Meredith Brown, Caleb Kornfein
May 1, 2020

The Journey to the Presidency - Topic Modeling of the 2020 Democratic Debates

I. Introduction

The process of becoming the President of the United States has progressively started earlier and earlier. For the 2020 election season, candidates to be the Democratic nominee for president began declaring their candidacy in early 2019. While the field of candidates grew to over 20 people[1], one of the most prominent ways for candidates to get their message to Americans is through live televised debates. These debates, beginning in June of 2019, take place across the country, are hosted and moderated by different television networks, and have cutoffs for what qualifies a candidate to participate.[2]

In each debate, candidates are asked questions about specific topics, such as healthcare, foreign policy, gun rights, and the economy. However, the candidates frequently drift off topic, choosing to talk about issues that will resonate with Americans and show them in a good light. It is the purpose of this project to explore the topics discussed in these debates and by which candidates. In this way, this project aims to tackle the key question that voters are tasked with addressing: how important to this candidate are the issues I care about?

To explore which candidates discuss which issues, this project will examine the transcripts from the Democratic debates using topic modeling. To implement topic modeling, this project will use a latent Dirichlet allocation model to infer the underlying topics of the debate. First, a corpus, or a collection of text documents, is defined. For this project, each candidate is assigned to a document, and each document includes all of their responses to any question during any of the Democratic debates. From this corpus, the model then attempts to identify the underlying, or latent, topics in the debates. The model assumes that each document comes from a generative probabilistic process. The process of model fitting implements a belief that a certain number of these latent topics exist, and each document is a mixture of topics. Likewise, there is a belief that a certain number of terms exist, and each topic is a distribution over these terms. With a distribution of terms per topic and a distribution of topics per document, this project is then able to analyze which topics occur most often for each candidate. In other words, the created latent Dirichlet allocation model will be used to assess which candidates discussed which topics the most, hopefully providing voters with a resource to evaluate the candidates on the issues they most care about.

---

[1] https://www.nytimes.com/interactive/2019/us/politics/2020-presidential-candidates.html
[2] https://www.uspresidentialelectionnews.com/2020-debate-schedule/2020-democratic-debate-schedule/

II. Data

The data to be used for this project comes from the Kaggle dataset "Democratic Debate Transcripts 2020".[3] The data in the dataset is provided publicly on Rev.com, which produces transcripts of audio recordings. The initial dataset contained six columns as variables for each observation: date, which debate the speech occurred at, section of the debate, speaker, text of the speech, and speech time. First, the dataset was filtered to only include the categorical variable speaker and text variable speech, as they are the only ones needed for the creation of the model. As the questions from the moderators and audience were also included, it was necessary to transform the data such that only responses from the candidates were left. Most importantly, the responses from this filtered data were collapsed by speaker, aggregating the responses made by each speaker throughout all of the debates.

With the transcript data filtered and organized by candidate, the data was primed to be assessed as a corpus. By definition, a corpus is a collection of documents, and each document is a text document filled with terms. The aggregation of each speaker's responses was translated into a representative document, which was then added to this project's corpus. This process resulted in 20 documents, one for each candidate that participated in the Democratic debates. In reviewing these documents, it was apparent that some modifications were necessary to make the analysis impactful. Many of the terms are uninformative, include punctuation, or vary only by the case of a letter. Therefore, some obligatory text preprocessing was performed, including: stripping whitespace and punctuation; converting letters to lowercase; removing common english stop words, such as "the" and "a"; removing numbers; and removing other uninformative terms that came up, such as "can" and "people".

The final corpus encompassed 20 documents and 8070 total terms. For purposes of exploring the data, a wordcloud (*Figure 1*) of the terms appearing more than 175 times was created. The more frequently it occurred, the larger the term in the word cloud.

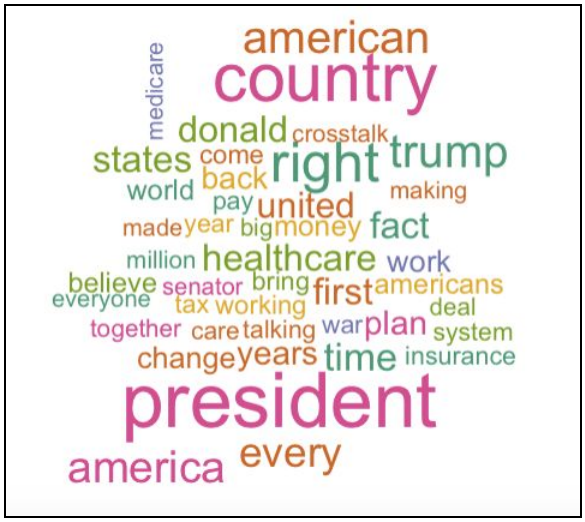

Figure 1: *Wordcloud of terms occurring more than 175 times*

[3] https://www.kaggle.com/brandenciranni/democratic-debate-transcripts-2020

For the purposes of model implementation and posterior analysis, the R package *topicmodels*[4] was used to run latent Dirichlet allocation. The *lda* function provided by this package takes as input an object in the form of a document-term-matrix. A document-term-matrix is organized with the documents as rows and the terms as columns. The value at the intersection of a given i^{th} row and j^{th} column represents how many times speaker *i* used term *j* in his or her responses at the Democratic debates. Thus, the final data preparation step was converting the final corpus from text preprocessing to a document-term-matrix using the *dtm* function provided by the aforementioned R package.

III. Model

For this project, a latent Dirichlet allocation (referred to hereafter as LDA) model was chosen for analysis because it is a useful model for identifying and understanding the underlying topics in a collection of texts. As the goal of this project is to identify the key topics in the candidates' responses, this model is useful in providing insight into which topics each candidate speaks about the most. In LDA, a bag-of-words policy, in which not the ordering of the terms in each document, but rather the frequency of the terms is considered important, is applied. The input data is the corpus of text-processed documents in the form of a document-term-matrix.

LDA works well because of the conjugacy between Dirichlet and Multinomial distributions. The Dirichlet distribution is a multivariate generalization of the Beta distribution that takes in a hyperparameter *alpha*, representing a vector of real probabilities. The Multinomial distribution is a multivariate generalization of the Binomial distribution, taking in a number of trials *n* and a vector of real probabilities *p* for a set number of categories *K*.

Each topic can be visualized as a distribution over all of the sample terms, where each term is assigned a probability and the sum of the probabilities of the terms is one. The process of generating a new document involves first sampling *K* topics, or distributions, over the sample terms, using a Dirichlet distribution with hyperparameter *beta*. Sampling *K* topics forms a matrix with *K* rows and *N* columns, where *N* is the number of terms in the sample space. A_{ij} represents the probability of drawing word *j* under topic *i*.

Each document can be visualized as a mixture of different topics. Thus, for each document, a sample is taken from the document's topic proportions, regulated by another Dirichlet distribution with hyperparameter *alpha*. Once the topic proportions are known, a sample can be drawn to identify which word index in the document belongs to which topic using a Multinomial distribution. Given a word index and a topic, it is easy to sample from the topic distribution, simply a Multinomial distribution with a singular trial using the associated probability vector sampled above. Thus, a process has been developed for generating documents,

[4] https://cran.r-project.org/web/packages/topicmodels/index.html

pictured below in *Figure 2*[5]. In *Figure 2*, the nodes are the random variables, arrows show dependence, and the plates show when something is replicated more than one time.

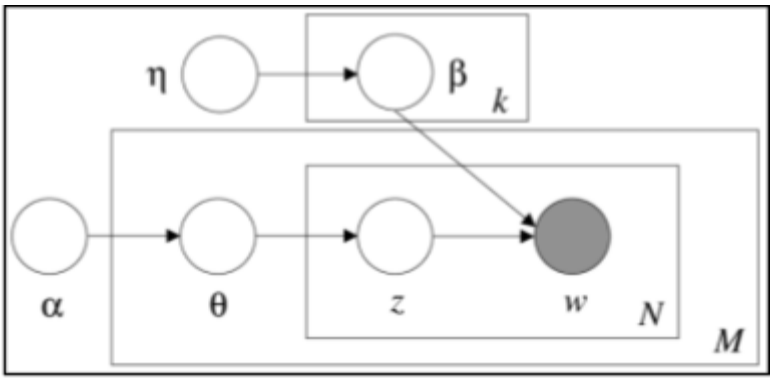

Figure 2: *Latent Dirichlet Allocation process*

k ~ Number of Topics
V ~ Vocabulary Size
M ~ Number of Documents
N ~ Number of Words in particular Document
w ~ Word in Document from 1,..., N
z ~ Topic from topics 1, …, k
Θ ~ Vector of Topic proportions
∝ ~ Topic Proportions parameter
β ~ Topic Concentration Parameter

This project is interested in the posterior: $p(topics, proportions, assignments \mid documents)$, where the joint distribution $p(topics, proportions, assignments, documents)$ is characterized by:

$$p(\beta,\theta,\mathbf{z},\mathbf{w}) = \left(\prod_{i=1}^{K} p(\beta_i \mid \eta)\right)\left(\prod_{d=1}^{D} p(\theta_d \mid \alpha)\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d)\, p(w_{d,n} \mid \beta_{1:K}, z_{d,n})\right)$$

[6]

For initial model development, weak prior values of 0.01 were chosen for *alpha* and *beta*, corresponding with small prior sample sizes. Small concentration parameters also push the density of the Dirichlet simplex to the corners. This is consistent with other LDA literature that sets small values for the priors. Likewise, the use of these prior values has the effect of making the Dirichlet prior distributions symmetric. To find the optimal number of topics, the R package *ldatuning*[7] was used, as it supplies four metrics for scoring the LDA model with the given number of topics.

---

[5] https://www.researchgate.net/figure/Graphical-model-representation-of-the-smoothed-LDA-model_fig3_221620547

[6] https://user.eng.umd.edu/~smiran/LDA.pdf

[7] https://cran.r-project.org/web/packages/ldatuning/index.html

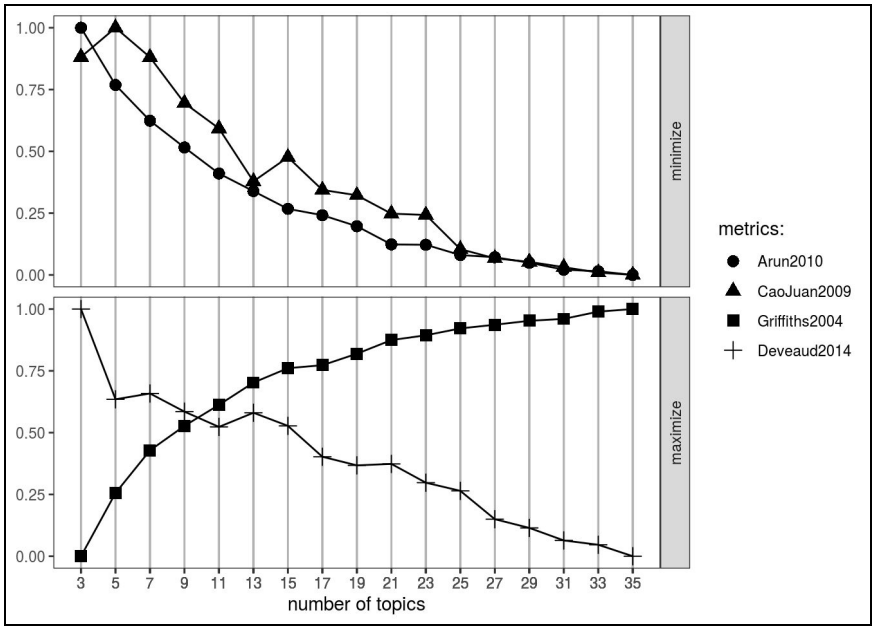

Figure 3: *Results from scoring metrics*

These four scoring algorithms represent four different approaches for evaluating the optimum number of topics in an LDA model. *Arun2010* proposes an empirical measure which assumes LDA operates through a matrix factorization mechanism. The corpus is split into two separate matrix factors, and this metric analyzes how good the split was by computing a form of symmetric KL-Divergence of distributions derived from the matrix factors, which hinges on the number of topics chosen. *CaoJuan2009* proposes a method of choosing the ideal topic number that adaptively selects the best LDA model based on density. The *Griffiths2004* metric uses a form of Bayesian model selection to determine the number of topics. *Deveaud2014* proposes an unsupervised approach to learning the optimal number of topics through Latent Concept Modeling. In reviewing *Figure 3*, the measures graphed in the top half are those which aim to be minimized, while those in the bottom are to be maximized by the choice of number of topics. Through this analysis, this project chose to set the number of topics at 13, as this value appeared to best minimize the *Arun2010* and *CaoJuan2009* metrics while maximizing the *Griffiths2004* and *Deveaud2014* metrics.

After an initial burnin period of 200 samples, the model converged to a desired density region of posterior probabilities. The results of the model help to address the key question of which candidates spoke about which topics most frequently. The posterior distribution of terms per topic identifies soft clusters of terms that provide information about a similar category. Further, the posterior distribution of topics per documents identifies which topics are most likely to occur in a document and at what frequency. Visualizations and analysis of this posterior, specifically, will be foremost in drawing conclusions about the nature of this corpus.

IV. Results

In the previous section, it was found that the optimal number of topics given this corpus is 13. As a reminder, each topic is a probability distribution over the term sample space, which contains 8070 terms. Displayed below in *Figure 4* are the top 20 words associated with each of the topics found by the model.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13
money	right	bill	fact	industry	right	country	fight	justice	back	president	country	states
jobs	washington	someone	deal	major	america	war	work	women	time	first	american	united
kids	american	work	idea	today	every	lives	big	work	years	donald	healthcare	common
right	climate	bills	here's	war	united	bring	money	children	stage	trump	trump	nation
communities	experience	everyone	he's	year	come	everyone	families	issue	didn't	change	system	fight
china	generation	agreement	middle	joe	time	americans	military	states	better	single	companies	policies
problems	order	back	cost	wealth	tax	national	everyone	plan	school	win	years	rights
immigrants	sense	election	ever	maybe	care	face	america	department	everybody	talking	working	criminal
americans	community	better	plan	fossil	together	texas	means	families	problem	plan	party	issue
economy	kind	bring	move	world	crosstalk	laws	government	folks	mayor	state	end	stage
who's	presidency	comes	states	fuel	country	meet	democrats	rights	city	house	americans	literally
build	politics	lead	entire	million	made	foreign	debt	united	senator	making	dollars	cause
ohio	somebody	rural	position	change	pay	member	build	black	stop	senate	policy	deal
start	future	sanders	guy	half	states	everyone	corporations	ability	job	thank	saying	children
forward	deliver	bernie	making	street	public	mental	works	elected	world	place	tonight	poverty
michigan	happen	voting	whether	billionaires	medicare	reserve	every	fight	live	point	issue	start
freedom	ready	experience	period	legislation	million	democracy	plan	vice	across	support	create	values
wall	black	means	end	earth	year	members	teacher	nation	ago	white	believe	communities
countries	majority	step	happened	billion	three	every	corruption	access	small	big	economy	issues
hands	whether	truth	immediately	human	trade	already	great	believe	taxes	home	income	mine

Figure 4: *Top 20 terms per topic*

Here are this paper's interpretations of each topic, derived from the top words chart:

Topic 1: Job outsourcing and immigration
Topic 2: Generic ability to deliver as President
Topic 3: Passing bills, legislature, experience
Topic 4: Efficacy, ability, plans, preparedness
Topic 5: Loosely climate/sustainability, wealth inequality, emerging industries
Topic 6: Healthcare and taxes
Topic 7: Foreign policy and war
Topic 8: Anti-corruption, big business
Topic 9: Justice, civil rights issues
Topic 10: The past, reflecting backwards
Topic 11: Generic Presidential aspirations, ability to win
Topic 12: Healthcare and working Americans
Topic 13: Crime and poverty, community issues

Viewing these key terms for each topic is helpful, but does not give a full picture for each topic. To understand more as to what these topics constitute, the R package *LDAvis*[8] was used to visualize the distribution of the terms over each topic, the frequency of the terms per topic, and the overlap between the topics.

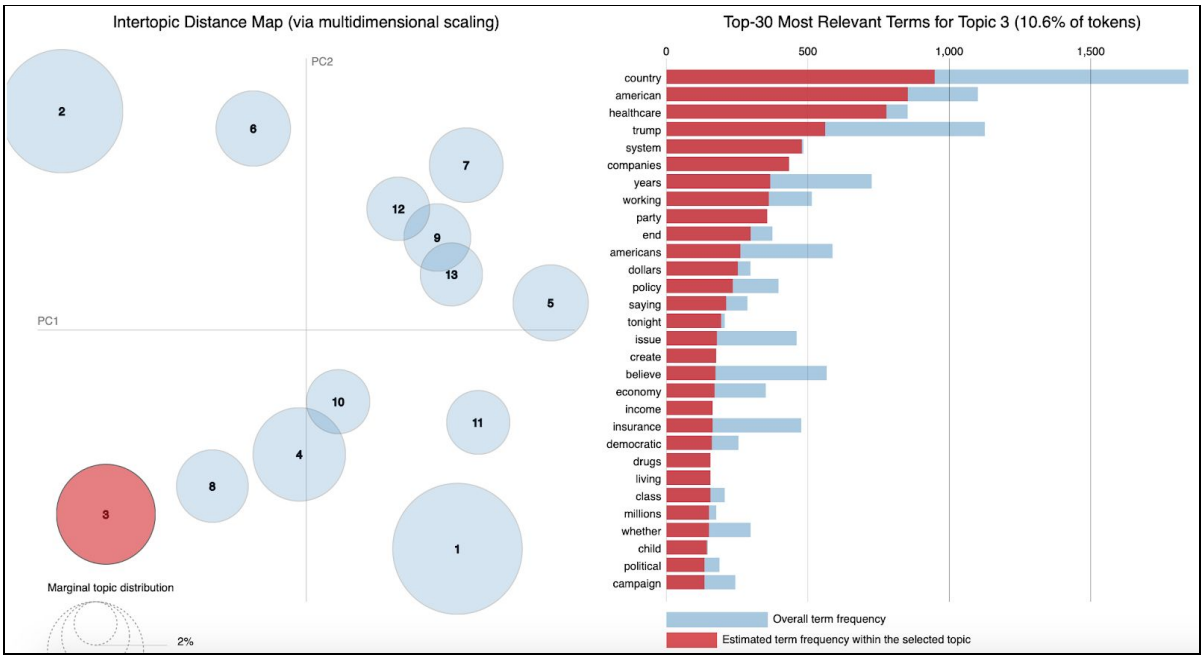

Figure 5: *Screengrab of LDAvis visualization for Topic 3*

Figure 5 showcases the visualization produced with *LDAvis*, running as a web application through this project's github repository at https://bl.ocks.org/meredithb3/30583d25bb2a606bba9c0f374f401c57. The left-hand side of the visualization plots the topics using multidimensional scaling on inter-topic distances.[9] The distance between the topic circles can be interpreted as the distance between the topics. Therefore, in analyzing the visualization, it can be concluded that there is relative separation between the posterior topic assignments. This speaks to the effectiveness of the model in assigning terms to topics, as little redundancy in the topics conveys the idea that the topics represent comparatively unique concepts. The *LDAvis* visualization is a good tool for viewing which topics were the most frequent, their distance from one another, and the most frequent words associated with them; however, in the visualization, the numbers corresponding to the bubbles represent the topics by frequency, rather than by topic number. For example, the bubble labeled 2 corresponds to the second most frequent topic, in this case, topic 11. The order of the topics is as follows: *6, 11, 12, 10, 9, 4, 8, 7, 1, 2, 13, 5, 3*.

[8] https://cran.r-project.org/web/packages/LDAvis/index.html
[9] https://slides.cpsievert.me/ldavis/#9

On the right-hand side of the visualization, more granular information about the distribution of the terms per topic is provided. The red bars indicate term frequency within the selected topic, here the third most frequent topic, topic 12. The blue bars reference the overall frequency of the term across the corpus. As a result, here we can see that the term "country", which appears approximately 2,000 times in the corpus, has about half of its occurrences associated with topic 12. For example, in highlighting this term on the right-hand side, the topic circles on the left are resized to show the topics which encompass this term most.

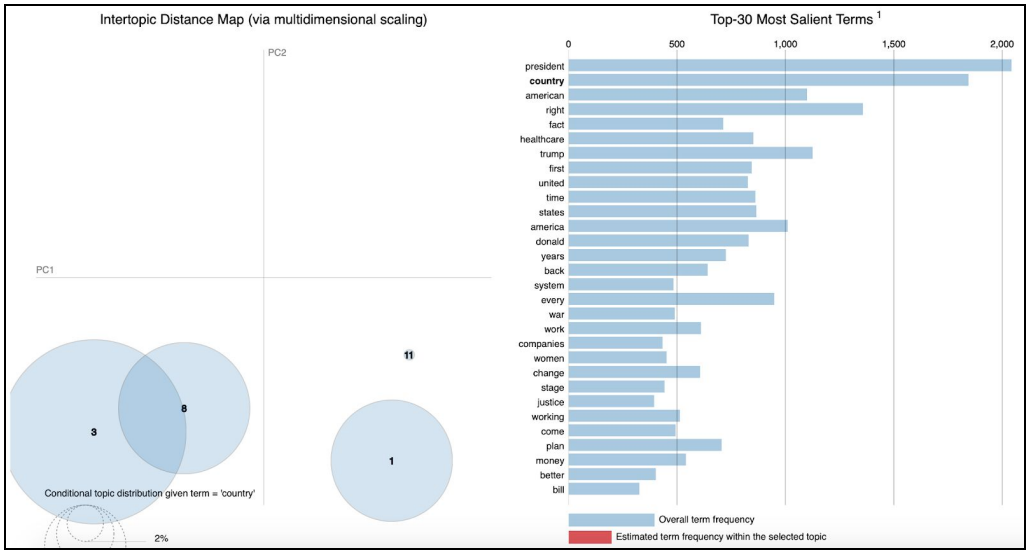

Figure 6: *Screengrab of LDAvis visualization for term "country"*

However, as *Figure 6* further shows, terms like "drugs" and "living" are almost entirely associated with topic 12, so they may be more fully encapsulated within the topic. To address this issue of overcompensating for well-dispersed terms, Term Frequency-Inverse Document Frequency (TF-IDF) was applied to the posterior term distribution. This process "[adjusts] the frequency of a term...for how rarely it is used," thus adjusting the topic proportions for the corpus. The "re-ranked" topics appear in *Figure 7* below, with the first 4 terms for each topic appearing as its pseudo-name.

Topic 1	money-jobs-kids-right
Topic 2	right-washington-american-climate
Topic 3	bill-someone-work-bills
Topic 4	fact-deal-idea-heres
Topic 5	industry-major-today-war
Topic 6	right-america-every-united
Topic 7	country-war-lives-bring
Topic 8	fight-work-big-money
Topic 9	justice-women-work-children
Topic 10	back-time-years-stage
Topic 11	president-first-donald-trump
Topic 12	country-american-healthcare-trump
Topic 13	states-united-common-nation

right america every come	0.18814034
president first donald trump	0.16902953
country healthcare american trump	0.12081550
back time stage years	0.09161095
justice women work children	0.05924945
kids money jobs right	0.05770772
fight big families military	0.05463195
country war lives bring	0.05459029
fact deal heres idea	0.05206291
washington right generation experience	0.04167829
states united common nation	0.04062802
bill someone bills work	0.03616174
industry major year war	0.03369330

Figure 7: *Concatenated titles, Term-Frequency-Inverse Document Frequency probabilities*

In comparing the re-ranked topics from *Figure 7* with the proportional circles in *Figure 5*, it is interesting to focus in particular on topic 4. From examining *Figure 5*, topic 4 is represented as the 6^{th} most frequent topic. However, upon reranking for more rare and substantial terms, topic 4 drops to represent the 9^{th} largest percentage of the text in the corpus. In this way, it is shown that the conclusions drawn from the *LDAvis* visualizations must be considered granularly and with caution.

After considering the relevant features and distribution of terms over the topics, this project turned to analyzing the distribution of topics over the documents in the corpus. As previously mentioned, each document in the corpus represents the responses from a given participant in the 2019-2020 Democratic debates. Each document consists of some combination of the 13 topics defined by the distribution of the terms. The posterior analysis produced a matrix wherein the rows represent the rank *i*, and the columns represent the *j* candidates. The colors represent topics, meaning that the color value at the intersection of row *i* and column *j* shows that that corresponding topic was the i^{th} most talked about issue for candidate *j*. To understand more as to what these values constitute, a chart (*Figure 8)* was created which can be read top-down. For example, in the top row, the j^{th} value means that for candidate *j*, this was the topic he or she discussed the most. The topics generally near the top of the chart were discussed more frequently, while those closer to the bottom were discussed less.

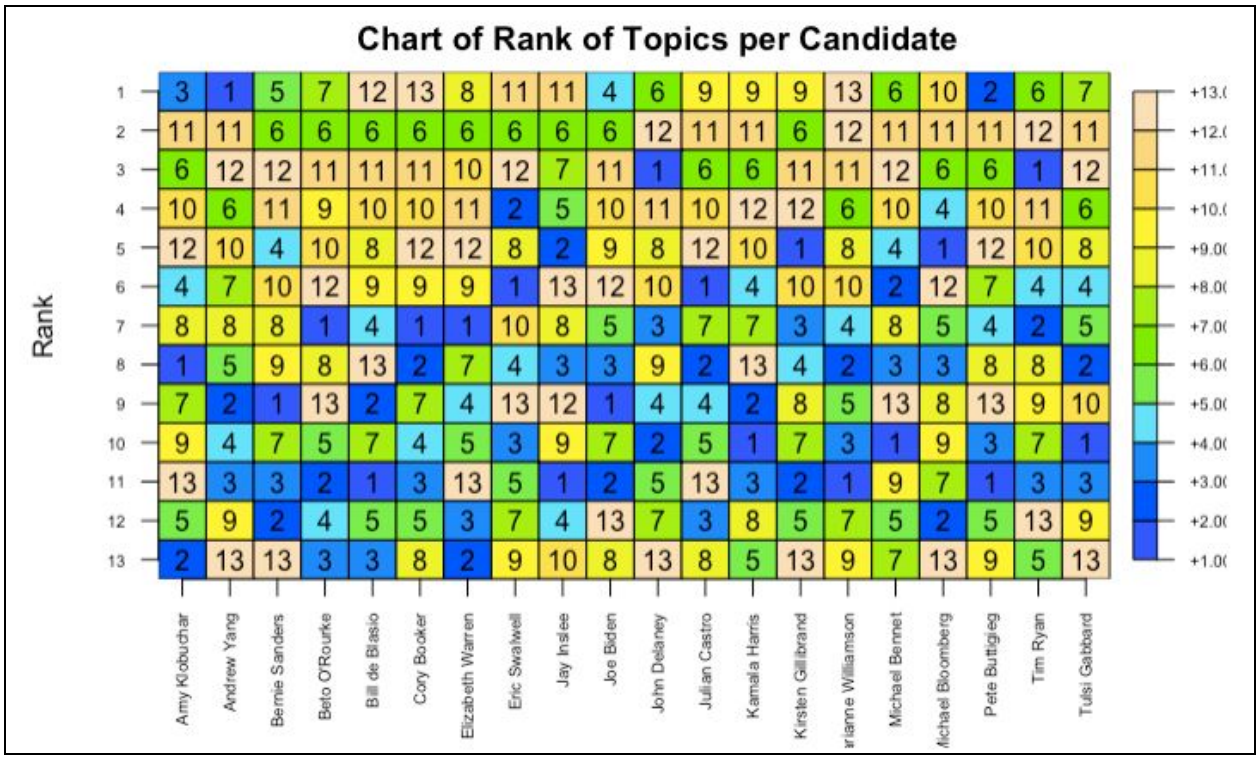

Figure 8: *Chart of the ranking of topics per candidate*

From *Figure 8*, we can more easily and accurately compare the distribution of topics across documents. As the visualization shows, topics 6 and 11 appear to maintain importance across candidates, while topics 1,3, and 5 seem to be relatively less talked about, with a few candidates having them as a high priority versus a majority of candidates regarding them with relative unimportance.

These findings can be explored further when concentrating on the posterior probabilities of the topics per candidate, rather than simply the relative importance of the topics. Using the *posterior* function from the *topicmodels* R package, the matrix values in *Figure 8* are transformed into the desired posterior topic probabilities, shown for the topics below in *Figure 9*:

	1	2	3	4	5	6	7
Amy Klobuchar	2.402308e-02	1.415901e-02	2.592522e-01	3.284272e-02	1.659601e-02	0.19763077	2.402308e-02
Andrew Yang	3.410190e-01	2.556631e-02	2.035427e-02	2.258800e-02	3.201932e-02	0.10895901	4.393256e-02
Bernie Sanders	1.385191e-02	7.417280e-03	1.254317e-02	3.653673e-02	3.021017e-01	0.23775538	1.297942e-02
Beto O'Rourke	3.152649e-02	2.607051e-02	1.970520e-02	2.182697e-02	2.849539e-02	0.23036679	2.809862e-01
Bill de Blasio	1.798601e-02	2.483456e-02	8.560691e-06	3.168312e-02	6.001044e-03	0.21316977	2.141029e-02
Cory Booker	3.759683e-02	3.084331e-02	2.386468e-02	2.746655e-02	1.621069e-02	0.16096107	2.904237e-02
Elizabeth Warren	2.139588e-02	9.952105e-03	1.234037e-02	1.910713e-02	1.373353e-02	0.21852738	2.089833e-02
Eric Swalwell	7.143070e-02	1.309315e-01	2.975039e-05	4.168030e-02	2.975039e-05	0.19043227	2.975039e-05
Jay Inslee	1.436513e-05	6.609398e-02	3.592720e-02	1.436513e-05	7.758608e-02	0.25140419	1.264275e-01
Joe Biden	1.915218e-02	1.404519e-02	2.209852e-02	3.043577e-01	2.877689e-02	0.21567295	1.551836e-02
John Delaney	1.362402e-01	1.114700e-02	5.227353e-02	1.543101e-02	8.568026e-06	0.34787042	8.568026e-06
Julian Castro	4.815902e-02	3.894679e-02	2.135981e-02	3.266573e-02	2.764087e-02	0.15284344	3.936553e-02
Kamala Harris	2.103985e-02	2.332654e-02	1.578046e-02	4.207741e-02	2.286692e-06	0.17264751	3.315932e-02
Kirsten Gillibrand	8.177118e-02	8.516944e-06	5.707204e-02	4.770341e-02	8.516944e-06	0.18567791	2.811443e-02
Marianne Williamson	1.005905e-05	3.018720e-02	2.817539e-02	3.722853e-02	3.018720e-02	0.11568909	1.005905e-05
Michael Bennet	2.334535e-02	6.184593e-02	3.617888e-02	9.334640e-02	1.166684e-05	0.23684855	1.166684e-05
Michael Bloomberg	5.360630e-02	1.540850e-02	3.081084e-02	7.208911e-02	3.204303e-02	0.14971690	2.218553e-02
Pete Buttigieg	2.529476e-02	2.291017e-01	2.789699e-02	4.215723e-02	2.394159e-02	0.12105696	6.099741e-02
Tim Ryan	1.578425e-01	3.802071e-02	2.231992e-02	7.438044e-02	8.263575e-06	0.19668135	3.223621e-02
Tulsi Gabbard	2.884866e-02	3.565920e-02	2.524308e-02	4.607532e-02	3.846354e-02	0.05889517	3.004691e-01

	8	9	10	11	12	13
Amy Klobuchar	3.145015e-02	1.717625e-02	7.995818e-02	0.24196107	0.04421542	1.671206e-02
Andrew Yang	3.872052e-02	1.712777e-02	6.552531e-02	0.14866981	0.12186502	1.365307e-02
Bernie Sanders	2.650306e-02	1.505159e-02	3.064740e-02	0.08485102	0.21234403	7.417280e-03
Beto O'Rourke	2.940472e-02	6.456551e-02	6.365618e-02	0.11791290	0.05668464	2.879850e-02
Bill de Blasio	5.993340e-02	4.794843e-02	1.326993e-01	0.14554031	0.27223854	2.654670e-02
Cory Booker	1.553534e-02	4.907781e-02	9.657754e-02	0.15870990	0.09387614	2.602378e-01
Elizabeth Warren	3.212228e-01	2.338610e-02	1.204094e-01	0.10677641	0.09931307	1.293744e-02
Eric Swalwell	1.160563e-01	2.975039e-05	6.250558e-02	0.24398298	0.13985660	3.004790e-03
Jay Inslee	4.885582e-02	2.443509e-02	1.436513e-05	0.28731702	0.03161766	5.029233e-02
Joe Biden	1.257203e-02	5.264223e-02	1.204080e-01	0.14329124	0.03820517	1.325950e-02
John Delaney	5.827114e-02	2.142863e-02	5.484393e-02	0.11053610	0.19193235	8.568026e-06
Julian Castro	4.187377e-06	2.420346e-01	1.185069e-01	0.18550498	0.06783969	2.512845e-02
Kamala Harris	1.555179e-02	2.906408e-01	4.367810e-02	0.20557587	0.10587611	3.064396e-02
Kirsten Gillibrand	4.685171e-02	1.958982e-01	7.155085e-02	0.14479657	0.14053810	8.516944e-06
Marianne Williamson	7.947653e-02	1.005905e-05	4.125215e-02	0.16598433	0.23539175	2.363977e-01
Michael Bennet	4.084561e-02	2.101198e-02	1.061799e-01	0.20884813	0.13768040	3.384551e-02
Michael Bloomberg	2.464990e-02	2.403381e-02	3.813681e-01	0.15587784	0.03820396	6.160936e-06
Pete Buttigieg	3.435053e-02	2.258843e-02	1.116889e-01	0.18725780	0.08150301	3.216465e-02
Tim Ryan	3.388892e-02	3.306256e-02	9.669209e-02	0.12726732	0.16610612	2.149356e-02
Tulsi Gabbard	5.849455e-02	2.283936e-02	3.405672e-02	0.20992897	0.14102230	4.006202e-06

Figure 9: *Posterior topic probabilities by candidate*

To interpret these values, a proportional bar chart was created, wherein each document represents a bar and the probabilities of the topics within each document total to 1.

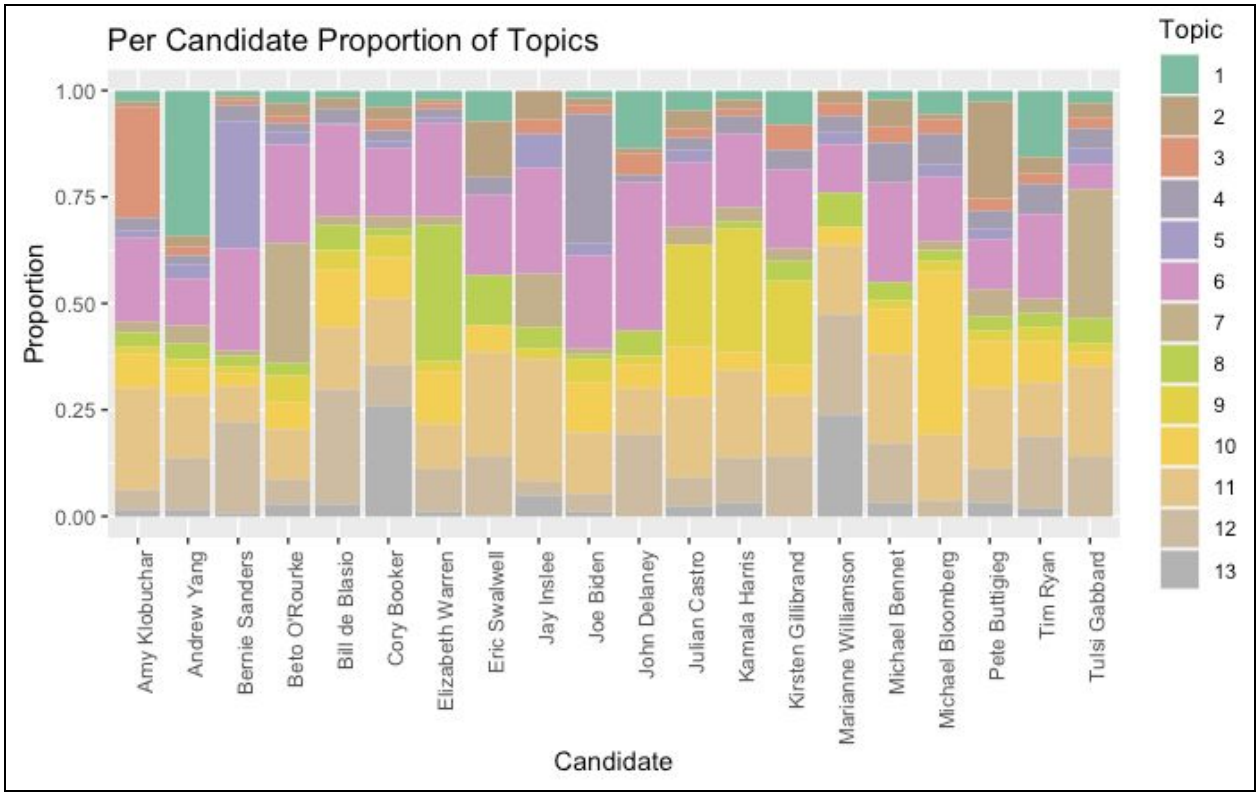

Figure 10: *Proportional bar graph of topic proportions per candidate*

This graph is the culmination of the results of this paper's model. In using the topic interpretations on page 7, this graph can be used to draw a direct link from the terms used in this corpus to the topics that are most discussed by each candidate. Each color in this graph represents 1 of the 13 topics, providing a quick visual aid to compare across candidates on who spent the most time discussing which topic.

Several conclusions can be drawn from this visualization. First, it can be used to draw conclusions about the time spent discussing a topic by all candidates. For example, topic 7, which this paper identified as "foreign policy and war", appears to take up little time for each candidate. On the other hand, topic 6, which this paper identifies as "healthcare and taxes", appears to take up a significant proportion of each candidates' responses. When considering this analysis in conjunction with the *LDAvis* visualizations, the idea that topics 6 and 11 constitute a large portion of the terms in the corpus and that they constitute a large proportion of the responses by the candidates join together to draw the conclusion that the terms in topics 6 and 11 are those which candidates use to appeal to voters, using these terms, aggregated into these 2 topics, to connect with voters.

Likewise, this visualization can be used to address the primary question of this paper: how important to this candidate are the issues I care about? In examining each bar individually, the proportions for each topic can be interpreted as the emphasis put on that concept by that specific candidate. For example, in focusing on the bar for Joe Biden, it appears that this candidate put emphasis on topics 4, 6, and 11. In comparison, the bar representing Elizabeth Warren appears to show a greater emphasis on topic 8. These interpretations of this visualization prove useful in association with the names identified for each topic on page 6, as they allow the viewer to directly correlate the responses of the candidates in the corpus to a candidate's attentiveness and focus on specific concepts. Thus, in analyzing key terms in each topic and topic proportions across documents, the LDA model was able to identify topics that are important to each candidate, providing voters with a final summative visualization by which to assess a candidate's focus on specific topics.

V. Conclusion

There are a few insightful conclusions that can be drawn from our model. First and foremost, the topics that constituted the largest portion of the corpus were health care and taxes, generic ability to deliver as President, and preparedness/reflecting on experience. As one might expect, the candidates discuss their individual qualifications and experience quite frequently.

A second conclusion is the infrequency with which candidates discuss foreign policy or broader international issues. Only one topic, topic 7, substantially addressed the international community and addressed US foreign policy and wars more broadly. The three candidates who spent the most time on this issue were Tulsi Gabbard, Jay Inslee, and Beto O'Rourke, with Beto spending around 28% of his time on the topic. Outside of these three though, the major candidates - Elizabeth Warren, Joe Biden, Bernie Sander - all hovered around 1-2%. This speaks to a broader issue of American-centric discourse that occurs at Presidential debates.

When looking at the major candidates, Bernie Sanders was an outlier, spending 75% of his time on topics 5, 6, and 12, which spanned topics such as climate/sustainability, wealth inequality, emerging industries, healthcare, taxes, and working Americans. In contrast, Joe Biden spent nearly 70% of his time on topics 4, 6, and 11, discussing efficacy, ability, plans, taxes, healthcare, presidential aspirations, and ability to win. Meanwhile, Elizabeth Warren spent about 32% of her time exclusively on topic 8, which was anti-corruption and big businesses. This could be an insight for an issue-oriented voter, noting that Bernie Sanders or Elizabeth Warren spent substantially more time covering issues.

The power of this model comes from its ability to be realistically applied. The model matters because it can act as a resource for individuals to evaluate the candidates on. For voters who might be low on time, unable to view the debates, or simply looking for a clear and efficient way to evaluate information about the debates, the model and the resulting visualizations could act as a resource for quick information on each of the candidates.

In evaluating the LDA model, one of the key strengths is the interpretability of the posterior. The posterior provides a myriad of different topics that, while unintelligible to a computer, carry meaningful human significance. For example, from a quick glance at the terms that construct topic 7, it is easy to see that it describes foreign policy and, more specifically, foreign aggression. Using this information, the human behind the model can then make actionable insights. However, the topics the model spits out are both a strength and a weakness, since it is hard to know when or if the model is working well, and many of the topics are not as clear. The difficulty is compounded by the fact that there are no cut and clear metrics for determining how good of a job the model is doing, only in evaluating an optimal number of topics prior to running the model. Additionally, the results are predicated off the assumption that the initial words were exchangeable through a bag-of-words policy, which is not always an accurate assumption. In real life, sentence structure and the placement of terms relative to each other matters. Moreover, because of the assumption of independently distributed exchangeability, the model does not capture correlations between certain words, such as "New" and "York", that would naturally go together. Overall, the produced LDA model is useful for learning real world topics and being able to visualize them so as to understand features of the larger corpus more quickly.

That being said, there is much more to look into. For example, it might be of use to analyze the Democratic debates sequentially so as to see how the proportions of topics discussed varies as the different campaigns progressed. Additionally, future work on this project could include the implementation of non-negative matrix factorization and anchor words to better assess the goodness of fit of the topics. The anchor word assumption assumes "for each topic there exists a word (the anchor word) that occurs with non-zero probability only if the document is about that specific topic."[10] In applying these techniques, future developments could expand on the LDA basis of this model, providing additional insights into the latent topic distribution across this corpus.

Link to Github for this project: https://github.com/meredithb3/Stat360-Topic-Modeling

[10] https://cs.stanford.edu/~rishig/courses/ref/l9b.pdf

Works Cited

Blei, David & Ng, Andrew & Jordan, Michael. (2001). Latent Dirichlet Allocation. The Journal of Machine Learning Research. 3. 601-608. https://www.researchgate.net/figure/Graphical-model-representation-of-the-smoothed-LDA-model_fig3_221620547

Burns, Alexander, Matt Flegenheimer, Jasmine C. Lee, Lisa Lerer, and Jonathan Martin. "Who's Running for President in 2020?" The New York Times. The New York Times, January 21, 2019. https://www.nytimes.com/interactive/2019/us/politics/2020-presidential-candidates.html.

Carson Sievert and Kenny Shirley (2015). LDAvis: Interactive Visualization of Topic Models. R package version 0.3.2. https://github.com/cpsievert/LDAvis.

Ciranni, Branden. "Democratic Debate Transcripts 2020." February 2020. Distributed by Kaggle. https://www.kaggle.com/brandenciranni/democratic-debate-transcripts-2020.

"Democratic Party Debate Schedule: 2020 Primary Debates." Election Central, April 30, 2020. https://www.uspresidentialelectionnews.com/2020-debate-schedule/2020-democratic-debate-schedule/.

Grün B, Hornik K (2011). "topicmodels: An R Package for Fitting Topic Models." Journal of Statistical Software, *40*(13), 1-30. doi: 10.18637/jss.v040.i13 (URL: https://doi.org/10.18637/jss.v040.i13).

Miran, Sina. "Latent Dirichlet Allocation (LDA) for Topic Modeling." Accessed April 16, 2020. https://user.eng.umd.edu/~smiran/LDA.pdf.

Murzintcev Nikita (2020). ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters. R package version 1.0.2. https://github.com/nikita-moor/ldatuning.

Sievert, Carson, and Kenny Shirley. "LDAVis: A Method for Visualizing and Interpreting Topics." Accessed April 26, 2020. http://cpsievert.github.com/slides/LDAvis.

Silge, Julia, and David Robinson. " Analyzing Word and Document Frequency: Tf-Idf." Text Mining with R, March 7, 2020. https://www.tidytextmining.com/tfidf.html#.

Voss, Catalin. "Building Topic Models Based on Anchor Words." Stanford University, May 13, 2014. https://cs.stanford.edu/~rishig/courses/ref/l9b.pdf.