Abstract

As political attitudes have diverged ideologically in the United States, political speech has diverged linguistically. The ever-widening polarization between the US political parties is accelerated by an erosion of mutual understanding between them. We aim to make these communities more comprehensible to each other with a framework that probes community-specific responses to the same survey questions using community language models (COMMUNITYLM). In our framework we identify committed partisan members for each community on Twitter and fine-tune LMs on the tweets authored by them. We then assess the worldviews of the two groups using prompt-based probing of their corresponding LMs, with prompts that elicit opinions about public figures and groups surveyed by the American National Election Studies (ANES) 2020 Exploratory Testing Survey. We compare the responses generated by the LMs to the ANES survey results, and find a level of alignment that greatly exceeds several baseline methods. Our work aims to show that we can use community LMs to query the worldview of any group of people given a sufficiently large sample of their social media discussions or media diet.

1 Introduction

Political polarization is a prominent component of politics in the United States (Poole and Rosenthal, 1984; McCarty et al., 2016; Heltzel and Laurin, 2020). Previous studies have shown growing polarization in social media (Bail et al., 2018; Demszky et al., 2019; Darwish, 2019) and substantial partisan and ideological differences in media diet (Bozell, 2004; Gil de Zúñiga et al., 2012; Hyun and Moon, 2016). Li et al. (2017) show that partisanship makes reliable predictions about an individual’s word understanding. R. KhudaBukhsh et al. (2021) used modern machine-translation techniques to demonstrate that the left and right communities use English words differently. Milbauer et al. (2021) extended the method to uncover worldview and ideological differences between 32 Reddit communities. These studies are word-level analyses based on Word2vec word embeddings, and none of them use pre-trained language models.

Prompting is a standard technique to make pre-trained language models generate texts conditioned on prompts. Recent work has shown that, through prompt engineering, pre-trained language models can achieve good zero-shot performance on NLP tasks from sentiment classification to reading comprehension (Radford et al., 2019; Brown et al., 2020) and mine factual or commonsense knowledge (Petroni et al., 2019; Davison et al., 2019; Jiang et al., 2021; Talmor et al., 2020). Through prompting, Palakodety et al. (2020) used a fine-tuned BERT (Devlin et al., 2019) model with fill-in-the-blank cloze statements to mine insights and compare prediction differences between Indian regional and national YouTube news channels. Feldman et al. (2021) fine-tuned GPT-2 on COVID-19 tweet corpora to mine user opinions.

However, none of these studies fine-tune GPT-style language models on community data to probe community worldviews. In this work, we focus on Republican and Democratic Twitter communities and conduct a feasibility study using fine-tuned GPT-2 partisan language models to generate community responses and to predict community stance.
As exemplified in Table 1, we observe clear partisan differences. In this sociopolitically fragmented society, our motivation is to provide a simple and flexible interface for people to probe each other’s worldviews on topics of interest and to encourage constructive dialogue. We demonstrate through our experiments and analyses that the proposed method is a reliable tool to probe community opinions. The contribution of the work is as follows:

- We present a simple COMMUNITYLM framework based on GPT-2 language models to mine community insights by fine-tuning or training the model on community data. This study focuses on Democrat and Republican communities on Twitter but can be easily extended to probe insights from any community based on their public discourse or media diet.

- We use ANES questions as prompts and find that GPT-generated opinions are predictive of community stance towards public figures and groups. We experiment with 4 types of prompts and find that the fine-tuned COMMUNITYLM with an “X is the” prompt outperforms all the baselines (including pre-trained GPT-3 Curie) in predicting community stance.

- We analyze the errors made by community language models and demonstrate the capability of the models to probe community preferences towards public figures by ranking.

2 Partisan Twitter Data

We construct a Twitter dataset containing 4.7M tweets (100M word tokens) by Republican and Democrat communities respectively. We first sample 1M active U.S. Twitter users before and after the 2020 presidential election. We adapt the standard method (Volkova et al., 2014; Demszky et al., 2019) to estimate their political affiliation from the political accounts they follow. Specifically, we update the list of Twitter handles of US politicians from Demszky et al. (2019) by adding current federal officeholders and governors from Ballotpedia. The final list has 457 Republican and 473 Democratic politician Twitter handles. To identify committed partisan users, we adopt the following rules: a user is labeled as a Democrat if they followed no fewer than 6 Democratic politicians and no Republican politician from the list in February 2022, whereas a person is labeled as a Republican if they followed no fewer than 2 Republican politicians and no Democratic politicians. We choose these thresholds because there are 69% Democratic users and 26% Republican users on Twitter (2.65:1). This step predicts 182,788 Democratic-leaning and 72,186 Republican-leaning users (2.53:1).

2 Tweet Pre-processing. We use the tweet tokenizer from Nguyen et al. (2020) to process all the data. This tokenizer converts user mentions and web/url links into special tokens @USER and HTTPURL. We delete HTTPURL from the tweets because it does not contain useful community information. We do not lower the case but filter out tweets with less than 10 tokens, producing 7,554,409 Democratic and 4,759,441 Republican tweets. We randomly sample from Democratic tweets to ensure both partisan communities have the same number of 4,759,441 tweets for training language models to ensure a fair community model comparison.

U.S. Twitter User Sampling. We first sample a subset of active Twitter users from the “decahose”, Twitter’s 10% sample of tweets. We define active U.S. users as those who posted at least 10 original tweets before and after the 2020 presidential election period (2020-07-01 to 2021-06-31). We then use Litecoder to extract user locations from their profile location strings and filter out users not based in the U.S. We construct the follow graph of the resulting set of 1,074,650 Twitter users.

Partisan Assignment. We follow previous studies (Volkova et al., 2014; Demszky et al., 2019) to estimate the party affiliation of Twitter users from the political accounts they follow. Specifically, we update the list of Twitter handles of US politicians from Demszky et al. (2019) by adding current federal officeholders and governors from Ballotpedia. The final list has 457 Republican and 473 Democratic politician Twitter handles. To identify committed partisan users, we adopt the following rules: a user is labeled as a Democrat if they followed no fewer than 6 Democratic politicians and no Republican politician from the list in February 2022, whereas a person is labeled as a Republican if they followed no fewer than 2 Republican politicians and no Democratic politicians. We choose these thresholds because there are 69% Democratic users and 26% Republican users on Twitter (2.65:1). This step predicts 182,788 Democratic-leaning and 72,186 Republican-leaning users (2.53:1).
3 Framework

We present a simple COMMUNITYLM framework which adapts GPT-style language models to mine community insights. This framework consists of four steps: (1) fine-tune or train GPT language models on community data, (2) design prompts based on survey questions, (3) generate community responses with language models, (4) aggregate community stance based on responses.

3.1 Model Training and Fine-tuning

We pick GPT-2 with 124M parameters and experiment with two training strategies on the partisan community data: (1) fine-tune a pre-trained GPT-2 model, (2) train a GPT-2 model from scratch. For both settings, we adopt training epoch 10 and batch size 24 on Nvidia GeForce GTX 1080 12GB. The greedy decoding is used for GPT-2. Otherwise, we use the default training parameters. The pre-trained GPT-2 model was released in February 2019, trained on data that cuts off at the end of 2017. We also use GPT-3 Curie as one of our baselines, which used training data up to Oct 2019. Therefore, neither pre-trained model used any data beyond the the start date of the ANES survey.

We adopt 10 epochs because GPT-2 was not pre-trained on the Twitter domain and had a steady loss decrease across all epochs. We checked all synthetic tweets (lowercased) generated by the fine-tuned GPT-2 with “X is/are the”. The percentages of synthetic tweets appearing in training data are 64.93% and 69.56% for Republican and Democratic models. For researchers who want to adapt our approach with a lower repetition rate, we suggest moving away from the greedy decoding algorithm and reducing the epoch number.

3.2 Prompt Design

We design discrete prompts based on survey questions to probe community insights towards public figures and groups. The American National Election Studies (ANES) are academically-run national surveys of voters in the United States. We adopt the ANES 2020 Exploratory Testing Survey, conducted between April 10, 2020 and April 18, 2020 on 3,080 adult citizens from across the United States, because this survey captures recent political changes in the US. We adapt all 30 questions from “FEELING THERMOMETERS” section of the ANES survey, which asks participants to rate people or groups from 0 (“not favorable”) to 100 (“favorable”) with the question “How would you rate ____?” The questions cover 30 items in two categories (a) 16 people: Donald Trump, Barack Obama, Joe Biden, Elizabeth Warren, Bernie Sanders, Pete Buttigieg, Kamala Harris, Amy Klobuchar, Mike Pence, Andrew Yang, Nancy Pelosi, Marco Rubio, Alexandria Ocasio-Cortez, Nikki Haley, Clarence Thomas, Dr. Anthony Fauci, (b) 14 groups: blacks, whites, Hispanics, Asians, illegal immigrants, feminists, the #MeToo movement, transgender people, socialists, capitalists, big business, labor unions, the Republican Party, the Democratic Party. For each item “X”, we experiment with four types of discrete prompts: (1) “X”, (2) “X is/are”, (3) “X is/are a”, (4) “X is/are the”. These names are copied from the survey verbatim except for “whites”, “blacks”, “Hispanics”, and “Asians” because “whites” and “blacks” also refer to other named entities such as “Blacks Clothing Company” and “Whites TV shows”. Instead, we translate the names of these four groups into “White people”, “Black people”, “Hispanic people”, “Asian people”. We also provide the count number of each item in Appendix A.

3.3 Community Response Generation

For each community, we use the corresponding language model to generate 1000 responses given the prompts. We use Hugging Face’s TextGenerationPipeline and apply the same decoding strategy by setting do_sample to true, temperature to 1.0, and max_length to 50. If one response contains multiple sentences, we use the first line in the response and remove the remaining tokens, because a response with multiple sentences may have mixed sentiments, making it hard to identify the overall sentiment.

3.4 Community Stance Aggregation

After response generation, we save them locally and compute the community stance for each prompt by aggregating the sentiment of the synthetic responses. Specifically, we use the state-of-the-art Twitter sentiment classifier “cardiffnlp/twitter-roberta-base-sentiment-latest” (Barbieri et al., 2020; Loureiro et al., 2022) on the SemEval-2017 benchmark (Rosenthal et al., 2017) to classify each
generated response into -1 (“Negative”), 0 (“Neutral”), and 1 (“Positive”). We take the average sentiment of the generated responses as the community’s stance score towards the person or group. We also show the results of a popular lexicon-based sentiment classifier VADER in Appendix C.

4 Evaluation

Task Formulation. The ANES survey has self-reported party affiliation from participants. We use responses from Republican and Democratic participants and calculate their average ratings towards each of 30 items (persons and groups). These average ratings are provided in Appendix B. If the average rating of Republican participants is higher than that of Democratic participants toward one item (e.g., Joe Biden), it is labeled as “R”. Otherwise, the item is labeled as “D”. 70% items are labeled “D” and 30% “R”. The 9 items with “R” label are Donald Trump, Mike Pence, Marco Rubio, Nikki Haley, Clarence Thomas, whites, capitalists, big business, and the Republican Party. The task asks a model to predict which community is more favorable towards an item. To address the data imbalance, we prefer weighted F1 to accuracy as a measure of model performance.

Baselines. We evaluate the performance of trained and fine-tuned COMMUNITYLM (GPT-2) against 4 baselines. The first baseline is Frequency Model which counts the frequency of an item’s name in each community’s data and classifies the community with higher word frequency to be the label. The second baseline is Keyword Retrieval which uses keywords to retrieve tweets containing the keywords from each community’s data, computes the average community stance, and selects the community with a higher stance score. Keyword Retrieval (full) means using the full names as keywords and Keyword Retrieval (surname) means using the surname of people. The third and fourth baselines use pre-trained GPT-2 and pre-trained GPT-3 Curie respectively. “[CONTEXT]” is a preceding context “As a Democrat/Republican, I think”, which is concatenated with the prompts to generate partisan responses on each item. We compute the average community stance on 1000 synthetic responses and pick the community with a higher stance score. It is noted that we also fine-tune or train GPT-2 on the aggregate partisan tweets and show their results in Appendix D.

Overall Performance. First, we observe that fine-tuned COMMUNITYLM with “X is the” prompt achieves the best performance in both accuracy (97.33%) and weighted F1-score (97.29%) on the task. The same model’s performance is sensitive to the prompt design and the longest prompt out of the four seems to work the best. “X” alone is bad, because it will result in many responses like “X @USER”, “X???” “X”, which are common Twitter posts and are too short to interpret their attitudes. Second, fine-tuned COMMUNITYLM outperforms trained COMMUNITYLM from scratch. It indicates that pre-training GPT-2 is helpful, probably because pre-training injecting the general knowledge about the named entities into GPT-2. Third, we find that Keyword Retrieval (surname) is a strong baseline in both accuracy (93.33%) and F1 (93.33%), but its performance is also sensitive to the selection of keywords. As we see, the weighted F1 performance of Keyword Retrieval (full), which uses a strict full name matching (e.g., “Joe Biden”), drops to 87.00%. In contrast, language models are able to learn the associations between different names for the same person and generalize without worrying about name forms. Last, fine-tuned COMMUNITYLM outperforms pre-trained GPT-2 and GPT-3 baselines. It is worth noting that the performance of pre-trained GPT-3 Curie is consistently better than pre-trained GPT-2. GPT-3 with the “X is/are” prompt achieves the same score as the Keyword Retrieval (surname) baseline.

Error Analysis. The rule-based Keyword Retrieval (surname) baseline misses “illegal immigrants” and “big business”. The fine-tuned COMMUNITYLM with “X is/are the” misses “White people”. The pre-trained GPT-3 with “X is/are” prompt misses “Dr. Anthony Fauci” and “Asian people’. It is interesting because the top 5 items with the closest average rating gap between ANES partisan participants are Asian people (5.5%), White people (5.9%), Hispanic people (7.7%), Dr. Anthony Fauci (8.4%), and Black people (9.7%).

Ranking Public Figures. We use the average community stance scores computed on the generated tweets from the fine-tuned COMMUNITYLM model to rank 16 public figures for each community, hoping to understand how they perceive these people. In Figure 1, we observe that Republican politicians are rated poorly by the Democratic model and vice versa. Overall, the ratings from the Republican model are more negative than the Democratic model. Interestingly, we find that An-
| Model                                | Prompt                               | Accuracy   | Weighted F1  |
|--------------------------------------|--------------------------------------|------------|--------------|
| Frequency Model                      | —                                    | 53.33      | 54.50        |
| Keyword Retrieval (Full)             | —                                    | 86.67      | 87.00        |
| Keyword Retrieval (Surname)          | —                                    | 93.33      | 93.33        |
| Pre-trained GPT-2                    | “[CONTEXT] + X"                      | 74.00±2.79 | 66.52±5.56   |
| Pre-trained GPT-2                    | “[CONTEXT] + X is/are”               | 72.00±1.83 | 64.63±2.35   |
| Pre-trained GPT-2                    | “[CONTEXT] + X is/are a”             | 75.33±1.83 | 68.47±3.35   |
| Pre-trained GPT-2                    | “[CONTEXT] + X is/are the”           | 77.33±2.79 | 74.71±3.22   |
| Pre-trained GPT-3 Curie              | “[CONTEXT] + X”                      | 83.33      | 83.88        |
| Pre-trained GPT-3 Curie              | “[CONTEXT] + X is/are”               | 93.33      | 93.50        |
| Pre-trained GPT-3 Curie              | “[CONTEXT] + X is/are a”             | 83.33      | 83.88        |
| Pre-trained GPT-3 Curie              | “[CONTEXT] + X is/are the”           | 83.33      | 84.02        |
| Trained COMMUNITY LM                 | “X”                                  | 90.00±0.00 | 89.63±0.27   |
| Trained COMMUNITY LM                 | “X is/are”                           | 90.00±0.00 | 89.82±0.00   |
| Trained COMMUNITY LM                 | “X is/are a”                         | 86.00±1.49 | 86.25±1.50   |
| Trained COMMUNITY LM                 | “X is/are the”                       | 90.67±2.79 | 90.49±2.68   |
| Fine-tuned COMMUNITY LM              | “X”                                  | 84.67±2.98 | 84.46±3.18   |
| Fine-tuned COMMUNITY LM              | “X is/are”                           | 96.00±2.79 | 96.00±2.79   |
| Fine-tuned COMMUNITY LM              | “X is/are a”                         | 91.33±1.83 | 90.83±2.05   |
| Fine-tuned COMMUNITY LM              | “X is/are the”                       | 97.33±1.49 | 97.29±1.52   |

Table 2: Performance of different approaches in accuracy to predict which community is more favorable towards 30 persons or groups from the ANES survey. Approaches based on GPT-2 are repeated five times to compute the average and standard deviation. GPT-3 is only run once for cost concern. Frequency Model and Keyword Retrieval methods are deterministic. The weighted average F1 is used because of data imbalance.

drew Yang is rated quite highly by both models, likely because of the sampling bias of Twitter. It is noted that “Andrew Yang” is also ranked 1st by the Democrat community and 3rd by the Republican community with the retrieval approach.

![Diagram](image)

Figure 1: Left and right rankings of 16 public figures by their average stance scores calculated on synthetic tweets from their fine-tuned COMMUNITY LM models.

5 Conclusion

In this paper, we present a simple COMMUNITY LM framework to evaluate the viability of fine-tuned GPT-2 community language models in mining community insights in the context of political polarization between Republicans and Democrats. We adopt ANES survey questions and experiment with four types of prompts to generate community responses through GPT-2, showing that generated opinions are predictive about which community is more favorable towards selected public figures and groups. Our results show that fine-tuned COMMUNITY LM (GPT-2) outperforms the baseline methods. We analyze the model errors and run qualitative analyses to demonstrate that GPT-2 community language models can be used to rank public figures and probe word choices.

There are a few limitations in the current approach. First, language models can synthesize unreliable responses. Structured knowledge (Wang et al., 2021; Yasunaga et al., 2021) can be used to reduce nonsensical or unfaithful generation. Therefore, it is important that we use statistical patterns rather than individual synthesized tweets to draw conclusions (Feldman et al., 2021). Second, language models are shown to be sensitive to prompt design in our experiments and are also vulnerable to negation and misprimed probes (Kassner and Schütze, 2020). In the future, we plan to develop a systematic approach to design effective prompts and evaluate the robustness of COMMUNITY LM. Third, we focus on the classic red and blue polarization and do not consider a more fine-grained segmentation of U.S. politics. We hope to extend this work to study multiple sociopolitical communities in America and surface their unheard voices.
Ethical Considerations

We propose a general framework to probe community insights and observe differences between the Democratic and Republican communities on Twitter. While we do not discuss how to react to these findings, the intention of our research is encourage people to escape from their echo chambers, hear voices from other communities, and engage in constructive communication. One reasonable ethical concern is that by using a language model to predict community opinions, instead of asking individuals from the community directly, don’t we risk erasing individual voices? To that concern we would like to emphasize that our model is no substitute for deeper engagement with a community; as discussed in the limitation paragraph, the language model is just an entry point for understanding a community’s perspective. It serves to synthesize the points expressed by the speakers in the training data more effectively than we know how to do by hand. Any automated or semi-automated prediction system risks misinterpreting or “erasing” an expressed opinion, and we show in our work that the simpler methods of doing so are more error-prone, and hence measurably more unfair than the proposed approach in the paper.

Acknowledgements

We would like to thank anonymous reviewers for their helpful comments on our paper. We also want to thank Belén Carolina Saldías Fuentes, Will Brannon, Suyash Fulay, Wonjune Kang, Hope Schroeder, and Shayne Longpre for their discussion and feedback in the early stages of the project.

References

Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. Proceedings of the National Academy of Sciences, 115(37):9216–9221.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1644–1650, Online. Association for Computational Linguistics.

L Brent Bozell. 2004. Weapons of mass distortion: The coming meltdown of the liberal media. Crown Forum.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dharwad, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Kareem Darwish. 2019. Quantifying polarization on twitter: the Kavanaugh nomination. In International conference on social informatics, pages 188–201. Springer.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1173–1178, Hong Kong, China. Association for Computational Linguistics.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philip Feldman, Sim Tiwari, Charissa SL Cheah, James R Foulds, and Shimei Pan. 2021. Analyzing covid-19 tweets with transformer-based language models. arXiv preprint arXiv:2104.10259.

Homero Gil de Zúñiga, Teresa Correa, and Sebastian Valenzuela. 2012. Selective exposure to cable news and immigration in the US: The relationship between fox news, cnn, and attitudes toward Mexican immigrants. Journal of Broadcasting & Electronic Media, 56(4):597–615.
Gordon Heltzel and Kristin Laurin. 2020. Polarization in america: Two possible futures. Current opinion in behavioral sciences, 34:179–184.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media, volume 8, pages 216–225.

Ki Deuk Hyun and Soo Jung Moon. 2016. Agenda setting in the partisan tv news context: Attribute agenda setting and polarized evaluation of presidential candidates among viewers of nbc, cnn, and fox news. Journalism & Mass Communication Quarterly, 93(3):509–529.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. Transactions of the Association for Computational Linguistics, 9:962–977.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7811–7818, Online. Association for Computational Linguistics.

Ping Li, Benjamin Schloss, and D Jake Follmer. 2017. Speaking two “languages” in america: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently. Behavior research methods, 49(5):1668–1685.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online. Association for Computational Linguistics.

Nolan McCarty, Keith T Poole, and Howard Rosenthal. 2016. Polarized America: The dance of ideology and unequal riches. mit Press.

Jeremiah Milbauer, Adarsh Mathew, and James Evans. 2021. Aligning multidimensional worldviews and discovering ideological differences. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4832–4845, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 9–14, Online. Association for Computational Linguistics.

Shripani Palakodety, Ashiqu R KhudaBukhsh, and Jaime G Carbonell. 2020. Mining insights from large-scale corpora using fine-tuned language models. In ECAI 2020, pages 1890–1897. IOS Press.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Keith T Poole and Howard Rosenthal. 1984. The polarization of american politics. The journal of politics, 46(4):1061–1079.

Ashiqu R. KhudaBukhsh, Rupak Sarkar, Mark S. Kamlet, and Tom Mitchell. 2021. We don’t speak the same language: Interpreting polarization through machine translation. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17):14893–14901.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. Transactions of the Association for Computational Linguistics, 8:743–758.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 186–196, Baltimore, Maryland. Association for Computational Linguistics.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. Transactions of the Association for Computational Linguistics, 9:176–194.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 535–546, Online. Association for Computational Linguistics.
Appendix

A Keyword counts

The Keyword Retrieval baseline method retrieves tweets containing the keywords. Here we show the list of full and surname keywords and their counts in tables 3 and 4, respectively, for the Republican and Democratic tweets. For corresponding items between these two tables (e.g. “Asian people” in Table 3 to “Asian” in Table 4) there is a consistent increase in counts, especially for “Asian people” “Anthony Fauci”, “Hispanic people”, “labor unions”, “Clarence Thomas”. Some items in Table 4 might have too many counts. For example, we observe that “Trump” has 150,000+ counts in both partisan tweets, which can take a relatively long time for sentiment classifiers to run.

| Keyword                   | Question     | Dem | Repub |
|---------------------------|--------------|-----|-------|
| Asian people              | twitter      | 81  | 21    |
| Joe Biden                 | fbiden1      | 4177| 5377  |
| big business              | fbigbusiness | 321 | 291   |
| Black people              | fblack       | 3199| 1278  |
| Pete Buttigieg            | ftbuttigieg1 | 982 | 521   |
| capitalists               | ftcapitalists| 279 | 197   |
| Anthony Fauci             | ftfauci1     | 102 | 85    |
| feminists                 | ffemale1     | 351 | 628   |
| Nikki Haley               | fhillary1    | 169 | 274   |
| Kamala Harris             | fharris1     | 1711| 1450  |
| Hispanic people           | fhisp        | 28  | 21    |
| illegal immigrants        | ftillegal    | 251 | 2233  |
| Amy Klobuchar             | fklobuchar1  | 451 | 193   |
| labor unions              | flaborunions | 68 | 47    |
| the #MeToo movement       | ftmetoo      | 105 | 84    |
| Barack Obama              | ftobama1     | 684 | 929   |
| Alexandria Ocasio-Cortez  | ftocasio1    | 410 | 534   |
| Mike Pence                | fpence1      | 911 | 502   |
| the Republican Party      | ftrepublican | 1683| 838   |
| Martin Luther             | fmluther1    | 166 | 132   |
| Bernie Sanders            | ftsanders1   | 4572| 2711  |
| socialists                | fsocialists  | 627 | 2697  |
| Clarence Thomas           | ftthomas1    | 157 | 132   |
| transgender people        | fttranspl    | 165 | 38    |
| Donald Trump              | fttrump1     | 8501| 5479  |
| Elizabeth Warren          | ftwarren1    | 3112| 1897  |
| White people              | fwhite       | 3625| 1362  |
| Andrew Yang               | fyang1       | 585 | 489   |

Table 3: Counts of full names for each person and group in Republican and Democratic tweets.

| Keyword                   | Question     | Dem | Repub |
|---------------------------|--------------|-----|-------|
| Asian                     | ftsian       | 2961| 1917  |
| Biden                     | fbiden1      | 26539| 21438 |
| big business              | ftbigbusiness| 321 | 291   |
| Black people              | fblack       | 3199| 1278  |
| Buttigieg                 | ftbuttigieg1 | 3514| 1348  |
| capitalist                | ftcapitalists| 1393| 941   |
| Democratic Party          | ftdemocratic | 2677| 3611  |
| Fauci                     | ftfauci1     | 931 | 1219  |
| feminist                  | ffemale1     | 1686| 1470  |
| Haley                     | fhillary1    | 831 | 712   |
| Harris                    | ffharris1    | 6753| 5416  |
| Hispanic                  | fhisp        | 1173| 1693  |
| illegal immigrant         | ftillegal    | 312 | 2835  |
| Klobuchar                 | fklobuchar1  | 1958| 384   |
| labor union               | flaborunions | 110 | 47    |
| #MeToo movement           | ftmetoo      | 114 | 102   |
| Obama                     | ftobama1     | 15390| 55163 |
| Ocasio-Cortez             | ftocasio1    | 751 | 1792  |
| Pelosi                    | fpelosi1     | 5985| 15844 |
| Pence                     | fpence1      | 5818| 3021  |
| Republican Party          | ftrepublican | 2251| 1079  |
| Rubio                     | ftrubio1     | 508 | 502   |
| Sanders                   | ftsanders1   | 16091| 6568 |
| socialist                  | ftsocialists | 3182| 12666 |
| Thomas                    | ftthomas1    | 2136| 1248  |
| transgender               | fttranspl    | 1309| 1469  |
| Trump                     | fttrump1     | 188170| 150589|
| Warren                    | ftwarren1    | 16954| 6969 |
| White people              | fwhite       | 3625| 1362  |
| Yang                      | fyang1       | 4443| 1433  |

Table 4: Counts of surname names for each person and group in Republican and Democratic tweets.

“Black people” (9.7%). These items have very close ratings and we confirm in our error analysis that they are also challenging to the GPT-2 models. It is worth noting that the survey was done in early 2020 and at that time “Dr. Fauci” as a topic was not as divisive as it is today on Twitter.

C How well does the system perform using a lexicon-based sentiment classifier?

In the main paper, we use a state-of-the-art pre-trained BERT Twitter sentiment classifier to classify tweets. Some researchers may be concerned that neural sentiment models may learn and reflect biases in the training data and prefer using lexicon-based approaches. Therefore, we also use VADER (Hutto and Gilbert, 2014), a popular rule-based model for sentiment analysis of social media texts, and report the performance of our models with VADER in Table 6. Overall, we show that these models perform slightly worse with VADER, but we still see that fine-tuned COMMUNITYLM with “X is the” perform the best (93.33%) out of these models. This performance is on par with the Keyword Retrieval (surname) approach. We conjecture that using prompts like “X is the” creates many
### Table 5: Average rating of each item (person or group) from Republican and Democratic participants in the ANES survey.

| Item                  | Dem       | Repub     |
|-----------------------|-----------|-----------|
| Trump                 | 68.95     | 63.44     |
| Hispanic              | 71.22     | 77.16     |
| Black                 | 76.22     | 66.51     |
| Native American       | 31.52     | 43.01     |
| Capitalists           | 46.68     | 60.53     |
| Big Business          | 43.14     | 57.38     |
| Labor Unions          | 60.67     | 44.87     |
| Haley                 | 29.86     | 47.03     |
| Thomas                | 29.95     | 48.63     |
| Yang                  | 39.28     | 29.19     |
| Klobuchar             | 30.04     | 22.17     |
| Socialists            | 61.21     | 25.72     |
| Transgender People    | 63.22     | 35.06     |
| Socialists            | 54.00     | 24.11     |
| Illegal Immigrants    | 56.17     | 28.25     |
| Female                | 63.74     | 32.73     |
| Confederate           | 52.79     | 21.66     |
| Harris                | 52.12     | 18.63     |
| Sanders               | 50.00     | 16.49     |
| Warren                | 59.84     | 20.46     |
| Biden                 | 66.30     | 24.40     |
| Sanders               | 63.77     | 20.50     |
| Pelosi                | 61.16     | 10.10     |
| Democratic Party      | 71.24     | 24.34     |
| Pence                 | 24.09     | 71.12     |
| Republican Party      | 25.02     | 74.47     |
| Obama                 | 81.29     | 29.99     |
| Trump                 | 17.06     | 77.83     |

Table 6: Performance of different approaches with VADER in predicting which community is more favorable towards 30 persons or groups from the ANES survey. Experiments are repeated five times to compute the average and standard deviation.

| Model                  | Prompt                          | Accuracy     |
|------------------------|---------------------------------|--------------|
| Keyword Retrieval (Full) | —                               | 76.67        |
| Keyword Retrieval (Surname) | —                               | 93.55        |
| Pre-trained GPT-2       | “[CONTEXT] + X”                  | 76.67 ± 0.00 |
| Pre-trained GPT-2       | “[CONTEXT] + X is/are a”        | 76.67 ± 1.49 |
| Pre-trained GPT-2       | “[CONTEXT] + X is/are the”      | 76.67 ± 1.83 |
| Pre-trained GPT-2       | “[CONTEXT] + X is/are the”      | 74.67 ± 3.06 |
| Trained COMMUNITYLM     | “X”                             | 91.33 ± 3.80 |
| Trained COMMUNITYLM     | “X is/are a”                    | 94.67 ± 3.80 |
| Trained COMMUNITYLM     | “X is/are the”                  | 92.00 ± 3.80 |
| Fine-tuned COMMUNITYLM  | “X”                             | 91.33 ± 3.83 |
| Fine-tuned COMMUNITYLM  | “X is/are a”                    | 92.00 ± 3.83 |
| Fine-tuned COMMUNITYLM  | “X is/are the”                  | 93.33 ± 2.36 |

### D Is fine-tuning or training GPT-2 on combined Twitter data performing better than pre-trained GPT-2?

In the main paper, we use pre-trained GPT-2 and GPT-3 to predict the community stance. In addition, we also experimented with training and fine-tuning GPT-2 on the combined Twitter corpus (Republican and Democratic tweets). By contrast with COMMUNITYLM, which fine-tunes two GPT-2 models on partisan Twitter data, in this variant we only train or fine-tune one GPT-2 model on the aggregate of the partisan tweets. Similar to what we did in the pre-trained GPT-2 setting, we use [CONTEXT]+prompt to generate responses. The results are quite interesting, because the performance of the resulting models are worse than the pre-trained GPT-2, even below the majority baseline of 70%. We conjecture that this is because the combined data of partisan tweets neutralizes the sentiment that the models were supposed to learn towards the public figures and groups.

| Model                  | Prompt                          | Accuracy     |
|------------------------|---------------------------------|--------------|
| Trained GPT-2 (combined)| “[CONTEXT] + X”                  | 48.67 ± 8.37 |
| Trained GPT-2 (combined)| “[CONTEXT] + X is/are a”        | 50.67 ± 7.23 |
| Trained GPT-2 (combined)| “[CONTEXT] + X is/are the”      | 52.67 ± 2.79 |
| Trained GPT-2 (combined)| “[CONTEXT] + X is the”          | 55.33 ± 6.37 |
| Fine-tuned GPT-2 (combined)| “[CONTEXT] + X”                  | 53.33 ± 4.08 |
| Fine-tuned GPT-2 (combined)| “[CONTEXT] + X is/are a”        | 50.67 ± 6.35 |
| Fine-tuned GPT-2 (combined)| “[CONTEXT] + X is/are the”      | 52.67 ± 2.79 |
| Fine-tuned GPT-2 (combined)| “[CONTEXT] + X is the”          | 38.00 ± 8.37 |