Unleashing the Power of Twitter: A Data Analysis of the US Senate's Social Media Strategies with Unsupervised Machine Learning

Miguel Cozar [1] **Carlos Munoz** [1] **Kai Shu** [2]

Abstract

Social media, such as Twitter, plays a crucial role in political discourse and communication. It is the window of voters to their candidates, and what senators publish may determine their success in the elections. A deep analysis is needed to comprehend the current situation and generate strategies to reach the audience. This paper joins the creation of a self-made dataset, using machine learning topic models, analyzing how geography influences the political landscape, and employing a proposed popularity metric to explain the current political landscape and provide insights about the most influential senators and their discourse.

1. Introduction

With 80 million active users worldwide, Twitter provides a unique opportunity for politicians to reach out to their constituents and shape public opinion. It is, therefore, crucial to understand the dynamics of Twitter in the context of politics.

This paper presents a deep data analysis of the US Senate's Twitter activities. We introduce several data collection techniques to build our dataset of the Twitter network of the current Senate. We also collect a large corpus of tweets posted by US Senators on Twitter. We perform a topic modeling of the corpus leveraging a fine-tuned version of BERTopic (Grootendorst, 2022). Finally, we design advanced popularity metrics to perform further analysis.

Our primary objective is to identify the network's most relevant nodes (senators), study the influence of external factors, e.g., the senator's geographic location, and determine which topics are more important to the voters. By leveraging the power of unsupervised machine learning, we can discover meaningful patterns and structures in large amounts of data generated by US Senators without requiring prior knowledge or labeling of the data.

Our analysis uncovers outstanding findings. First, with network statistics, we discover that the Democratic and Republican senator sub-networks have very different structures. Second, we identify 13 topics whose influence and popularity vary dramatically. Third, we detect a disconnection between the number of followers of some senators and their influence inside and outside the Senate. We explain the disconnection with a temporal popularity analysis, correlating it with political events. Fourth, we establish a correlation between time and the influence of topics. Finally, we include the factor of geography as a concept to explain how the US Senate works.

In conclusion, this paper provides a simplified approach to analyzing social media in Politics while going beyond the current state of the art. Our findings improve the quality of view on the political landscape, how US citizens see their senators, and how the parties work together on Twitter. It can inform stakeholders, such as journalists or researchers, on important issues and trends. It is also helpful for guiding US politicians and parties on what topics they should cover to maximize their influence among voters and their career strategies.

2. Related Work

In 2018, Russell studied the polarization in the US Senate, hypothesizing that Republican party members were more likely to be polarized (2018). Newly in 2020, Russell studied how senators used Twitter to articulate their agenda (2020). Our study extends a much broader period and covers the senators' relations. Researchers have stated how race, ethnicity, and geography can be differential for election competition (Mccarty et al., 2019; Lee & Rogers, 2019). Other works focus mainly on the economic capacity of the voters and conclude that senators that take sides gain influence (Lax et al., 2019). However, they leveraged data from outside social media while we took a broader approach.

[*]Equal contribution [1]Illinois Institute of Technology and Universidad Politécnica de Madrid, Chicago, United States [2]Illinois Institute of Technology, Chicago, United States. Correspondence to: Miguel Cozar <mcozartramblin@hawk.iit.edu>, Carlos Munoz <cmunozlosa@hawk.iit.edu>, Kai Shu <kshu@iit.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Topic modeling is also another recurrent research goal. Jelodar *et al.* employed LDA applied to plenary speeches in the US Senate and a dataset of politically oriented users' tweets (2019). Ylä-Anttila *et al.* modeled Twitter debates on climate change (2022). Our approach includes senate-specific tweets and proposes a modern approach, BERTopic, that has overcome traditional topic modeling techniques. Sia *et al.* outlined a methodology for clustering word embeddings for unsupervised document analysis and concluded that it yielded a greater diversity within topics compared to LDA. Egger *et al.* performed a comparative study between topic modeling tools (2022), remarking on the versatility and support of embedding models of BERTopic, and stating the unreliability of LDA.

Regarding popularity metrics, De Vries *et al.* discussed how fan pages nurture relationships with customers (2012), stating the influence of likes in their publications. This previous work supports our model of retweets and likes as a metric for influence. Pancer and Poole studied the Twitter messages for Hilary Clinton and Donald Trump after the 2016 elections. Our approach is slightly different, focusing more on what matters senators discuss rather than how they say it (2016).

3. Methodology

This section covers gathering and representing the data, performing topic modeling, and analyzing.

3.1. US Senator Network Creation

We take a list of senators, their parties, and states by scrapping Wikipedia. After, we build the network using Networkx and Twitter's official API. We leverage a publicly available list of zip codes to assign a random zip code belonging to the senators' state, and Geoapify to obtain a pair of coordinates per zip code. We utilize Plotly and Mapbox for representing the graph. Note that the current structure of the Senate comprises 49 Democrats, 48 Republicans, and 3 Independent senators. Since the three independent senators caucus with the Democratic party, we include them as Democratic. We achieve a complete network of 100 geolocated nodes and 2,378 edges.

3.2. Tweet Corpus Creation

We gather a dataset of 238,645 tweets from March 2008 to March 2023. We also include in the dataset the publisher of the tweet and the number of likes and retweets. We use this dataset in combination with the one in the previous section to map each senator with their publications, likes, retweets, geographical coordinates, and other features. We have more recent tweets than old ones to represent the current landscape better. Another reason is due to the year of arrival at the Senate of some of the senators after 2008. The most represented year is 2022, with around 67K tweets, and the least represented is 2008, with 109.

3.3. Topic Modeling

After preprocessing, we feed *multi-qa-mpnet-base-dot-v1*, a pre-trained BERT model with the obtained data, obtaining 768-dimensional embedding vectors. We select this sentence transformer because it is the best performer in semantic search, according to sbert.net (Reimers & Gurevych, 2019).

For the topic modeling, we leverage BERTopic (Grootendorst, 2022). This algorithm reduces the dimensions of the embedding with uMap and performs clustering of the documents using HDBSCAN. Once the clusters are detected, with c-TF-IDF, the algorithm assigns the topics. With the previous information, we manually label each topic according to their highlighted words, obtaining a topic distribution. The algorithm leaves one topic to outliers we catalog as *Unclassified*.

3.4. Popularity Score

We compute a correlation matrix to determine which variables we should use to measure a topic's popularity. We detect that retweets and likes are highly correlated and, as per (de Vries et al., 2012), are a consistent measure of influence. However, a high number of tweets with a low number of interactions could fool the model. Therefore, we include a penalization factor as a quotient.

$$\text{Popularity}_{\text{topic}} = \frac{\sum_{i=1}^{n}(\text{retweets}_i + \text{likes}_i)}{n} \tag{1}$$

n being the number of tweets assigned to a topic.

We further develop the expression for measuring a senator's popularity. We include the number of the senator's followers as one extra parameter, leaving the equation as follows.

$$\text{Popularity}_{\text{sen}} = \frac{\sum_{i=1}^{n}(\text{retweets}_i + \text{likes}_i + 0.25 \cdot Followers_{\text{sen}})}{n} \tag{2}$$

n being the number of tweets published by the senator.

Hence, we use the first one to calculate the popularity of topics per senator by filtering the number of tweets by just the ones published by certain senators. The second one is used to analyze the popularity of senators as individuals.

With these two expressions, we compute all the necessary popularity metrics.

4. Results and Discussion

This section exposes the different results that we obtain. All data and results are publicly available on our Kaggle repository. We also have a public repository with the code of the project on Github.

4.1. Network Metrics

The in-degree distribution illustrates how many senator-followers a senator has. The average in-degree is 23.78. Democratic senator Cory Booker from New Jersey is the most followed senator, with 50 senator followers, whereas Republican Lisa Murkowski from Alaska has only two colleagues following her.

Table 1 illustrates how the in-degrees do not depend on the global amount of followers. The power of social media does not always spread to the US Senate, and some of the most followed senators, such as Mitt Romney or Ted Cruz, have quite an in-degree average. In contrast, Democrats seem to be lobbying more, with examples such as Chuck Schumer or Booker. Finally, if we separate between parties, the in-degree average for Republicans is 20.2, whereas the mean in-degree for Democrats is 27.21. These numbers indicate that the Democratic party network is more cohesive and more likely to perform as a team.

Table 1. List of the ten most followed senators with their internal follows

Senator	In-degree	Followers
Bernie Sanders	38	12.5M
Elizabeth Warren	38	7.04M
Ted Cruz	24	5.93M
Cory Booker	50	4.83M
Rand Paul	39	4.61M
Marco Rubio	41	4.52M
Chuck Schumer	48	3.45M
Lindsey Graham	36	2.21M
Mitch McConnell	34	2.18M
Mitt Romney	23	2.05M

4.2. Topic Modeling

We configure BERTopic with a minimum cluster size of 150 to avoid outliers while generating a reasonable amount of clusters. We obtained 167 clusters and manually labeled them, reviewing the most relevant words of each topic, obtaining a total amount of 13 main topics: Climate Change (7.7K), Economy (9K), Education (4.3K), Elections (5K), International Politics (9.5K), Justice (5.6K), National Issues (42K, including tweets of senators speaking about their state issues), Public Health (11.5K, including tweets about Covid19 pandemics), Religion (1K), Security (15K), Social Issues (16.5K), Technology (3.1K) and Others (20.5K, includes tweets that do not fit in the previous categories). We leave the 88K outliers as *Unclassified*.

4.3. Popularity Metrics

The analysis of the popularity metrics is overwhelming, considering the number of features involved. We select the most important figures to understand the critical points of the work.

Figure 1 shows the topic distribution. The apparent difference in popularity for several topics between both parties is surprising. Democrats are three times more relevant than Republicans in Economy and significantly more popular in electoral tweets. Republicans are four times more popular in Technology. A possible explanation is the increased debate about users' rights on social media caused by Former President Trump's ban from Twitter. The Democratic Party shows significantly more popularity. **Democrats have 45% more average popularity rate than Republicans**, according to our measurements based on the overall popularity, computed through the weighted average popularity of all topics.

Figure 2 compares popularity and the followers of the 32 most popular senators. We normalize the popularity score by dividing the popularity score by the maximum score, so the maximum score is 1. We selected only senators overcoming a 0.05 popularity score.

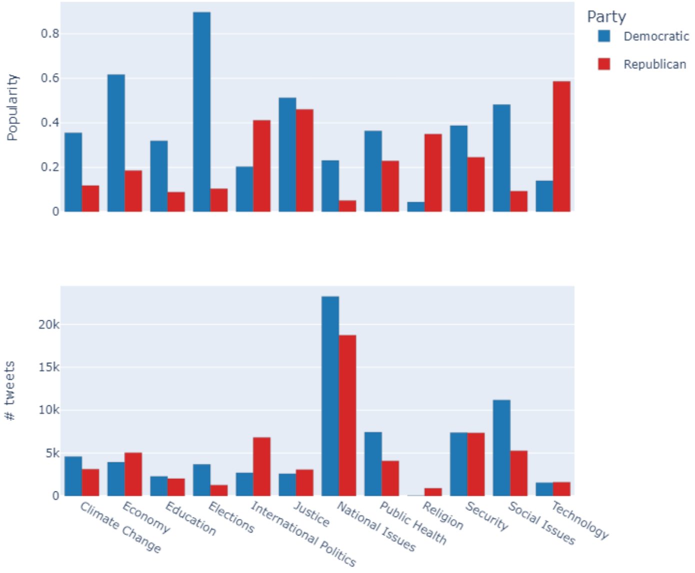

Figure 1. Popularity value versus the number of tweets published for Democrats and Republicans (shared x-axis).

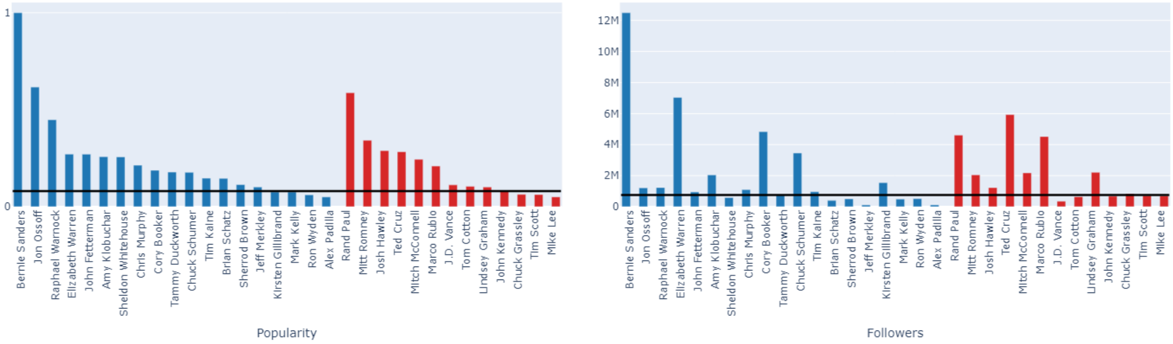

Figure 2. Popularity value for the 32 most popular senators on the left. Number of followers for the 32 most popular senators on the right. The black line represents the mean value in each case.

The figure exposes that some senators' tweets are significantly popular, but their number of followers does not align with it. We mentioned Booker and Schumer in Section 4.1 as two of the most inner-followed, but they do not reach those broad audiences among regular users.

On the Republican side, Mitt Romney has the party's second most popular metric, with fewer followers than others like Ted Cruz or Marco Rubio. Precisely Ted Cruz, the most followed Republican, cannot generate that much engagement. Ted Cruz also gets limited support from the Senate (his in-degree is one of the lowest in Table 1), another symptom of the Republican party's lack of cohesion, which the Democrats do not show.

Figure 3 shows a comparison between senators Rand Paul (Republican) and Jon Ossoff (Democrat) popularity metrics over the last three years. We selected this window because it is the one with more tweets gathered. It is a representative example of how a senator with average followers can compete in popularity with a more mainstream senator. While Senator Paul remains constant over time, we appreciate that Senator Ossoff has two prominent peaks triggering his levels of popularity.

The first peak is the biggest, comprising November and December 2020, where his most popular tweets are electoral campaign and tweets referencing the Capitol attack. Some examples of popular tweets are: *Today's insurrectionist attack on the U.S. Capitol was incited by Trump's poisonous lies; flagrant assault on our Constitution. The GOP must discard and disavow Trump once and for all, end its attacks on the electoral process; commit fully to the peaceful transfer of power.* or *Today, as I was sworn in, I held in my jacket pocket copies of the ships' manifests recorded at Ellis Island when my Great Grandfather Israel arrived in 1911 and my Great Grandmother Annie arrived in 1913. A century later, their great grandson was elected to the U.S. Senate*

The second peak was on January 2022, when Jon Ossoff led new legislation to ban Congress members from trading with stocks (oss). His most popular tweet in this period is *Tonight I introduced legislation to ban Members of Congress (and our spouses) from trading stocks. 3/4 of Americans agree!*. Again, we can see a correlation between events happening in the Senate and tweets published.

On the other hand, Rand Paul's popularity was constant, we appreciate an increase in 2023, where he overcomes Jon Ossoff, but we do not find a correlation between events and popularity.

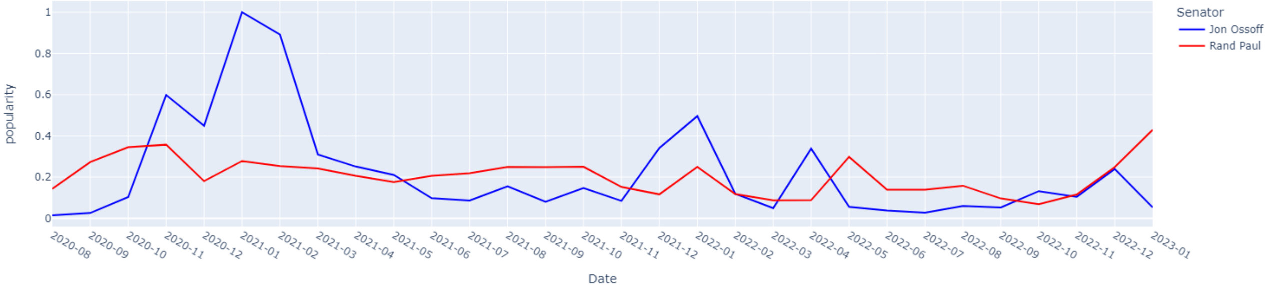

Figure 3. Popularity of senator Rand Paul (Republican) versus Jon Ossoff (Democrat) over time

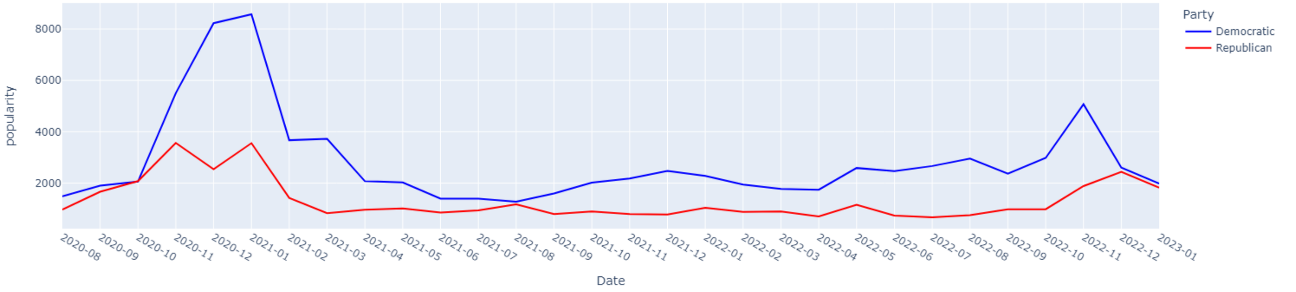

Figure 4. Popularity evolution over time for both parties

Figure 4 validates the argument that political events dramatically affect popularity, de Democrat party has the lead on popularity, and one key factor was the 2021 national election and the posterior Capitol assault. We observe a tendency for equality before and a constant difference after, which remains in 2023. If senators promote their work within the Senate heavily on Twitter, especially leading new initiatives, the public will likely support their actions.

Finally, Figure 5 aims to include geography in the previous factors. In addition, it facilitates the study of the influence between senators and their relations. The East Coast hosts the most popular Democrats. On the contrary, the Republicans' most popular senators gather in the Midwest Central states such as Texas or Kentucky.

Democrats have historically dominated both US coasts, but the West Coast's popularity is secondary in the Senate. Only Brian Schatz from Hawaii (technically West Coast) entered the top.

Another remarkable aspect is that the Democratic network appears to have significantly more interconnections than the Republican. Blue nodes are almost entirely fully connected, and red nodes have few connections.

5. Conclusion

We gathered data from different sources, focusing on Twitter. We employed unsupervised learning to perform a topic modeling of tweets, minimizing human effort. Finally, we used all the generated data to analyze the US Senate paradigm deeply.

We uncovered a significant difference between the two parties in the Senate, and to strengthen our findings, we introduced a topic, geographical and temporal analysis.

From the results, we conclude that the Democratic party seems to be leading the race of Twitter, with a more compact network and the majority of the most influential senators. In addition to their dominant popularity, they add a strong connection between nodes. Note that **a strong interconnection is vital for information diffusion**. Democratic senators may leverage this power to spread their message and continue gaining popularity, transferring power to less popular peers. On the other hand, Republicans lag in terms of network cohesion. Geographical factors also appear in the US Senate, with a general dominance of the East Coast over the West Coast for the Democratic party and a solid base in the Mid West for the Republicans.

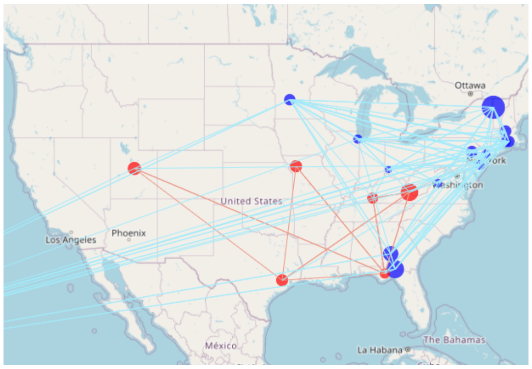

Figure 5. US map territory with the top 20 most popular senators and their connections.

Finally, our popularity metric shows that few senators receive high levels of popularity while the vast majority have poor ratings. However, senators with significantly higher levels of popularity are bound to have a more significant potential for influence. This last statement is particularly relevant because social network recommendation algorithms suggest similar tweets and users based on popularity measures. There is a particular dis-concordance between the popularity and engagement of some senators and their number of followers.

This last concept is fatherly explained by adding a temporal dimension. We studied how events can affect senators' popularity. We detected a high correlation between social media and events, that allows some senators to gain engagement for windows of time and average a global popularity higher than other senators with more followers We proved it with the particular case of Senator Jon Ossoff.

6. Limitations

During the development of the paper, we identified some limitations essential to consider for readers. The performance metric can improve its accuracy by adding more factors. For instance, our study skips the number of mentions, commentaries, and images in senators' publications. Twitter's API has also limited the work. The generated dataset includes a significant percentage of the tweets the Senate published during this time window. The usage of BERTopic as a topic model led to the appearance of some outliers. Techniques like LDA or NMF (Egger & Yu, 2022) can reduce the number of outliers. However, we decided to prioritize using transformers and density-based clustering because of its novel approach and proven reliability. One last limitation of BERTopic relates to the algorithm's stochastic components, which may affect the reproducibility of the topic modeling stage.

7. Future Work

For future work, we plan to include several lines of research. The most promising improvement could be the inclusion of multi-modal models, allowing to input both text and images.

Secondly, one exciting line of research is the implementation of graph neural networks. Graph neural networks (GNNs) are machine learning models that process and analyze structured data represented as graphs. Taking the Senate as a graph, we can leverage models to approximate discourses and state similarities between senators within the network structure and their discourse. For instance, we could measure how close popular senators' discourses are between them.

Including community detection algorithms could also help better understand the US Senate's structure. We want to study the popularity distribution among the senators through time and focus on how senators influence each other.

Depending on the trending topics of each. Sentiment analysis of tweets is something that could add value. Finally, a more state-focused analysis of some candidates could shed light on our *National Issues* topic.

Finally, the line of research that we are currently following is the study of toxicity discourse within the Senate. We are employing Debiasing Adversarial Models to analyze the number of toxic tweets published by senators, their distribution, and their temporal evolution.

8. Acknowledgements

This work was possible thanks to the collaboration between the Illinois Institute of Technology and Universidad Politécnica de Madrid. We thank them for providing the necessary resources to develop the research work.

We also thank the organizations providing the open-source data necessary for our analysis: Twitter with their API for building the network and retrieving the tweets, Wikipedia, Geoapify, Mapbox and the individual Stephen Pike for his zipcode dataset.

References

Sens. ossoff, kelly introduce bill banning stock trading by members of congress - u.s. senator for georgia jon ossoff. URL https://www.ossoff.senate.gov/press-releases/sens-ossoff-kelly-introduce-bill-banning-stock-trading-by-members-of-congress/.

de Vries, L., Gensler, S., and Leeflang, P. S. Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of Interactive Marketing*, 26(2):83–91, 2012. ISSN 1094-9968. doi: https://doi.org/10.1016/j.intmar.2012.01.003. URL https://www.sciencedirect.com/science/article/pii/S1094996812000060.

Egger, R. and Yu, J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, 7:886498, 5 2022. ISSN 22977775. doi: 10.3389/FSOC.2022.886498. URL /pmc/articles/PMC9120935//pmc/articles/PMC9120935/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/.

Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78: 15169–15211, 6 2019. ISSN 15737721. doi: 10.1007/S11042-018-6894-4/TABLES/11. URL https://link.springer.com/article/10.1007/s11042-018-6894-4.

Lax, J. R., Phillips, J. H., and Zelizer, A. The party or the purse? unequal representation in the us senate. *American Political Science Review*, 113:917–940, 11 2019. ISSN 0003-0554. doi: 10.1017/S0003055419000315. URL https://www.

cambridge.org/core/journals/american-political-science-review/article/abs/party-or-the-purse-unequal-representation-in-the-us-senate/286BFEAA039374759DE14D782A0BB8DD.

Lee, D. W. and Rogers, M. Measuring geographic distribution for political research. *Political Analysis*, 27:263–280, 2019. ISSN 1047-1987. doi: 10.1017/PAN.2019.14. URL https://www.cambridge.org/core/journals/political-analysis/article/abs/measuring-geographic-distribution-for-political-research/3EFD1595DB544E76FF26CFE9416AF032.

Mccarty, N., Rodden, J., Shor, B., Tausanovitch, C., and Warshaw, C. Geography, uncertainty, and polarization. *Political Science Research and Methods*, 7:775–794, 10 2019. ISSN 2049-8470. doi: 10.1017/PSRM.2018.12. URL https://www.cambridge.org/core/journals/political-science-research-and-methods/article/geography-uncertainty-and-polarization/4BF8B48947D113D85662123D207D3418.

Pancer, E. and Poole, M. The popularity and virality of political social media: hashtags, mentions, and links predict likes and retweets of 2016 u.s. presidential nominees' tweets. *Social Influence*, 11(4):259–270, 2016. doi: 10.1080/15534510.2016.1265582.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 3982–3992, 8 2019. doi: 10.18653/v1/d19-1410. URL https://arxiv.org/abs/1908.10084v1.

Russell, A. U.s. senators on twitter: Asymmetric party rhetoric in 140 characters. *American Politics Research*, 46:695–723, 2018. doi: 10.1177/1532673X17715619. URL https://doi.org/10.1177/1532673X17715619.

Russell, A. Senate representation on twitter: National policy reputations for constituent communication. 2020. doi: 10.1111/ssqu.12904. URL https://onlinelibrary.wiley.com/doi/10.1111/ssqu.12904.

Ylä-Anttila, T., Eranti, V., and Kukkonen, A. Topic modeling for frame analysis: A study of media debates on climate change in india and usa. *Global Media and Communication*, 2022:91–112, 2022. doi: 10.1177/17427665211023984. URL https://doi.org/10.1177/17427665211023984.

9. Author Contributions

- Miguel Cozar is a LatinXAI member since December 2022. He writes for the LXAI blog. In this paper, he was in charge of the data gathering of both the network and the tweet dataset. He performed the topic modeling and analyzed the results. He was also in charge of writing the manuscript and designing the poster.
- Carlos Munoz is also a LatinXAI member. He designed the popularity metrics and performed an analysis and interpretation of the results. He was in charge of writing the manuscript and designing the poster.
- Kai Shu is a Gladwin Development Chair Assistant Professor in the Department of Computer Science at Illinois Institute of Technology since Fall 2020. He supervised and mentored the entire project, as well as helped in the development of new research lines.