Democratic Debates Topic Modeling

Pouya Mohammadi and Hunter Gregory

April 30, 2020

1 Introduction

In recent years, we have seen the popularity of presidential primary debates increase, with the two most viewed debates both occurring during the current presidential election cycle. As more and more Americans tune in to these primary debates to gather information about who they will cast their votes for and declare their preferred party's nominee for president, it is increasingly important to understand how these candidates are using this growing platform to get out their message and convince voters of their platform and candidacy. One interesting point to consider when analyzing this problem is that politicians avoid answering the questions asked of them in interviews. Instead, candidates often use the question to segue into one of their main talking points and avoid answering difficult questions that would be difficult for even the best of candidates to answer **CITE**. We would like to explore if this same notion extends to debates as well. While the candidates are again being asked difficult questions, being surrounded by other candidates may force the candidates to be held more accountable as candidates who feel stronger about answering the questions asked could force the other candidates to remain on subject. However, it could also be the case that all candidates would like to avoid answering hard questions, and it would be within the best interests of the group to avoid answering difficult questions proposed by the moderators.

In this project, we will look to explore this question and see if candidates tend to talk about the same topics that moderators ask them about or if they tend to change the topic and discuss the topics that they feel most confident discussing. In order to conduct this analysis, we employed Latent Dirichlet Allocation (LDA) topic models on transcripts from all of the democratic debates in the 2020 election cycle thus far. Using this model, we form a posterior distribution of topics across all debates as well as a posterior distribution for any word given a topic. We visualize and authenticate our results, and conduct a more in-depth analysis to see how candidates respond to moderator questions. Finally, we include a case study on the two best performing candidates to see how the eventual nominee differed from his competitors.

We hope that this research will contribute to the ongoing discussion on holding leaders and politicians more accountable. Additionally, we could see uses for politicians to employ this research in framing the topics they discuss.

2 Data

The data used to conduct this analysis can be downloaded on Kaggle by clicking here. The dataset contains a written transcript of the entirety of the 2020 democratic primary debates. "This data comprises of data scraped from rev.com. All data is taken as is, and some transformations are made in order to add additional columns for speaker and speaking_time_seconds." The data is formatted so that each row is the beginning of a new speaker and all data is formatted chronologically by when the speech was given during a debate. The speeches are grouped by the date and debate as well. For our analysis, we will only use the ordering of the data, the speaker, and the physical speech, ignoring the speaking time variable as well as information about when a debate occurred.

Initial preprocessing included removing meaningless unicode characters, and cutting observations based on heuristics. We realized there were many instances where a moderator said "time's up" in the middle of a candidate's speech, and other instances where candidates would talk over each other until a moderator selected a designated speaker. We decided that these occurrences would only contribute noise to the topic modeling, so we aimed to remove these instances by removing all speeches with less than a certain threshold of words. As seen in Figure 1, the number of speeches seems to decrease quadratically until a threshold of twelve words, where we thought perhaps normal speeches began being removed. To test this hypothesis, we took a random sample of the 2,210 speeches removed at the twelve-word threshold and found that 9 out of 200 speeches had significant topic information. These numbers gave us an estimate that 99 out of the 2,210 speeches removed from this threshold were topic-rich, with a 95% confidence interval of [37, 161] (using an estimate of variance with a finite population correction factor). Thus, we used this threshold since it seemed to remove a lot of noise while removing barely any topic-rich speeches. After taking this survey, we realized that 157 speeches had unlabeled speakers, so we removed these as well since our documents are based on speaker identity. We believe removing these speeches from the dataset would hardly if at all affect the strong evidence that a twelve-word threshold is suitable for our means.

We also eliminated words that wouldn't help with topic modeling. Words that appeared once or twice were discarded since there is hardly any information associated with their few occurrences. We also eliminated eliminated contractions since no contractions have topic information. Furthermore, we set out to remove the most frequent words. We looked at the top 200 most frequent words, and decided which of these were important enough to keep in our dictionary. Finally, we removed punctuation and replaced each word with its lemmatized form so that our models wouldn't have to discover the similarity between words in their plural and singular forms for instance.

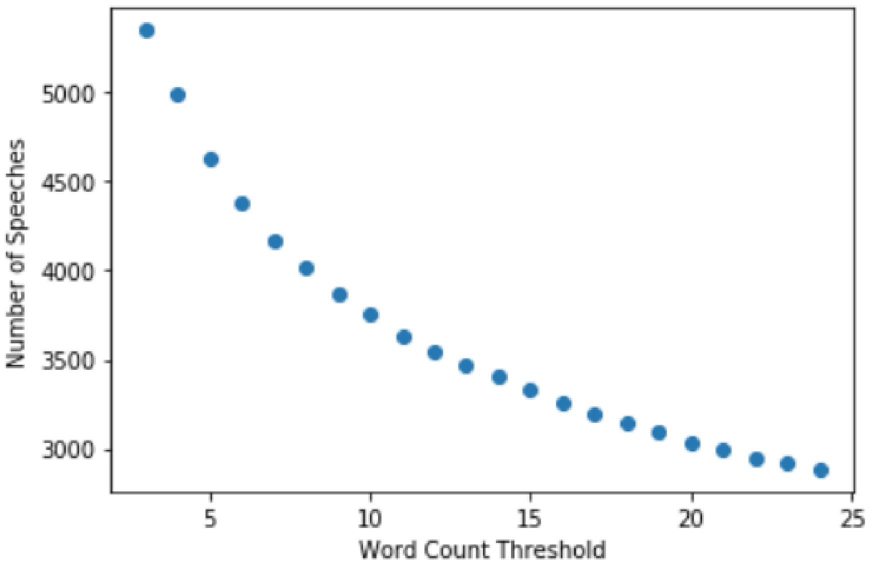


Figure 1: Dataset Reduction at Different Minimum Word Counts

3 Model

Several others have described the LDA probability model to its full extent, and we merely give a general overview to provide insight into the model. This framework assumes latent topic structure underneath the words within documents. Each document is represented as a bag of words, i.e. as the unique words in the document and their associated frequencies. This representation is simple and powerful enough for our goals, although we lose certain spatial relationships between words in this representation. In the LDA framework, we have priors on the distribution of topics among documents and distribution of words among topics. Each document consists of topics, and each word in a document is drawn from one of those topics. The posterior distribution is quite complicated for the model, so we end up using a collapsed gibbs sampler as provided by the lda package in R.

We chose to set uninformative priors for our model. We didn't know what the topics would be before fitting models. The α parameter is related to the distribution of topics for each document, so we set its value to a low scalar value (0.1), since there was no reason to favor certain topics over others. Additionally, a low value for this parameter lends itself to a document represented by fewer topics, which was desirable since each moderator question is only likely to pertain to one or two topics, and each response by all the candidates is likely to focus on only a couple topics. The other hyperparameter η is related to the distribution of words for each topic. Since there was no reason to favor certain words over others, we kept this as a low scalar value as well.

In order to analyze this problem from a few different perspectives and thus develop a more robust understanding of the issue, we ran the topic model on a few different corpora, defining the documents differently each time. For the first model that we ran, we treated each document as the concatenation of all contiguous sequences of the moderators and the concatenation of all contiguous sequences of the candidates. If different candidates spoke in between the speech of moderators, we concatenated their speeches as one document, doing the same for moderators. Once defining our documents in this way, we ran our model on these documents and compared the top topics discussed in each document of candidate speech with the top topics of preceding document of moderator speech. We defined the top topics to contain the most likely topics according to the documents Dirichlet distribution, choosing three to account for some error or variation in the model. We then looked to analyze how often candidates seemed to discuss the same topics moderators had just asked about.

For our second analyzing framework, we ran separate topic models on the candidate documents and the moderator documents, defining documents in the same manner as our first analysis. We then compared how these topics varied across the two groups of speakers. However, we believe that this analysis is less effective than our previous analysis because it does not include information about when each topic was discussed and gives a more broad level sense of topics discussed by each group.

Finally, we wanted to include a case study to analyze if the successful nominee discussed different topics than his competitors. For this case study, we chose Joe Biden, the eventual nominee, and Bernie Sanders, a candidate who we believed had similar name recognition to Biden and was polling fairly close to Biden for the majority of the primary. We hope that this selection would reduce latent factors such as polling position and name recognition that could influence the topics that candidates discuss. For this analysis, we recreated our vocabulary and our bag of words vectors independently for each candidate. We then ran the topic models and compared the top five topics that they each discussed in order to understand how their messaging differed. Theoretically, all leading candidates are given the chance to speak on each issue, so if they responded to the questions at hand, there would be an equal distribution of topics across each of their speeches. While the candidates are not always given the chance to respond equally within a given night, over all debates, it appears as if the speaking time per issue balances out between leading candidates. **CITE**

4 Results

4.1 Overall Debate Topics

We started with the full topic model including documents from both moderators and candidates. Before analyzing our model results, we wanted to ensure that the model fit was accurate. Therefore, we included a distribution map of our

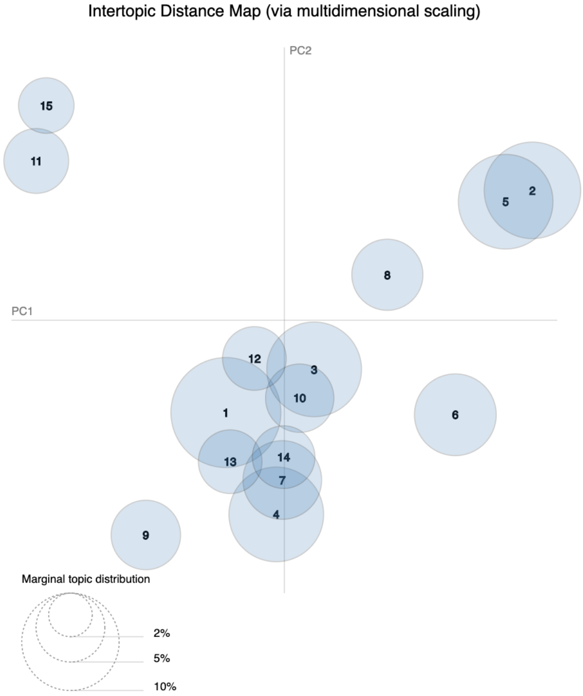
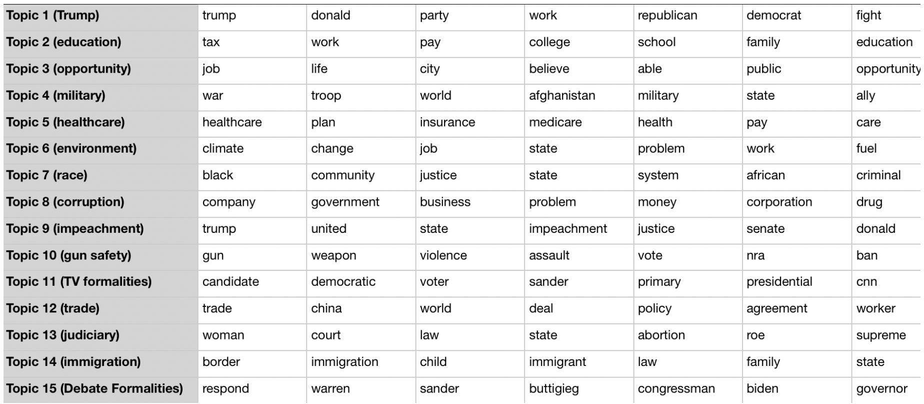


Figure 2: Intertopic Distance Map for Both Moderators and Candidates

data (figure 2) and manually checked that each of our 15 topics were coherent with their word distributions and that the topics aligned with a topic that we believed was relevant to politics. We also include the top words for each topic in figure 3. It can be seen that they are fairly coherent. For instance, topic 4 relates to the military while topic 5 relates to healthcare.

From these figures, we felt confident that the model was both capturing the distribution of topics and developing coherent topics well as there was not a large amount of overlap between topics and it is fairly easy to determine what political topic each distribution of words is referring to.

The topics covered in all debates range from healthcare to gun control, global affairs, military, and criminal justice. Not surprisingly, the most prominent topic in our posterior distribution (topic 1 in Table 3) is related to the upcoming election and the relationship between the Democratic and Republican party, particularly Donald Trump. In this table and the ensuing tables, topics closer to the top of the table are denser in the posterior distribution of topics for a given document. For a given topic, words closer to the left of the table are more likely in the posterior distribution of words.



Topic 1 (Trump)	trump	donald	party	work	republican	democrat	fight
Topic 2 (education)	tax	work	pay	college	school	family	education
Topic 3 (opportunity)	job	life	city	believe	able	public	opportunity
Topic 4 (military)	war	troop	world	afghanistan	military	state	ally
Topic 5 (healthcare)	healthcare	plan	insurance	medicare	health	pay	care
Topic 6 (environment)	climate	change	job	state	problem	work	fuel
Topic 7 (race)	black	community	justice	state	system	african	criminal
Topic 8 (corruption)	company	government	business	problem	money	corporation	drug
Topic 9 (impeachment)	trump	united	state	impeachment	justice	senate	donald
Topic 10 (gun safety)	gun	weapon	violence	assault	vote	nra	ban
Topic 11 (TV formalities)	candidate	democratic	voter	sander	primary	presidential	cnn
Topic 12 (trade)	trade	china	world	deal	policy	agreement	worker
Topic 13 (judiciary)	woman	court	law	state	abortion	roe	supreme
Topic 14 (immigration)	border	immigration	child	immigrant	law	family	state
Topic 15 (Debate Formalities)	respond	warren	sander	buttigieg	congressman	biden	governor

Figure 3: Top Words per Topic for Both Moderators and Candidates

4.2 Moderator versus Candidate Topics

In order to assess whether candidates were talking about issues prompted by the moderators, we employed quantitative and qualitative approaches. First, using our topic model discussed above, we looked at the posterior distribution of topics for each document. Since there were over 1,300 documents, we developed a “top-n” metric to measure similarity between a moderator prompt and the candidates’ response. As discussed in the Model section, in this metric, we originally considered the former document similar to the latter document if any of the former’s top three topics in its posterior distribution agree with the latter’s top three topics. We found 79.1% of the document pairs were similar under the top-three metric.

We thought that 79.1% was fairly high and provided some evidence against the stereotype that candidates dodge questions. We decided to extend our metric and found that 67.6% of document pairs were similar within the top two topics, and 42.6% of document pairs had the same most likely topic.

4.3 Separate Moderator and Candidate Topic Models

In our qualitative analysis of the similarity between moderator and candidate talking points, we created one topic model built solely on contiguous moderator speeches and another built on contiguous candidate speeches. Similarly to the full topic model in the prior subsection, these models showed coherence in topics and not too significant of similarity between topics (as seen in the maps in Figure 4 and Figure 5).

We find that the top topics for moderators were conflated with words related to TV formalities. For instance, in Table 6, topic three’s third most likely word was CNN. As another example, topic 4 consists of names and election terminology, likely originating from when moderators address candidates or talk

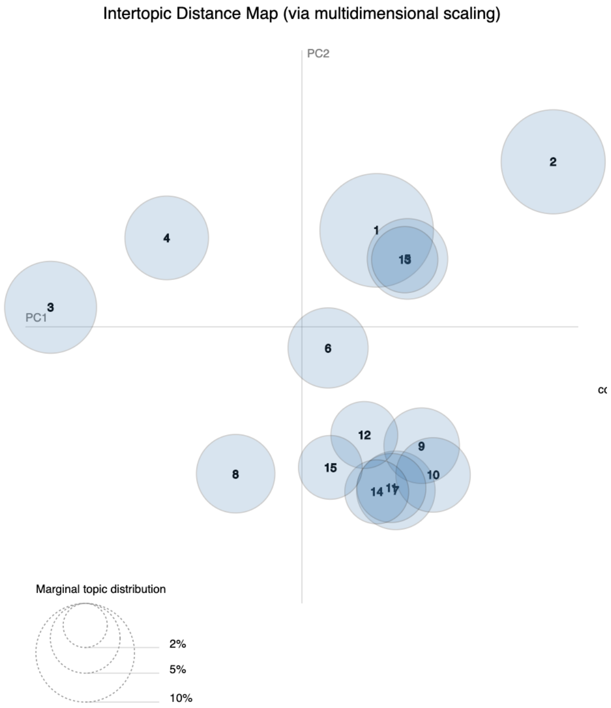


Figure 4: Intertopic Distance Map for Moderators

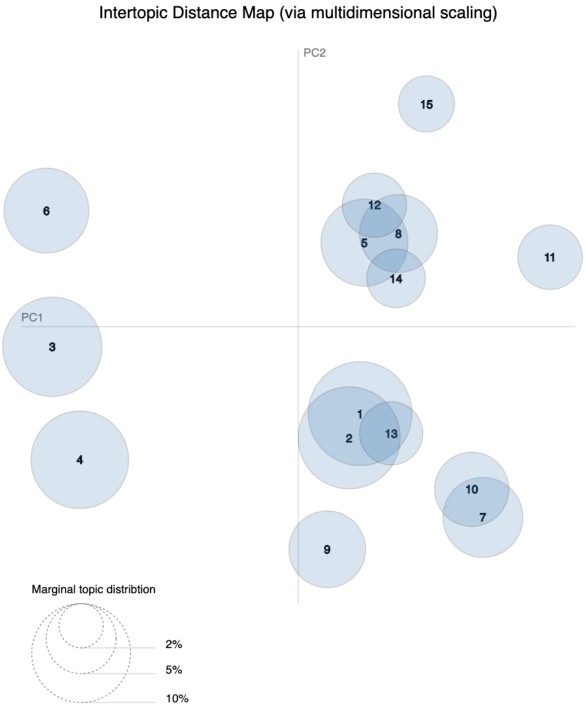
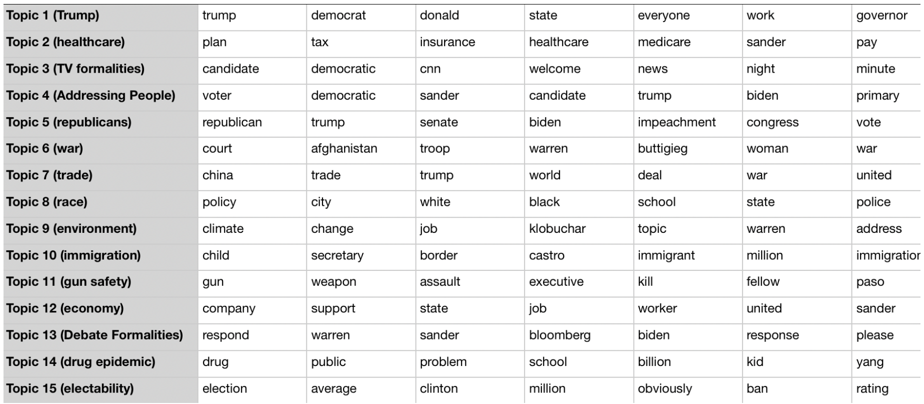


Figure 5: Intertopic Distance Map for Candidates



Topic 1 (Trump)	trump	democrat	donald	state	everyone	work	governor
Topic 2 (healthcare)	plan	tax	insurance	healthcare	medicare	sander	pay
Topic 3 (TV formalities)	candidate	democratic	cnn	welcome	news	night	minute
Topic 4 (Addressing People)	voter	democratic	sander	candidate	trump	biden	primary
Topic 5 (republicans)	republican	trump	senate	biden	impeachment	congress	vote
Topic 6 (war)	court	afghanistan	troop	warren	buttigieg	woman	war
Topic 7 (trade)	china	trade	trump	world	deal	war	united
Topic 8 (race)	policy	city	white	black	school	state	police
Topic 9 (environment)	climate	change	job	klobuchar	topic	warren	address
Topic 10 (immigration)	child	secretary	border	castro	immigrant	million	immigratior
Topic 11 (gun safety)	gun	weapon	assault	executive	kill	fellow	paso
Topic 12 (economy)	company	support	state	job	worker	united	sander
Topic 13 (Debate Formalities)	respond	warren	sander	bloomberg	biden	response	please
Topic 14 (drug epidemic)	drug	public	problem	school	billion	kid	yang
Topic 15 (electability)	election	average	clinton	million	obviously	ban	rating

Figure 6: Top Words per Topic for Moderators

of logistics.

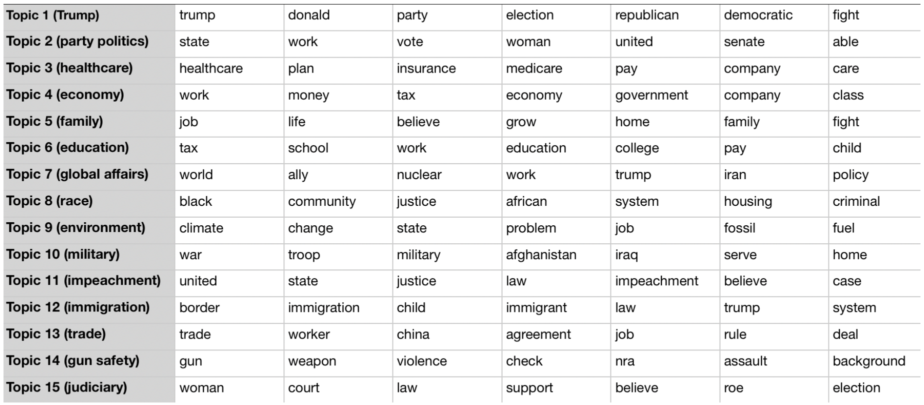
Besides these functional topics, we find that the topics in the moderator model are quite similar to those in the candidate model. Topic 11 for moderators (see Table 6) and topic 14 for candidates (see Table 7) both are about gun violence, and topic 8 for moderators and topic 8 for candidates both refer to criminal policy.

4.4 Case Study: Biden vs. Bernie

In our case study of Bernie Sanders and Joe Biden, we see that the two candidates share a few topics that they discuss, such as healthcare, gun safety, education, trade, and race. However, they also diverge on many topics discussed. For instance, Bernie talks much more about grassroots activism and corruption while Biden spends more time talking about Trump and his accomplishments with the Obama administration. Even when discussing the same topics, you see a difference in their tones as well. For instance, you see Bernie reference Joe Biden by name a lot in many of the topics he discusses, indicating that he is attacking Joe Biden's position on these issues. Most notably, this includes the Iraq war and trade. Meanwhile, Joe Biden seems to focus his own accomplishments and positions, instead of mentioning Bernie or any of the other candidates by name.

5 Conclusion

Overall, found that LDA topic models produced fairly coherent topics for political debates. Any of our topic models (especially the full model) can inform us that issues of this election cycle include gun control, criminal justice, trade with China, and more. We found evidence that candidates talk about topics at least



Topic 1 (Trump)	trump	donald	party	election	republican	democratic	fight
Topic 2 (party politics)	state	work	vote	woman	united	senate	able
Topic 3 (healthcare)	healthcare	plan	insurance	medicare	pay	company	care
Topic 4 (economy)	work	money	tax	economy	government	company	class
Topic 5 (family)	job	life	believe	grow	home	family	fight
Topic 6 (education)	tax	school	work	education	college	pay	child
Topic 7 (global affairs)	world	ally	nuclear	work	trump	iran	policy
Topic 8 (race)	black	community	justice	african	system	housing	criminal
Topic 9 (environment)	climate	change	state	problem	job	fossil	fuel
Topic 10 (military)	war	troop	military	afghanistan	iraq	serve	home
Topic 11 (impeachment)	united	state	justice	law	impeachment	believe	case
Topic 12 (immigration)	border	immigration	child	immigrant	law	trump	system
Topic 13 (trade)	trade	worker	china	agreement	job	rule	deal
Topic 14 (gun safety)	gun	weapon	violence	check	nra	assault	background
Topic 15 (judiciary)	woman	court	law	support	believe	roe	election

Figure 7: Top Words per Topic for Candidates

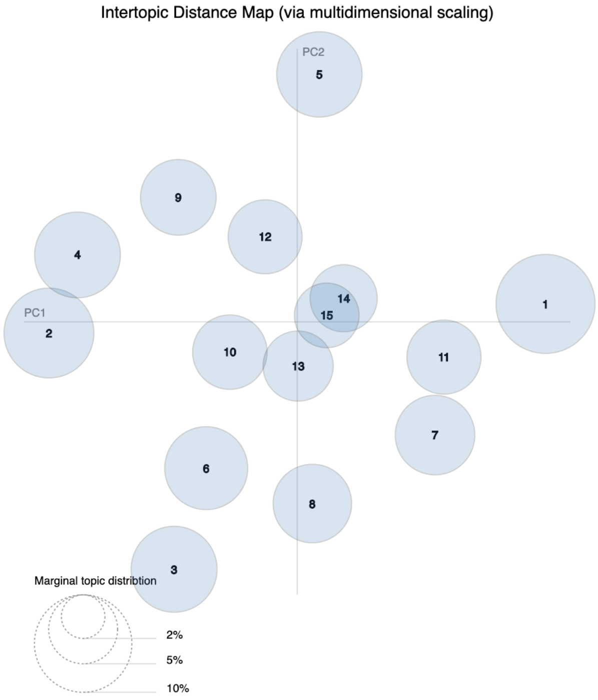
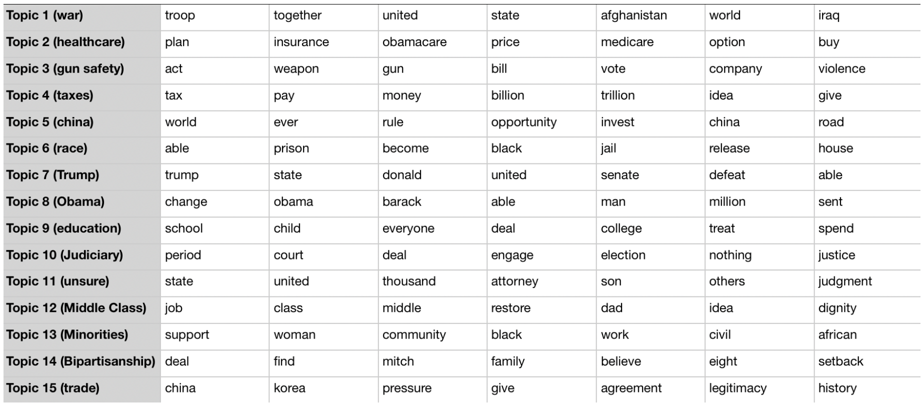


Figure 8: Intertopic Distance Map for Joe Biden



Topic 1 (war)	troop	together	united	state	afghanistan	world	iraq
Topic 2 (healthcare)	plan	insurance	obamacare	price	medicare	option	buy
Topic 3 (gun safety)	act	weapon	gun	bill	vote	company	violence
Topic 4 (taxes)	tax	pay	money	billion	trillion	idea	give
Topic 5 (china)	world	ever	rule	opportunity	invest	china	road
Topic 6 (race)	able	prison	become	black	jail	release	house
Topic 7 (Trump)	trump	state	donald	united	senate	defeat	able
Topic 8 (Obama)	change	obama	barack	able	man	million	sent
Topic 9 (education)	school	child	everyone	deal	college	treat	spend
Topic 10 (Judiciary)	period	court	deal	engage	election	nothing	justice
Topic 11 (unsure)	state	united	thousand	attorney	son	others	judgment
Topic 12 (Middle Class)	job	class	middle	restore	dad	idea	dignity
Topic 13 (Minorities)	support	woman	community	black	work	civil	african
Topic 14 (Bipartisanship)	deal	find	mitch	family	believe	eight	setback
Topic 15 (trade)	china	korea	pressure	give	agreement	legitimacy	history

Figure 9: Top Words per Topic for Joe Biden

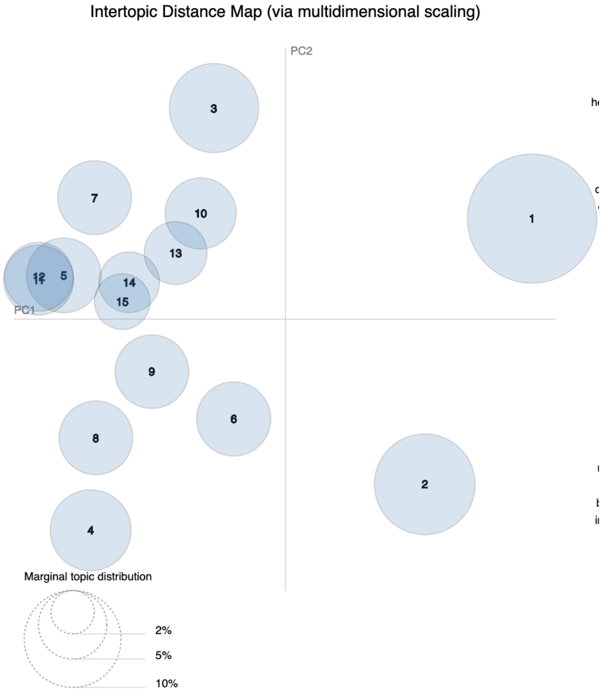
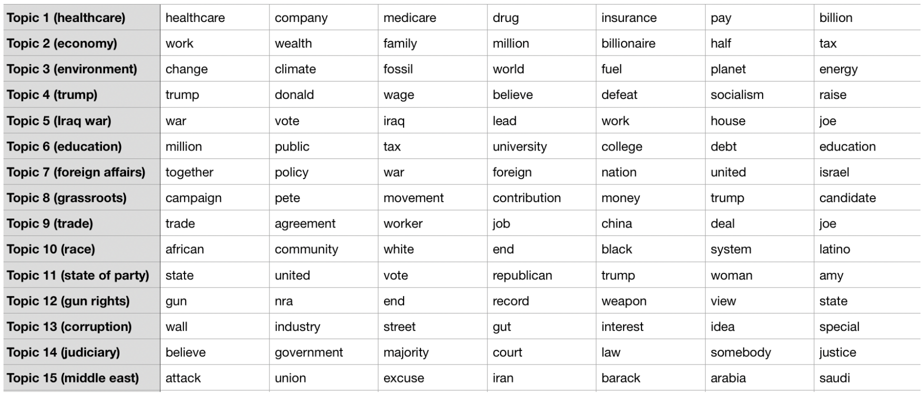


Figure 10: Intertopic Distance Map for Bernie Sanders



Topic 1 (healthcare)	healthcare	company	medicare	drug	insurance	pay	billion
Topic 2 (economy)	work	wealth	family	million	billionaire	half	tax
Topic 3 (environment)	change	climate	fossil	world	fuel	planet	energy
Topic 4 (trump)	trump	donald	wage	believe	defeat	socialism	raise
Topic 5 (Iraq war)	war	vote	iraq	lead	work	house	joe
Topic 6 (education)	million	public	tax	university	college	debt	education
Topic 7 (foreign affairs)	together	policy	war	foreign	nation	united	israel
Topic 8 (grassroots)	campaign	pete	movement	contribution	money	trump	candidate
Topic 9 (trade)	trade	agreement	worker	job	china	deal	joe
Topic 10 (race)	african	community	white	end	black	system	latino
Topic 11 (state of party)	state	united	vote	republican	trump	woman	amy
Topic 12 (gun rights)	gun	nra	end	record	weapon	view	state
Topic 13 (corruption)	wall	industry	street	gut	interest	idea	special
Topic 14 (judiciary)	believe	government	majority	court	law	somebody	justice
Topic 15 (middle east)	attack	union	excuse	iran	barack	arabia	saudi

Figure 11: Top Words per Topic for Bernie Sanders

somewhat related to what is asked of them (at least in debate format). There is evidence, however, that candidates stray from the question about a third of a time, if one measures similarity with a stricter metric. These findings seem to align with the popular stereotype that candidates are crafty with words by transitioning from what is prompted to what they want to talk about.

From a qualitative perspective, the overall topics discussed by moderators were quite similar to those discussed by candidates. This perhaps emphasizes that there's a symbiotic relationship between moderators and candidates, and that debates are a good forum for candidates to address what issues they want to and need to talk about.

Finally, from an closer look at the two top candidates, we found that the eventual nominee and successful candidate focused more on his own positions and less on other candidates, indicating to future candidates that it might be productive to focus on your own stances instead of attacking other candidates. Additionally, Joe Biden mentioned talked more about well-known politicians, such as Obama and Trump, than Bernie. This may indicate that viewers respond to individuals they are familiar with and candidates aligning themselves with or against these popular individuals could be beneficial to their candidacy.