Not Just Semantics: Social Distancing and COVID Discourse on Twitter

Austin van Loon
Stanford University
avanloon@stanford.edu

Sheridan A. Stewart
Stanford University
sastew@stanford.edu

Brandon Waldon
Stanford University
bwaldon@stanford.edu

Shrinidhi K Lakshmikanth
Uber ATG
klshrinidhi@uber.com

Ishan Shah
Stanford University
ijshah@stanford.edu

James Zou
Stanford University
jamesz@stanford.edu

Johannes Eichstaedt
Stanford University
eich@stanford.edu

Abstract

As we prepare for future waves of COVID-19, our ability to control the disease's spread will depend not only on our understanding of the biological processes that lead an individual to get sick from the virus, but also on our understanding of the psychological and sociological processes that lead individuals to either endorse or reject health behaviors which help reduce the spread of the virus. We argue that for different communities in the United States, the virus has taken on different meanings, and these disparate meanings shape communities' response to the virus. Using word embeddings, we demonstrate that in counties where residents social distanced less on average, the COVID discourse was more indicative of associations between the virus and the concepts of fraud, the political left, and more benign illnesses during the early, vital stages of the outbreak in the U.S. context. We also show that the different meanings the virus took on in different communities explains a substantial fraction of what we call the "Trump Gap", or the empirical tendency for more Trump-supporting counties to social distance less.

1 Introduction

The COVID-19 Pandemic is one of the most significant and devastating events in human history, with over 10 million confirmed cases and over 500,000 deaths globally (as of June 28th, 2020). Part of the reason this pandemic has been so deadly is that individuals vary in the degree to which they take the virus seriously and follow suggested governmental guidelines aimed at reducing the spread of the virus, including the practice of social distancing. The degree to which the virus will continue to spread—especially as we await a second wave of infections—will depend not only on our biological understanding of the virus, but also on our understanding of the social factors that explain the degree to which everyday individuals are willing to and do in fact engage in coordinated efforts to help slow the spread of the virus.

An especially interesting case of response to the virus is the United States, the country with the most deaths of any country in the world at the time of writing. What makes the variation in response to the pandemic in the United States particularly interesting is that this variation does not appear to be randomly distributed. Political party identification in particular seems to strongly predict the degree to which individuals endorse health behaviors recommended by authoritative organizations such as the CDC and WHO (Allcott et al. 2020; Painter and Qiu 2020). Rejecting these health behaviors, however, is not definitionally related to identifying as a Republican or supporting Donald Trump. We seek to build on this work by identifying, at least in part, the reason *why* we observe these behavioral disparities. We contend that in addition to the virus's objective properties, SARS-CoV-2 has taken on heterogeneous *socially constructed meanings*, which vary over the population and shape communities' response to the pandemic. Sociologists and other social scientists call this process the "social construction" of meaning (Berger and Luckmann 1966; Goldberg and Stein 2018). We further contend that this variation in how people perceive the virus and the larger context explains the link between political party identification and

these behavioral outcomes.

Research in sociology and anthropology (e.g. Geertz 1973; Rawlings and Childress 2019) finds that the meanings humans associate with different concepts tend to cluster in groups. One way that these meanings are clustered is by geography. Here, we leverage geographic variation in the content of discourse related to COVID-19 over the social media platform Twitter in order to explore the relationships between political identity, the meaning of COVID-19, and social distancing. To do this, we capture elements of what COVID-19 means to different populations using word embeddings, a technique that has been demonstrated to measure widely-held cognitive associations in groups (Bolukbasi et al. 2016; Caliskan et al. 2017; Garg et al. 2018; Kozlowski et al. 2019). This approach allows us to pick up on the disparate ways COVID-19 is interpreted by the residents of different counties. Further, we use Google Mobility Report data[1] to demonstrate that these disparate meanings both correlate with social distancing behavior and explain a substantial proportion of what we call the "Trump gap", or the tendency toward less social distancing within counties that exhibited greater support for Donald Trump in the 2016 presidential election.

First, using Word2Vec to capture interpretable differences in meanings attributed to the virus, we find that counties in which the virus is discussed in a way that is semantically similar to the concepts (a) the political left, (b) fraudulence, and (c) the flu and the common cold are less likely to social distance. We further corroborate these observational findings with experimentally manipulated texts and a BERT-based deep learning model, lending more internal validity to our results. Finally, we demonstrate with mediation analysis that inductively derived variations in the meanings associated with the virus explain nearly one-fifth of the association between support for Donald Trump and social distancing at the county level.

2 Data

In conjunction with a consortium of researchers at Stanford and UPenn, we curated a corpus of English-language, U.S.-based tweets that contained at least one of a set of coronavirus-related hashtags (for other work using related corpora, see Eichstaedt et al. 2015; Jaidka et al. 2020) that were created between February 28 and May 18.[2] We used the subset of these tweets for which we could identify a U.S. county of origin and which were not retweets or shares. After pre-processing (see below), this resulted in a final corpus of approximately 1.1 million coronavirus-related tweets originating from 181 U.S. counties from the end of February to the middle of May.

For our Word2Vec-based analyses (Sections 3 and 5), we preprocessed the Twitter data in several ways to ensure we were gaining the most accurate and potent signal possible. First, we removed stop words, URLs, punctuation, and all non-alphabetic characters. We then lowercased all letters. Next, we removed all tweets in which there were not at least two words. Finally, because we conduct analyses at the county level, we dropped all tweets from counties for which we had fewer than one-thousand tweets. In our BERT-based analyses (Section 4), we utilized a larger, less restrictive subset of the data that went through less preprocessing (described in Section 4).[3]

Using several publicly-available sources, we combined this data with county-level demographic information including income, population density, education, and Donald Trump's vote margin in each county during the 2016 US presidential election.

2.1 Measuring Social Distancing

To capture adherence to social distancing guidelines at the county level, we use data from Google COVID-19 Community Mobility Reports, which record daily human mobility levels in various settings such as workplaces, residential areas, and retail/recreation. We focus on the change in mobility in residential areas, a proxy for the amount of people staying at home.[4] For each county, we compute the average mobility scores for the week leading up to the observation period (February 21 to February 27) as well as the average mobility scores for the week just after the observation period (May 18 to May 24). We then take the latter

[1] https://www.google.com/covid19/mobility/

[2] We ended our observation period a week before the murder of George Floyd so that our outcome measure would not pick up any changes in mobility due to related protests

[3] Due to an error in our Twitter scraping pipeline, tweets in our dataset are truncated to a maximum length of 140 characters. We plan to replicate our studies with non-truncated tweets in future work.

[4] We acknowledge that this measure has limitations; it would not capture, for instance, the degree to which individuals in a county traveled to residential areas that were not their own home.

less the former as our measure of social distancing adherence.[5]

The resulting measure tells us the degree to which residents of each county adhered to social distancing guidelines over the observation period—specifically the degree to which residential mobility (which we interpret as being positively correlated to social distancing) increased over that time frame. In the counties we analyze in Sections 3 and 5, this measure ranges from 5.4 to 26.4 ($\mu = 14.5$, $\sigma = 4.0$). Because this measure captures within-county variation over time, the effect of county attributes that do not vary over time (and whose importance on the outcome do not vary over time) on our analyses should be minimized.

3 Deductive Word2Vec Analysis

We measure semantic associations between words used in a county's COVID discourse using the Word2Vec word embedding algorithm (Mikolov et al., 2013). The algorithm places each word that appears in a corpus in a high-dimensional space where proximity of two words in that space is proportional to the similarity of the words that appear in the linguistic contexts of those words. More intuitively, the Word2Vec algorithm creates a k-dimensional space in which semantically similar words, defined roughly by the interchangeability of those words, appear closer to one another than do semantically dissimilar words. We can then measure the semantic similarity of any two words by taking one less the cosine distance between the vectors assigned to those words corresponding to their position in the word embedding space.

The initial trouble with using Word2Vec to measure *variation* in meaning is that the algorithm requires a relatively large amount of text to reliably detect differences in the usage of words, and a model with a small number of examples will produce largely random noise. We combat this in several ways. First, we implement a novel measurement strategy whereby we take a fixed number of randomly selected tweets from each county and build a "baseline model". Then, for each county, we fine-tune that baseline model to create a "county-specific model". Instead of using only the similarity between words in the county-specific model, we take the difference between the similarity of two keywords (e.g., "coronavirus" and "hoax") in this model and the similarity of the same word pair in the baseline model. This allows us to leverage broadly consensual linguistic associations present across our corpus while also measuring how individual counties deviate from central tendencies in the useage of words. This process is illustrated in Figure 1.

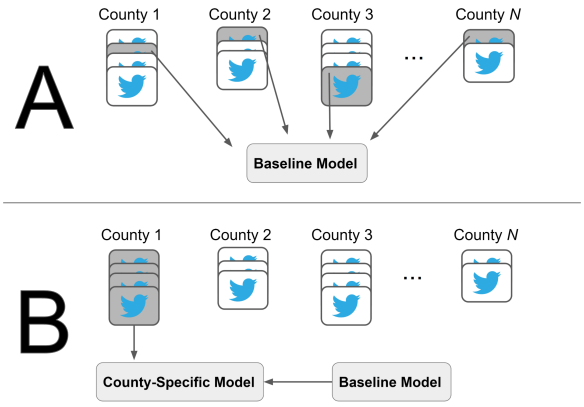

Figure 1: To build a baseline model and accompanying county-specific models, we (A) select a fixed-size, random sample of tweets from each county to train the baseline model and (B) create a copy of that baseline model for each county and fine-tune that baseline model on all tweets from each model's respective county.

Second, to get at any particular association between two concepts, we take the average similarity among a set of related words instead of just the similarity of a single representative word. These include the association of the virus ("coronavirus" and "covid") with (1) the political left ("democrat", "democrats", and "liberals"), (2) diseases that are less severe than COVID-19 ("flu", "influenza", and "cold"), and (3) fraudulence ("hoax", "fake", and "scam"). Finally, to combat the stochasticity in this measurement procedure, we create fifty separate baseline model/county-specific model pairs using random samples of 1,000 tweets from each county and take the average similarity change score across all fifty. We standardize each of these measures to ease interpretation.

Table 1 presents the results of a series of ordinary least squares (OLS) regression models predicting a county's change in residential mobility from these various linguistic associations. In models one, two, and three, we test whether our measured linguistic associations are correlated with our measure of the change in social distancing at the county level.

[5] Each individual measure is in comparison to a pre-COVID baseline, but since we take the difference in these scores, we essentially "net out" this baseline

Then, in models four, five, and six, we test whether these same associations are robust to the inclusion of a set of competitive control variables, including Trump's vote margin, population density, median income, and education.

Models one and four show the relationship between semantic association with the political left and our social distancing measure. Both with and without our competitive set of controls, counties with discourse more indicative of an association between the virus and the political left social distanced less over our observation period. Models two and five demonstrate that counties with discourse more indicative of an association between the virus and the concept of fraudulence social distanced less, and that this association is significant even in the face of controls. Finally, model three shows that counties in which COVID discourse exhibits a relatively stronger association between COVID-19 and less serious illnesses social distanced less over our observation period, but—as can be seen in model six—this association is not robust to the inclusion of our controls.

In the models without controls, a one standard deviation change in any of the three distances is associated with a sizeable decrease in social distancing (between -0.18 standard deviations and -0.22 standard deviations). Adding Trump vote margin and our control variables, we still see sizeable coefficients for two of our distance measures: a one standard deviation change in the association of the virus with the political left or with fraud is associated with a decrease in the change in social distancing of approximately 0.07 standard deviations and 0.09 standard deviations, respectively. Notably, a one standard deviation increase in Trump vote margin is associated with a decrease in our social distancing measure of approximately 0.4 standard deviations in models four, five, and six.

4 BERT Experiments

Observational results like the ones reported in Table 1 have their strengths, they also have inherent limitations. One major concern is that our Word2Vec model might pick up on linguistic information we do not intend for it to learn, potentially introducing omitted variable bias, biasing the parameter estimated in our OLS regressions. The problem is that while the language produced by county residents might vary in the linguistic feature of interest (e.g. likening the virus to the flu), it might also vary on many other dimensions. The fundamental problem is that we do not have experimental control over the generation of the text. To combat this, we combine deep-learning models using BERT (Devlin et al., 2018) with a minimal pair analysis using artificially-created data. The intuition is to train a deep-learning model to predict social distancing in a county from tweets in that county, and then feed texts into that model that we experimentally manipulate to only vary on the linguistic feature of interest. If this procedure replicates the findings from the above section, it will lend a great deal of internal validity to our conclusions.

Bidirectional Encoder Representations from Transformers (BERT) models are highly general in that they have been shown to achieve state-of-the-art performance on a large array of Natural Language Processing tasks. In this section, we apply a BERT fine-tuning approach to a wider sample of our county-level Twitter corpus. We find some evidence that our fine-tuned BERT models learn to associate our linguistic variables of interest (associating the virus with the political left, likening the virus to the flu, and questioning the reality or severity of the pandemic) with change in mobility on a larger sample of U.S. counties.

In contrast to analyses in Sections 3 and 5, in this section we drop only counties for which we have fewer than 512 tokens of Twitter data, leaving a total of 745 counties. We predict the same Google mobility metric described in Section 2.1 above.

Moreover, for this section the original tweets are lowercased but are otherwise unaltered so as to leverage pre-trained BERT's understanding of syntactic dependencies and compositional semantic meaning (that is, we do not employ the preprocessing described above): we leave in non-ascii letters, urls, hashtags, and an anonymized user tag (<user>). A dummy token—[NEWTWEET]—is prepended to each tweet in the corpus. We concatenate the tweets of each county into documents which serve as the basis of the analysis.

We use the `simpletransformers` library[6] to interface with the Huggingface transformers module (Wolf et al., 2019). We employ the `bert-base-uncased` model (12 transformer layers, 12 self-attention heads, hidden size 768, 110 million total parameters). The pooled output of the model is fed to a linear layer, with a mean squared error loss function to support regression.

[6] https://github.com/ThilinaRajapakse/simpletransformers

Table 1: Meanings of COVID-19 and Changes in Social Distancing

	(1)	(2)	(3)	(4)	(5)	(6)
Association with left	-0.210**			-0.0733*		
	(0.0737)			(0.0303)		
Association with fraud		-0.176*			-0.0870**	
		(0.0742)			(0.0299)	
Association with flu			-0.221**			-0.0420
			(0.0735)			(0.0303)
Trump vote margin				-0.416***	-0.417***	-0.435***
				(0.0349)	(0.0343)	(0.0344)
Other controls				✓	✓	✓
Constant	0.0	0.0	0.0	0.0	0.0	0.0
N	178	178	178	175	175	175
R^2	0.044	0.031	0.049	0.853	0.855	0.849

"Other controls" includes log population density, log median income, and education index

All measures are mean-centered and standardized

Standard errors in parentheses; excluded for constants

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

We employ a maximum sequence length of 512 (the longest sequence permitted by the model), a training batch size of 8, an evaluation batch size of 8, a learning rate of 4e-5 with Adam optimization (Adam epsilon value of 1e-8), and a training time of 10 epochs.[7]

4.1 Evaluation method

A 20% sample of counties are left for evaluation; the remainder of counties are used for training. The evaluation metric is accuracy (R^2) of predicted target values in the evaluation set against observed target values.

4.2 Prediction details

We fine-tune and compute predictions on the evaluation set with the `bert-base-uncased` model 30 times, generating 30 separate fine-tuned models. In each instance of fine-tuning, we first randomly sample a 512-token sequence from each concatenation of tweets from every county in the training set. A new random 512-token sequence is also sampled for counties in the evaluation set when generating predictions for fine-tuned models on the evaluation set (that is, no fine-tuned model sees the exact same Twitter data when making predictions on the evaluation set). We average the predictions of each model on the evaluation set to yield the ultimate predictions which we compare against our actual observed values.

4.3 Prediction Results

Figure 2 displays the predictions generated by the 30 BERT model runs against observed values in the evaluation set. The mean of these predictions achieve an R^2 of 0.43 on the evaluation set, that our BERT model is reasonably capable of predicting social distancing at the county level from tweets appearing in that county.

4.4 Experiment

Now that we have demonstrated that our deep-learning model can reasonably estimate our outcome in a county from tweet language originating from that county, we now move to assess whether our fine-tuned BERT models "learned" the semantic associations we measure in our deductive Word2Vec analyses (Section 3). We construct artificial minimal pairs of manually altered Twitter data and compare our models' predictions for each set

[7] We determined the epoch size by analyzing the evaluation loss over epochs for one random fine-tuning, discarded for future analyses. We found that most reduction in loss occurred after 5 epochs and that the loss was mostly level for the 5 subsequent epochs. The learning rate, batch size, and Adam epsilon value are defaults for the `simpletransformers` package.

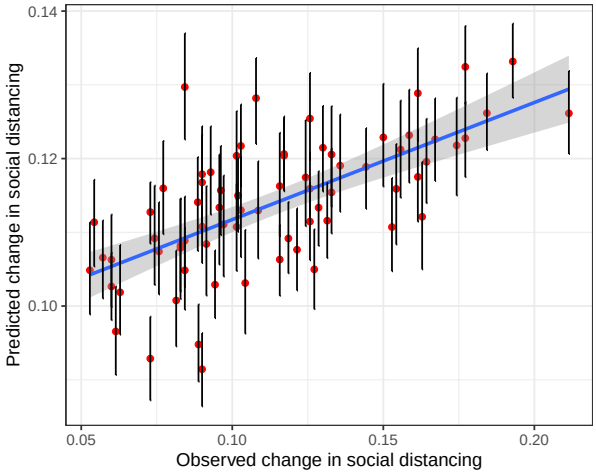

Figure 2: Average predicted change in social distancing (Google mobility metric, described in Section 2.1) from 30 fine-tuned BERT models, plotted against observed values in the evaluation set. Error bars indicate 95% confidence intervals of prediction means (displayed in red).

of data in the pair.[8] Minimal pairs differ according to one of the three following linguistic indicators of interest: blaming the pandemic on the political left, likening SARS-CoV-2 to the flu, and questioning the reality or severity of the pandemic.

For each comparison, we first create a *control* document of Twitter data, consisting of five manually-constructed tweets that suppress the linguistic indicator of interest; and an *experimental* document, where we manually alter tweets in the control document in order to evoke that linguistic indicator. Control and experimental sets were always less than 512 tokens long, in keeping with the constraints of our BERT architecture. These data are fed to our 30 fine-tuned BERT models to generate 30 predictions per set of data in each pair. Control and experimental documents used in this analysis can be found in the appendix.

4.5 Experiment Results

Model predictions on the control versus experimental sets of data are displayed in Table 2. For each pair of data, we compare the average predictions

[8] We take inspiration from the minimial pair analysis employed by Schuster et al. (2019), who probe the behavior of neural networks trained to predict the strength of pragmatic inference licensing in corpus data. For other recent applications of this type of analysis to neural network language modeling, see e.g. Ettinger et al. (2018), who probe the ability of neural nets to learn information about compositional semantic meaning; and Futrell et al. (2019), who probe their ability to learn syntactic representations.

Fine-tuned average predictions

Test	Control	Exper.	% Change
Flu	0.130	0.127	-2.31
Fraudulence	0.169	0.167	-1.18
Left	0.190	0.165	-13.2

Un-tuned predictions

Test	Control	Exper.	% Change
Flu	0.303	0.308	1.79
Fraudulence	0.287	0.308	7.34
Left	0.324	0.290	-10.6

Table 2: BERT model predictions on minimal pair sets described in Section 4.4. The rightmost column provides the percent change in mobility reduction predictions between the control and experimental sets for each analysis.

of our 30 fine-tuned models against the predictions of the `bert-base-uncased` model initialized without fine-tuning (identical architecture and random seed as our fine-tuned models). The fine-tuned models predict lower reductions in mobility for experimental documents vs. control documents on all three tests. To further contextualize these results, we look to the predictions of the un-tuned BERT model, which in the case of the "flu" and "fraudulence" experiments predicts a much *higher* reduction in mobility between our control and experimental data, counter to our predictions. Moreover, though the un-tuned model generates the predicted effect in the case of the "left" experiment, the magnitude of the effect is greater in the case of our fine-tuned models. In total, we take these results as evidence that the fine-tuned BERT models learned the linguistic correlates of social distancing that we explore elsewhere in the paper.

5 Mediation analysis

Now we have shown that the disparate meanings attributed to the virus are indeed associated with social distancing with two different methodological approaches. Here, we seek to demonstrate that these variations in meaning help explain the "Trump gap", or the empirical regularity that Trump vote margin in the 2016 presidential election is highly predictive of social distancing behavior during the crucial, early stages of the COVID outbreak in the U.S. Indeed, in our data, a linear model using that single variable to explain our outcome achieves an in-sample R^2 of 0.38.

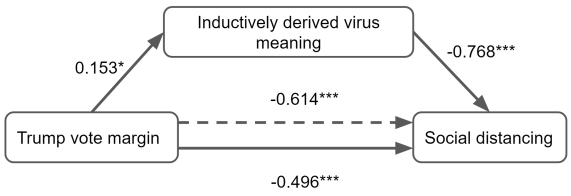

Figure 3: Results from structural equation modeling (SEM). Our inductively derived dimensions of virus meaning mediate 19.2% of the total effect of trump vote margin on our social distancing measure. All variables are mean-centered and standardized.

It is unlikely that the three semantic associations we test in the above sections account for the differences in meanings attributed to the virus across the U.S., and we do not claim to know all the different dimensions of meaning that have become imbued with political identity. We therefore take an inductive approach. Specifically, we measure the degree to which discourse in each county is indicative of cognitive associations (as above) between the virus and each of the one-thousand most frequent unigrams in our corpus (excluding "coronavirus" and "covid"). We then estimate the top three principle components of these measures, creating three orthogonal dimensions that explain the most variation in these one-thousand measures. These three dimensions succinctly describe the variation in meanings attributed to the virus in each county under analysis.

Before testing whether these three dimensions explain the Trump gap, we test the strength of the association between these dimensions and social distancing. In an OLS regression with these three variables predicting social distancing, all three coefficients are statistically significant at the $p < 0.001$ level. Further, the in-sample R^2 of the model is 0.36.

We are, however, primarily interested in the degree to which the meaning of COVID-19 explains the Trump gap. To test this, we identify how much of the variance in social distancing explained by Trump vote margin is mediated by these components. Using structural equation modeling, we find that 19.2% of the association between Trump vote margin and social distancing is mediated by these three components.

6 Discussion and Conclusions

In our deductive analses using Word2Vec, we demonstrated that counties social distanced less the more their COVID-19 discourse was indicative of cognitive associations between the virus and the concepts of (a) the political left, (b) fraud, and (c) less serious illnesses. We reaffirmed these findings using the contextual embedding BERT model, showing that a model trained on a large set of counties predicts a smaller increase in social distancing (i.e. less social distancing) for synthetic counties we created which (a) blame the left for the pandemic, (b) question the reality or severity of the pandemic, and (c) liken the virus with the flu, compared to control versions of these counties. Through mediation analysis, we show that the inductively derived heterogeneity in meanings the virus took on across the U.S. explains almost 20% of what we call the Trump gap, i.e. the empirical pattern that counties where residents supported Donald Trump in the 2016 election more social distanced less.

While it is difficult for us to make causal claims from the present analyses, we mitigate concerns about spuriousness by using within-county variation in social distancing as well as by controlling for several potential confounding variables. Additionally, we establish temporal precedence by using the change in the outcome measured after the observation period from which we measure our meaning-related variables. Our experiments described in Section 4 further mitigate these concerns.

Overall, our results confirm that the different meanings attributed to the virus shaped individuals' tendency to social distance. In fact, we demonstrate that these meanings explain a great deal of the association between Trump's vote margin and social distancing, an empirical regularity that has been demonstrated elsewhere. This means that if public health officials hope to increase adherence to social distancing (and potentially other health behavior-related) guidelines in the future, they must be mindful of the various meanings attributed to the virus in different communities, and attend to the dynamic process by which it might acquire new meanings.

These studies also provide a unique procedure for measuring the variation in meaning attributed to a novel concept in a population when limited text is available for each sub-population. Additionally, our results demonstrate the importance of the process of the social construction of meaning which sociologists have argued for over a century. Finally, we think this study pushes forward the field of computational social science, showing how tools from

computational linguistics can help demonstrate a process of social scientific interest and also help shed light on a phenomenon that is of great public health importance.

7 Appendix

7.1 Documents used in BERT minimal pair analysis

7.1.1 Mentioning the flu

Control document: [NEWTWEET] This virus is very different from the flu. [NEWTWEET] This virus is nothing like the flu. [NEWTWEET] This virus is much deadlier than the flu. [NEWTWEET] This virus is more dangerous than the flu. [NEWTWEET] I'm more afraid of this virus than I am of the flu.

Experimental document: [NEWTWEET] This virus is very similar to the flu. [NEWTWEET] This virus is just like the flu. [NEWTWEET] This virus is as deadly as the flu. [NEWTWEET] This virus is as dangerous as the flu. [NEWTWEET] I'm more afraid of the flu than I am of this virus.

7.1.2 Mentioning fraudulence

Control document: [NEWTWEET] The pandemic is real. [NEWTWEET] The pandemic needs to be taken seriously. [NEWTWEET] The virus is not a hoax. [NEWTWEET] The virus is not a scam. [NEWTWEET] The pandemic is not at all made up.

Experimental document: [NEWTWEET] The pandemic is fake. [NEWTWEET] The pandemic doesn't need to be taken seriously. [NEWTWEET] The virus is a hoax. [NEWTWEET] The virus is a scam. [NEWTWEET] The pandemic is completely made up.

7.1.3 Mentioning the political left

Control document: [NEWTWEET] People need to start taking responsibility during this pandemic. [NEWTWEET] The behavior of some people during this pandemic is totally reckless! [NEWTWEET] I blame this virus on careless people. [NEWTWEET] People are the cause of this virus. [NEWTWEET] I'm so angry at the people who caused this pandemic!

Experimental document: [NEWTWEET] Democrats need to start taking responsibility during this pandemic. [NEWTWEET] The behavior of some democrats during this pandemic is totally reckless! [NEWTWEET] I blame this virus on careless democrats. [NEWTWEET] Democrats are the cause of this virus. [NEWTWEET] I'm so angry at the democrats who caused this pandemic!

References

Hunt Allcott, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Y Yang. 2020. Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *NBER Working Paper*, (w26946).

Peter L Berger and Thomas Luckmann. 1966. *The social construction of reality: A treatise in the sociology of knowledge*. 1. Penguin Uk.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Johannes C Eichstaedt, Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, et al. 2015. Psychological language on twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify

100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Clifford Geertz. 1973. *The interpretation of cultures*, volume 5019. Basic books.

Amir Goldberg and Sarah K Stein. 2018. Beyond social contagion: Associative diffusion and the emergence of cultural variation. *American Sociological Review*, 83(5):897–932.

Kokil Jaidka, Salvatore Giorgi, H Andrew Schwartz, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2020. Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, 117(19):10165–10171.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Marcus Painter and Tian Qiu. 2020. Political beliefs affect compliance with covid-19 social distancing orders. *Available at SSRN 3569098*.

Craig M Rawlings and Clayton Childress. 2019. Emergent meanings: Reconciling dispositional and situational accounts of meaning-making from cultural objects. *American Journal of Sociology*, 124(6):1763–1809.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2019. Harnessing the richness of the linguistic signal in predicting pragmatic inferences. *arXiv preprint arXiv:1910.14254*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.