first
Scientific Research	research	technology	new	science	space	develop	national	advance	center	engineer
State and Local	state	district	local	city	build	county	park	center	area	residents
Foreign Policy	world	countries	nuclear	international	china	agreement	nations	treaty	economic	policy
Healthcare	health_care	care	health	medicare	medical	cost	seniors	plan	coverage	insurance
Judicial System	court	law	judge	case	constitution	justice	right	federal	legal	constitutional
Common Congressional Phrases	chairman	gentleman	amendment	committee	thank	want	think	issue	like	distinguish
Economy	job	need	help	economy	lose	create	economic	america	workers	million
Programs and Budget	program	fund	million	provide	budget	billion	need	appropriations	house	fiscal_year
Veterans	veterans	service	federal	legislation	benefit	transportation	system	employees	provide	act
Elections and Parties	vote	congress	house	pass	republican	us	american_people	rule	debate	members
Education	school	education	students	program	college	teachers	educational	public	student	train
Freedom	government	freedom	human_right	democracy	political	free	religious	soviet	right	cuba
Children's and Family Programs	children	program	families	provide	help	child	need	service	parent	live
The Navy	foreign	aid	ship	port	coast_guard	flag	nicaragua	navy	central_america	american
Positive Words	community	serve	service	honor	award	dr	university	state	member	recognize
Military Service	military	service	serve	army	veterans	war	defense	honor	country	soldier
Foreign Trade	trade	farm	farmers	market	price	agriculture	export	products	industry	american
Random Verbs 2	get	think	want	come	know	talk	us	way	see	try
Common Congressional Phrases 2	amendment	vote	amendments	senators	ask_unanimous_consent	committee	order	rule	resolution	debate
Common Congressional Phrases 3	amendment	provision	act	legislation	law	require	congress	section	state	change
Bankruptcy and Liability	claim	bankruptcy	relief	property	case	fee	liability	damage	action	file
Drug and Violent Crime	drug	crime	safety	gun	police	fire	law_enforcement	victims	violence	kill
War and Defense	war	iraq	troop	peace	force	military	must	attack	israel	resolution
Art and Culture	american	mexico	music	community	broadcast	arts	art	culture	library	museum
Finance	company	bank	business	loan	market	financial	industry	credit	small_business	businesses
Intelligence and Security	information	intelligence	act	report	commission	use	require	government	immigration	protect
Taxes and Budget	tax	budget	billion	percent	spend	pay	cut	increase	year	debt
Women's and Labor Rights	women	workers	right	labor	act	abortion	employers	discrimination	pay	employees
Common Congressional Phrases 4	report	letter	administration	record	hear	article	question	point	secretary	write
Common Congressional Phrases 5	house	committee	congress	resolution	congressman	representative	member	members	house_representatives	staff
Healthcare Studies	percent	health	disease	study	treatment	increase	national	cause	number	population

Ethnicity Model:

	0	1	2	3	4	5	6	7	8	9
Crime and Immigration	drug	crime	border	immigration	gun	violence	act	mexico	law_enforcement	victims
Military	veterans	service	serve	va	military	honor	men_women	country	families	sacrifice
Environment	water	land	environmental	forest	project	national	public	area	environment	park
Elections and Parties	vote	congress	political	republican	public	party	democratic	house	campaign	election
Random Verbs	get	money	want	know	dont	tell	back	come	let	pay
Judicial System	law	right	court	case	constitution	constitutional	judge	legal	federal	supreme_court
The Navy	fire	ship	port	navy	haiti	flag	coast_guard	officer	guam	beach
War and Defense	military	war	iraq	troop	defense	force	army	soldier	afghanistan	general
Judicial Nominations	vote	judge	nomination	senators	district	record	present	ms	confirm	district_columbia
Common Phrases	house	committee	resolution	rule	ask_unanimous_consent	follow	order	may	hear	members
Foreign Policy	government	peace	freedom	israel	must	soviet	democracy	jewish	continue	world
Healthcare and Women's Health	women	medical	care	health	doctor	patients	hospital	abortion	service	access
National Security	report	information	department	intelligence	government	commission	office	committee	administration	security
Positive Words	community	service	national	recognize	award	association	center	city	honor	year
Agriculture	farm	farmers	agriculture	food	agricultural	program	price	rural	arkansas	market
Economy	job	workers	economy	labor	create	economic	american	employees	americans	help
Legislation	legislation	act	provision	require	congress	federal	provide	process	section	change
Poverty and Welfare	children	program	families	welfare	child	poor	family	poverty	live	care
Common Congressional Phrases	amendment	chairman	gentleman	committee	offer	vote	want	amendments	thank	distinguish
Natural Disasters	emergency	home	help	texas	state	florida	flood	disaster	damage	communities
Federal Assistance Programs	program	fund	provide	million	house	service	include	assistance	federal	grant
Education	school	education	students	college	children	program	educational	teachers	young	learn
Foreign Trade	trade	company	industry	market	american	export	foreign	products	countries	import
Children's Diseases	children	health	child	disease	treatment	live	parent	national	research	help
Positive Words 2	serve	service	family	honor	dr	member	life	career	university	john
Infrastructure	transportation	safety	project	system	highway	construction	air	build	airport	truck
Random Verbs 2	think	want	come	know	get	issue	talk	way	see	country
Diplomacy and Security	world	nations	international	policy	countries	nuclear	must	administration	agreement	security
Finance and Business	bank	loan	financial	credit	small_business	market	business	company	small_businesses	capital
Energy	energy	oil	price	fuel	production	power	natural_gas	use	supply	gas
Scientific Research	research	technology	new	science	space	program	test	system	national	use
States	state	washington	colorado	california	alaska	indian	minnesota	nevada	governor	native
Budgets and Spending	budget	spend	billion	cut	deficit	congress	debt	year	tax	vote
Costs	percent	cost	increase	million	year	number	billion	rate	less	level
Taxes and Welfare	tax	pay	medicare	benefit	plan	social_security	income	cost	insurance	seniors
Great Nation	history	great	nation	day	american	live	first	world	americans	life

Appendix 7. Most commonly used topics in misclassified speeches