Decoding Tweets to Forecast Mid-Term Elections

Team Members: *Anant Vashistha, Arushi Jain, Nikos Galanos*

MOTIVATION AND OBJECTIVE

Social channels are the primary source of spreading information and understanding public opinion. Social media has impacted our society, political discourse, and media landscape through hashtag activism on recent affairs such as the Overturning of Roe vs. Wade, Covid-19, and Black Lives Matter. Many parties also routinely use these channels for political campaigns and marketing activities for maximum public outreach.

Twitter data shows that the election candidates and the general public were pretty active on Twitter before and during the US mid-term elections. We decided to leverage the tweets of the candidates and the public's reaction to those tweets as proxies for forecasting the mid-term election outcomes.

We developed a model to predict the outcome of US mid-term elections by identifying the critical topics (based on candidates' tweet data) and analyzing public sentiment from the replies to those tweets. The model used the actual election outcome as the dependent variable predicted using a combination of sentiments and topics as well as additional Twitter metadata such as likes, retweets, and followers as features for each candidate. The model can be expanded to any election, given that the required tweets of the involved candidates, their public replies, and additional metrics are collected and analyzed.

The report discusses a high-level view of the dataset (scraping, cleaning, and EDA), methodology (topic modeling, sentiment analysis, and feature engineering), model results, conclusion, and challenges.

DATASET

Data Scraping

- We collected candidate state and party information from Ballotpedia.
- For each candidate, we searched their Twitter handles and used Tweepy's User Lookup API and Tweet Lookup API to get the user information and tweet data from 15th Sep - 15th Oct and their reply data from 8th Oct - 15th Oct. The number of replies we could fetch was limited to a week due to the limited number of tweets we could scrape (25000) as a part of Twitter's Developer API's Elevated Access. Nevertheless, we fetched up to 200 tweets per user and up to 1000 replies for each user. However, due to these limitations, we couldn't fetch reply data for all the scraped tweets.
- We collected the election results information (number of votes, vote share, winner) from npr.org

Data Cleaning

We transformed the tweets and the replies data by converting all tweets to lowercase, removing hashtags, links, and mentions, and removing stop words and punctuations to create a cleaned corpus of tweets and their replies.

Exploratory Data Analysis

We conducted extensive EDA on the scraped data containing candidates' tweets and the public's replies to understand the party-wise and state-wise activity. The EDA was conducted on 93 candidates out of 129 candidates for the 33 states on approximately 9.6K candidate tweets and 41K public replies. As per the analysis, candidates from Pennsylvania were the most active (54k tweets) with 276K public replies.

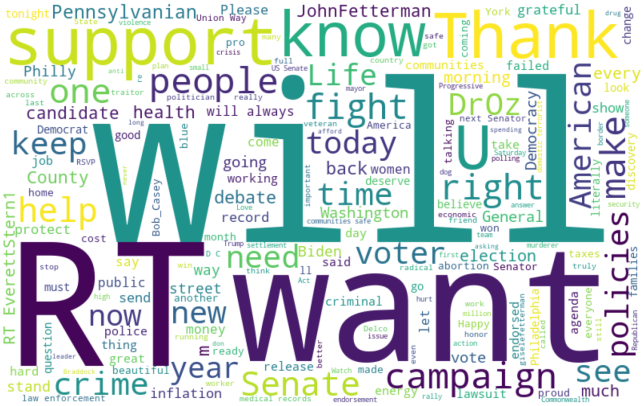


Fig. 1: Word Cloud on Tweets of Candidates from Pennsylvania

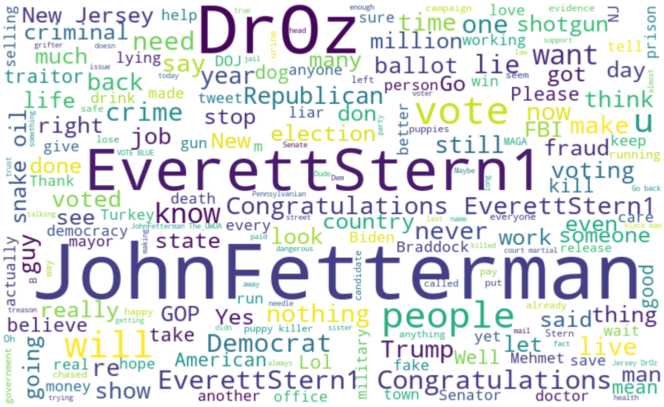


Fig. 2: Word Cloud on Replies to Candidates from Pennsylvania

As per the above word clouds on candidates' tweets and public replies data for Pennsylvania state, the candidates majorly talked about inflation, crime, women, while the public sentiments were relatively negative (fraud, traitor, lie).

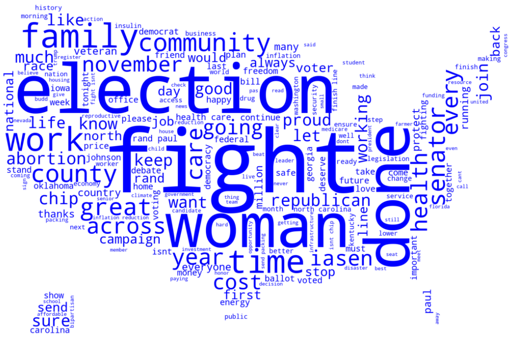


Fig. 3: Word Cloud on Democrats Tweets

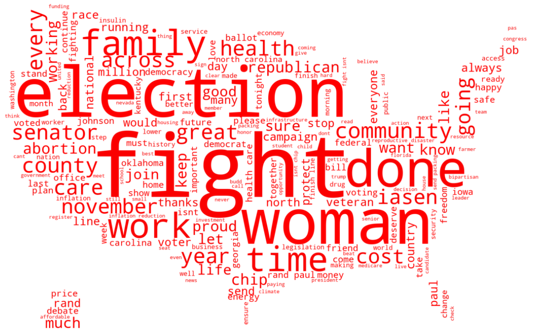
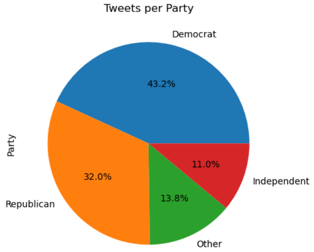


Fig. 4: Word Cloud on Republican Tweets



The above Democrats and Republican-specific word clouds showed a strong emphasis on critical topics such as women, family, abortion, health, etc. They can be further studied to understand party-wise agendas and important manifesto points.

The dataset contains 35 republicans, 33 democrats, 10 independents, and 15 "other" candidates. As per the pie chart (Fig. 5), Democrats (4154 tweets) are more active on Twitter as compared to Republicans (3083 tweets). In the dataset, out of 93 candidates, 27 are currently serving. A party-wise study reveals that out of 33 Democratic candidates running for mid-term elections, 13 are currently serving whereas out of 35 republicans, 14 are currently serving.

Fig. 5: Party-wise Tweet Distribution

METHODOLOGY

Topic Modeling

We first represent our tweets as a TF-IDF Vector (We use Inverse Document Frequency along with Term Frequency to retain only the most relevant terms across tweets) with a vocabulary of 3331 words.

We use Latent Dirichlet Allocation (LDA) to classify tweets (documents) into multiple topics and topics into multiple words using Dirichlet distributions. This model assumes that each document is a weighted collection of topics and each topic is a weighted collection of certain combinations of keywords that always appear together.

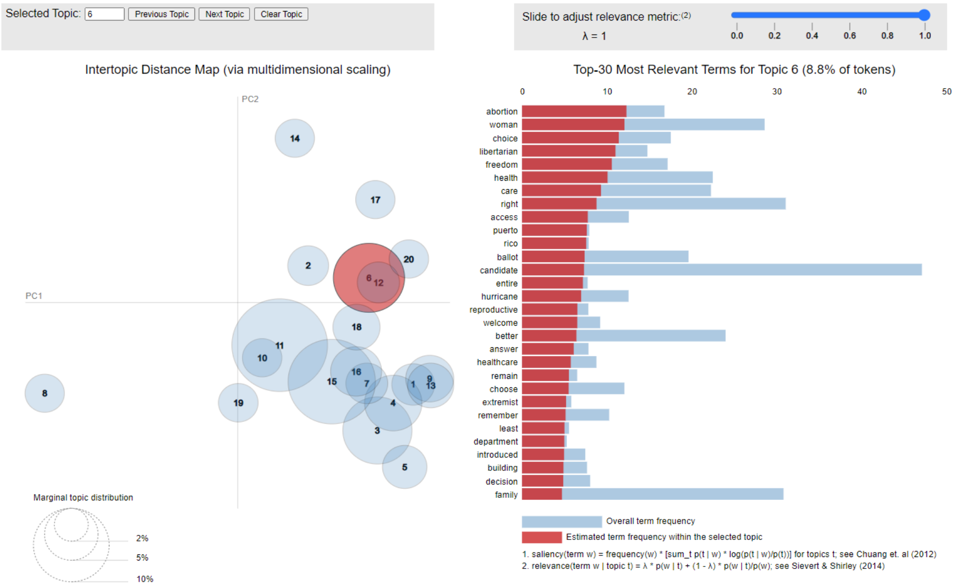

Selected Topic: 6
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
14
17
2
6
12
20
18
11
10
16
15
7
1
9
13
8
19
4
3
5
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 6 (8.8% of tokens)
0
10
20
30
40
50
abortion
woman
choice
libertarian
freedom
health
care
right
access
puerto
rico
ballot
candidate
entire
hurricane
reproductive
welcome
better
answer
healthcare
remain
choose
extremist
remember
least
department
introduced
building
decision
family
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 6: LDA Model - Results for Topic 6

For our corpus of candidate tweets, we ran the LDA model with a topic parameter of 20 topics. The resulting output produced a tweet-topic matrix where each cell represented the importance of a topic in a given tweet as well as a topic-keyword matrix where each cell represented the importance of a given keyword in a topic.

The given graph shows the PCA of our topics and the inter-topic distance among topics. Topic 6 is the highlighted topic, and on the right, we see the most commonly occurring terms for this topic.

To understand the individual topics better, we tried to extract some representative tweets where the following sample topics had maximum contribution (The topic names are not produced by the model but interpreted by us based on the keywords and associated tweets):

Topic 1 Top Keywords: *social, cost, medicare, security, senior, insurance, prescription, insulin*
Example: **@bittergertrude You are so right. No matter what your health conditions are, you deserve high quality healthcare with no out of pocket costs! I'm going to the Senate to do all I can to make that happen** *(Topic Contribution: ~80%)*

Topic 2 Top Keywords: *abortion, woman, choice, libertarian, freedom, health, care, right, access*
Example: **RT @PattyMurray: It's heartbreaking &amp; wrong that a patient from Texas had to fly to WA state just to get abortion care** *(Topic Contribution: ~73%)*

Topic 3 Top Keywords: corporation, profit, resolution, location, inflation, energy
Example: **Food and energy prices keep rising as inflation breaks expectations again** *(Topic Contribution: ~69%)*

Refer to the Appendix for more examples.
Since we have only 93 candidates and we plan to use topics as our features, we cannot have 20 topics. As such we combined similar-looking topics by manually looking at the top topic keywords, the top 30 most relevant terms from the topic graph, and sample representative tweets to arrive at **7** topics: **Election Campaigns, Inflation/ Economy, Healthcare, Abortion, Crime/ Race, Students/ Schools, and Current Events Tweets** (HurricaneIan/ Ukraine/ Iran) for each candidate tweet (Refer to the Appendix for further details)

Sentiment Analysis

We conducted sentiment analysis on the replies dataset containing 36K replies to gauge the public sentiment on the candidate's tweets using three different models: Textblob, Flair, and Distil Bert. The replies dataset was cleaned extensively during the analysis to remove hashtags, emojis, special characters, spaces, and unnecessary characters. The cleaned dataset containing more than 4 words was used for analyzing sentiments.

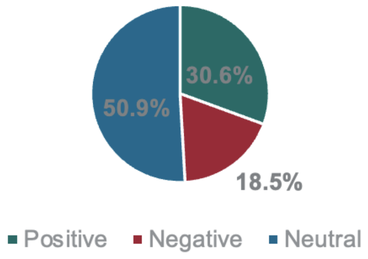
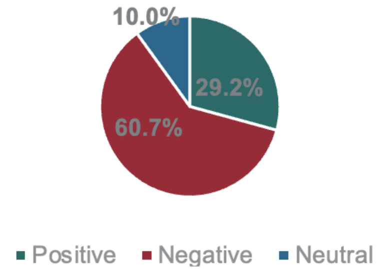
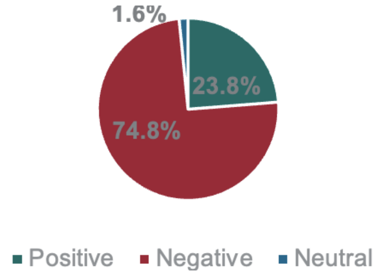
	Blob	**Flair**	**DistilBert**
Distribution			
Positive Example	@TheOtherMandela The Republican Party is the party of Parental rights, family values and Law and order. Ron Johnson will continue to support policies that help working class families rise in life. He is a Truth Teller who exposes corruption in Washington. He is the perfect candidate for WI.		
	Positive → 1	Positive → 0.9916	Positive → 0.997
Negative Example	@SenatorLankford Nothing on leader Pelosi's husband? Latest reporting was he was going to break her kneecaps. You and your goon party are responsible for this. And you can't even make a rudimentary comment because your goons will turn on you. Pathetic.		
	Negative → -0.1	Negative → -0.999	Negative → -0.99

Table 1: Sentiment Analysis - Techniques and Results

Feature Engineering

Our topic model is at a tweet level while the sentiment analysis is at a reply level. However, we must bring all our features to a senate candidate level. Recall that we do not have reply data for all the tweets due to our scrape limits. Hence, we must make some assumptions before extrapolating our reply dataset with sentiment scores to our tweet datasets. We arrive at candidate-level data in the following steps (refer to the Appendix for an example):

- Compute the average reply sentiment per candidate
- Multiply the average sentiment per candidate with the number of replies on each tweet to get the weighted tweet sentiment
- Multiply the weighted tweet sentiment with the topic contributions for each tweet to get topic sentiments for each tweet
- Average the topic sentiments per candidate on their tweets for each of the 7 topics

We repeat this exercise for our 3 sentiment models to get 3 datasets. We also use average likes per candidate, average retweets per candidate, and the number of candidates' followers as additional metadata information along with party information and information about whether the candidate is currently serving as a candidate (boolean). Hence, we arrive at **12 features for 93 candidates.**

Since we need winners at the state level, instead of predicting a binary variable of win v/s loss, we predict the continuous variable vote share for each candidate. Then we rank the predicted vote shares of each candidate by state, and the candidate with the greatest vote share is predicted as the winner for that state while the other candidates are predicted as the losing candidates.

RESULTS

We do the train-test split of 75-25 stratified by candidate party & election result and run various regression models to predict the candidate vote share with 5-fold cross-validation for each model. Once we have the predicted vote share per candidate, we group the candidates by state and declare the candidate with the highest vote share as the winner.

Baseline (Predicting everyone as not-elected): Train = 42/65, Test = 18/28						
Model	**TextBlob**		**Flair**		**DistilBert**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Linear	59/65	22/28	59/65	22/28	60/65	25/28
Lasso	60/65	23/28	59/65	22/28	60/65	25/28
Ridge	61/65	24/28	59/65	22/28	60/65	25/28
CART	62/65	23/28	56/65	21/28	61/65	26/28
Random Forest	62/65	21/28	64/65	23/28	61/65	26/28
Bagging	57/65	20/28	53/65	22/28	56/65	23/28
Adaboost	59/65	22/28	63/65	24/28	55/65	22/28

Most Important Features

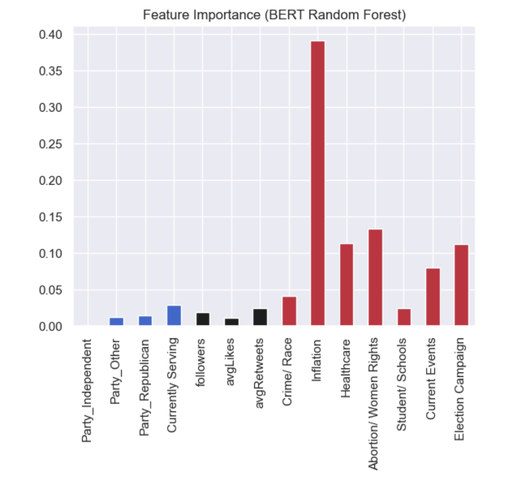
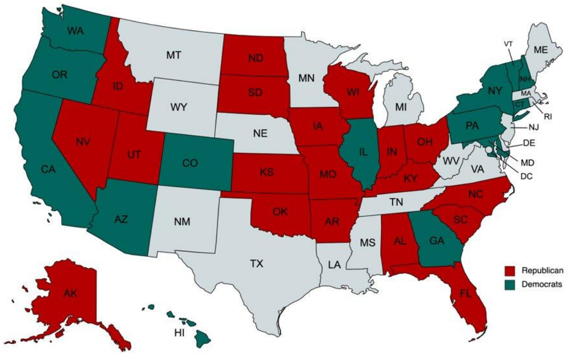


Table 2: Predictive Models for Predicting Winning Candidates



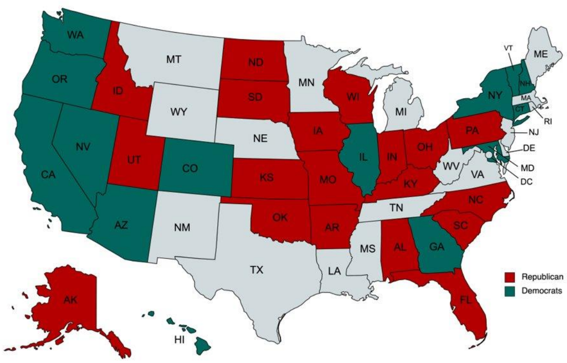


Fig. 7: Actual Results

Fig. 8: Model Results

The BERT model with Random Forest correctly classified the winning party in 31 out of 33 states and winning candidates in 30 out of 33 states. According to media reports, the incorrectly predicted states (party-level) were Nevada and Pennsylvania, out of which Nevada had a very close call. Based on the variable importances given by the Random Forest Model of DistillBert, the most important features were **Inflation, Abortion, and Healthcare**.

CONCLUSION

The results from our models show that social media can be a powerful tool in predicting important political and social outcomes. Through the limited data on tweets, replies, and Twitter metadata information of the contesting candidates, we could correctly predict the election results for 87 out of 93 candidates (61/65 in-sample and 26/28 out-of-sample). We also found that apart from the party information and the information about if the candidate is currently serving, the topics that the contesting candidates talk about are the most important indicators for the election results - Inflation, Abortion, and Healthcare, in our case. The significance of these topics also makes intuitive sense considering the fears of recession, the controversy around the overturning of Roe vs. Wade, and high healthcare costs in the USA.

CHALLENGES AND LIMITATIONS

Our analysis was based on the Twitter data of the candidates, and as such suffered from the limited number of tweets and replies that we could pull due to the scrape limits of the Twitter APIs. We also had limited data in terms of using only this year's candidates for training and predictions. Additionally, since reply data was missing for most tweets of each candidate, we had to make strong assumptions about the reply sentiments. Also, we had a lot of missing location data and as such we could not leverage the geographical data of the people who replied to the candidates and assumed that all replies will affect the election outcomes instead of the replies limited to the state of the candidate.

APPENDIX

More examples of important topics with dominant contributions in sample tweets:

Topic Name	Keyword Distribution	Sample Tweets	Topic Contribution
Healthcare	0.012*"social" + 0.011*"cost" + 0.011*"medicare" + 0.010*"security" + 0.010*"senior" + 0.009*"insurance" + 0.009*"discus" + 0.009*"prescription" + 0.008*"farmer" + 0.008*"insulin"	@bittergertrude You are so right. No matter what your health conditions are, you deserve high quality healthcare with no out of pocket costs! I'm going to the Senate to do all I can to make that happen.	0.794
		More than one million North Carolinians live with diabetes. But Congressman Budd voted against capping insulin costs because he's fighting for his Big Pharma donors – not North Carolinians.	0.787
		We're still fighting with insurance companies today, just like so many North Carolinians. It shouldn't be this way – and it doesn't have to be.	0.732
Abortion/ Reproductive Rights/ Women's Rights	0.015*"abortion" + 0.014*"woman" + 0.014*"choice" + 0.013*"libertarian" + 0.013*"freedom" + 0.012*"health" + 0.011*"care" + 0.010*"right" + 0.009*"access" + 0.009*"puerto"	Reproductive health care decisions should be made by patients, not politicians. I will always fight to support a woman's right to choose.	0.789
		If Ted Budd's nationwide abortion ban becomes law, women will die. It's clear Budd is too extreme for North Carolina.	0.779
		RT @PattyMurray: It's heartbreaking &amp; wrong that a patient from Texas had to fly to WA state just to get abortion care.	0.73
		@janecoaston "right to sex" is not fundamentally different than "right to education" or "right to housing" if you support one of these, you should really support them all	0.718
		Abortion is health care. I'll vote to codify Roe into law in the Senate.	0.716
Crime/ Race	0.011*"chase" + 0.010*"police" + 0.009*"border" + 0.009*"terrorist" + 0.008*"black" + 0.007*"criminal" + 0.007*"crime" + 0.007*"statement" + 0.007*"southern" + 0.006*"enforcement"	'I just think it's B.S,' Lowry said of ads portraying [@CheriBeasleyNC] as pro law enforcement. 'She's for the criminals and not for the law officers, which my brother was one of them that got killed…' #ncsen #ncpol	0.714
		NKY law enforcement backs Rand Paul, who says he is against 'defunding the police'	0.702
		It's absurd to think someone entering the US by breaking the law is going to follow instructions when it's time to check in with ICE. This report @JohnCornyn &amp; I requested makes it clear that the southern border remains open. #BidenBorderCrisis	0.667
Inflation/ Economy	0.008*"corporation" + 0.008*"easy" + 0.007*"profit" + 0.007*"based" + 0.007*"created" + 0.007*"resolution" + 0.007*"location" + 0.007*"inflation" + 0.006*"energy"	Food and energy prices keep rising as inflation breaks expectations again.	0.693
		I'm Michelle Lewis, the write-in Unaffiliated candidate for the US Senate in NC. I'm running because we need to become an energy independent nation. Follow me if you agree we shouldn't be dependent on other countries for our energy needs.	0.698
		RT @ActionDemocrat: Mean while Democrats are delivering while Republicans scheme to take benefits away	0.683
School/ Children	0.023*"read" + 0.020*"page" + 0.018*"worker" + 0.015*"website" + 0.012*"speak" + 0.011*"sent" + 0.011*"school" + 0.011*"education" + 0.009*"save"	For our kids to reach their full potential, we have to invest in education. Providing our classrooms with the resources they need and empowering our teachers will set our kids up for success. Defunding public education isn't the answer.	0.657
		@mattmccoin Part of that is parents going to all parent teacher meetings. There is no way you would believe any of the BS politically motivated noise makers are spouting. Teachers and Librarians are on our side and not indoctrinating our children.	0.552

Table A.1: Example Tweets for Sample Topics

Bringing reply and tweet-level features to candidate-level features: An Example

We will

- Compute the average reply sentiment per candidate: Table A.2 shows how we have computed the average reply sentiment for Candidate 1 for whom reply data was available only for 2 out of 6 tweets.

Candidate	Tweet ID	ReplyID	Reply Sentiment
1	3	1	0.6
1	3	2	-0.7
1	6	3	1.0
Average Sentiment			0.3

Table A.2: Computing Average Reply Sentiment per Candidate

- Multiply the average sentiment per candidate with the number of replies on each tweet to get the weighted tweet sentiment: Table A.3 shows computing the individual tweet sentiment based on the average sentiment of the candidate and the number of replies for all tweets of Candidate 1.

Candidate	Tweet ID	Number of Replies	Tweet Sentiment
1	1	100	0.3 * 100 = 30
1	2	50	0.3 * 50 = 15
1	3	20	0.3 * 20 = 6
1	4	100	0.3 * 100 = 30
1	5	30	0.3 * 30 = 9
1	6	80	0.3 * 80 = 24

Table A.3: Computing Weighted Tweet Sentiments

- Multiply the weighted tweet sentiment with the topic contributions for each tweet to get topic sentiments for each tweet: Table A.4 shows the computations of the topic sentiments with 3 topics for Candidate 1 based on the corresponding tweet sentiments.

Candidate	Tweet ID	Tweet Sentiment	Topic 1		Topic 2		Topic 3	
			Contribution	Score	Contribution	Score	Contribution	Score
1	1	30	0.3	9	0.3	9	0.4	6
1	2	15	0.4	6	0.3	4.5	0.3	4.5
1	3	6	0.1	0.6	0.1	0.6	0.8	4.8
1	4	30	0.8	24	0	0	0.2	6
1	5	9	0.7	6.3	0.3	2.7	0	0
1	6	24	0.2	4.8	0.3	7.2	0.5	12

Table A.4: Computing Weighted Topic Sentiments Per Tweet

- Average the topic sentiments per candidate on their tweets for each of the 7 topics: Table A.5 shows the candidate-level features after averaging the tweet level sentiments across topics for each topic.

Candidate	Average Topic 1 Sentiment	Average Topic 2 Sentiment	Average Topic 3 Sentiment
1	8.616	4	5.55

Table A.5: Computing Topic-level Features For Each Senator