

Measuring Dimensions of Self-Presentation in Twitter Bios and their Links to Misinformation Sharing

Navid Madani¹, Rabiraj Bandyopadhyay², Michael Miller Yoder³, Stefan D. McCabe⁴, Briony Swire Thompson⁵, Kenneth Joseph¹

¹ University at Buffalo

²GESIS Leibniz Institute for the Social Sciences

³University of Pittsburgh

⁴George Washington University

⁵Northeastern University

smadani@buffalo.edu, rabiraj.bandyopadhyay@gesis.org, mmyoder@pitt.edu, stefanmccabe@gmail.com, b.swire-thompson@northeastern.edu, kjoseph@buffalo.edu

Abstract

Social media platforms provide users with a profile description field, commonly known as a “bio,” where they can present themselves to the world. A growing literature shows that text in these bios can improve our understanding of online self-presentation and behavior, but existing work relies exclusively on keyword-based approaches to do so. We here propose and evaluate a suite of simple, effective, and theoretically motivated approaches to embed bios in spaces that capture salient dimensions of social meaning, such as age and partisanship. We evaluate our methods on four tasks, showing that the strongest one out-performs several practical baselines. We then show the utility of our method in helping understand associations between self-presentation and the sharing of URLs from low-quality news sites on Twitter, with a particular focus on explore the interactions between age and partisanship, and exploring the effects of self-presentations of religiosity. Our work provides new tools to help computational social scientists make use of information in bios, and provides new insights into how misinformation sharing may be perceived on Twitter.

Introduction

On social media sites like Tumblr (Yoder et al. 2020), Twitter (Li et al. 2020; Rogers and Jones 2021; Pathak, Madani, and Joseph 2021), TikTok (Darvin 2022), and Parler (Bär, Pröllochs, and Feuerriegel 2023a), the profile description field (or *bio*) asks users to describe themselves in a single text box. Users commonly do so with a combination of delimited phrases, each of which can range from unigrams like “Republican” to more complex expressions like “2020 Election Truth Seeker” (Pathak, Madani, and Joseph 2021).

In most social settings, people aim to present to others only a single social identity (MacKinnon and Heise 2010; Schröder, Hoey, and Rogers 2016). Examples of identities include social roles, like “doctor” and “mother”, group memberships like “Democrat” and “Yankees fan”, and social categories like “black people” and “women” (Tajfel and Turner 1979). Bios thus present a unique setting where

we can observe individuals explicitly labeling themselves with multiple social identities (Marwick and boyd 2011).¹ With respect to the language of self-presentation (Johnstone 1996), bios are therefore an important tool for social scientists for three reasons.

First, bios provide insight into patterns of shared identity. For example, bios have been used to show that people who label themselves coffee snobs tend to be the same that label themselves marketing gurus (Pathak, Madani, and Joseph 2021). They have also been used to better understand the increasing entanglement of partisanship and culture in social identities in the United States (Essig and DellaPosta 2024; Phillips, Carley, and Joseph 2024). Second, bios allow us to study links between the identities people select for themselves and behavior. Users’ choices on how to label themselves can be linked to, for example, decisions on who chooses to re-blog whom on Tumblr (Yoder et al. 2020), or how people respond to marketing campaigns (Ng and Cruickshank 2023). Finally, bios can help us identify people who share a certain characteristic. To this end, prior work has also explored how bios can help identify individuals such as Qanon supporters (Bär, Pröllochs, and Feuerriegel 2023a) and journalists (Zeng et al. 2019), characterize the demographics associated with the self-presentation of particular identities (Pathak, Madani, and Joseph 2021), and even reflect current social movements (Rogers and Jones 2021; Hare and Jones 2023).

Existing quantitative work on bios has, however, has focused almost exclusively on characterizing bios via the presence of particular phrases, whether through dictionary-based methods (Rogers and Jones 2021; Zeng et al. 2019; Bär, Pröllochs, and Feuerriegel 2023b) or through analysis of patterns in phrases extracted with regular expressions (Yoder et al. 2020; Pathak, Madani, and Joseph 2021; Essig and DellaPosta 2024; Ng and Cruickshank 2023). While findings from these approaches can be illuminating, they also make it difficult to study how bios align with particular dimensions of importance to social scientists, such as partisan-

¹Bios really express phrases that *signal identity*, what Pathak, Madani, and Joseph (2021) call *personal identifiers*. Here, we retain the phrase (social) identity as a familiar and concise shorthand.

ship (Kozlowski, Taddy, and Evans 2019). Acknowledging these challenges and opportunities, Jiang, Ren, and Ferrara (2023) develop an approach that uses modern NLP methods to infer ideological leanings implied by bios. However, their work focuses only on learning political ideology and also requires data beyond the bio itself to perform estimation. There thus exists a need for a method that can help us to measure self-presentation in bios in ways that 1) let us move beyond keyword-based methods and 2) beyond political ideology, 3) without requiring additional data.

To this end, the goals of this paper are two-fold. The first goal of the present work is to develop and evaluate three methods to project English-language social media bios onto multiple dimensions of social meaning, without retraining a new embedding model for each new desired dimension. Our methods are technically straightforward, in that we extend existing work on how to scale text along dimensions of meaning like gender and race using projections of text embeddings (Wilkerson and Casas 2017). However, we build on this work by developing approaches specifically for the study of social media bios. Our models are *grounded in the relevant social science*, in that each method we present draws on social identity theory (Tajfel and Turner 1979) to learn bio embeddings by making use of the fact that many bios consist of multiple, clearly delimited social identities (Marwick and boyd 2011; Pathak, Madani, and Joseph 2021). As such, unlike in standard embedding models where we aim to embed words that mean the same thing closely together, our models aim to create embedding spaces (and projections of them) where phrases that are *applied to the same people* are close together in embedding space.

After describing our three approaches, we present an extensive validation across four tasks: 1) predicting which identities will appear in the same bio, 2) how projections of embeddings for specific identities within bios onto dimensions of age, gender, and partisanship correlate with human judgements, 3) whether projections using the embeddings of *entire bios* also correlate with human judgements, and 4) whether projections using the embeddings of entire bios correlate with other measures of user ideology, and show similar associations to behavior. Across all evaluation tasks, we find that the most effective model is an SBERT (Reimers and Gurevych 2019) model fine-tuned on bios. In Evaluation 1), we show that this model is best able to make predictions about identities the model has not seen, emphasizing its ability to generalize to new and unseen identities. Evaluations 2) and 3) show that this model effectively captures *perceptions* of how individuals present themselves, which is critical for understanding how other users interpret a particular bio (Heise 1987; Pathak, Madani, and Joseph 2021). Finally, *while our focus is on what is implied to others when a user self-presents, rather than inferring some “truth” about a user*, we show that estimates of user ideology by our method correlate with other known approaches for doing so. Moreover, we show that associations with behavioral patterns (of misinformation sharing) using our method are consistent with other approaches.

Having identified a single model to embed bios, we continue to the second goal of our work, which is to explore

associations between self-presentation in bios and the rate of low- (relative to high-) quality news shares on Twitter. Specifically, we focus on two research questions that have not been addressed in the existing literature. First, perhaps the most well-established finding in this literature is that misinformation sharing is more prevalent among older, right-leaning individuals (Grinberg et al. 2019; Guess, Nagler, and Tucker 2019; Brashier and Schacter 2020; Osmundsen et al. 2021; Nikolov, Flammini, and Menczer 2021). However, due in part to sample size restrictions, existing work has theorized (Grinberg et al. 2019) but not tested empirically the existence of an *interaction* between age and partisanship. In the present work, using two different large datasets of Twitter users, we indeed show for the first time that such an interaction effect exists between self-presentation of age and partisanship, in that self-presenting as older *and* Republican has a multiplicative association with misinformation sharing. Second, while scholars have *suggested* that religiosity is a critical dimension of the self associated with misinformation sharing online, no empirical work has addressed this point (Druckman et al. 2021). To this end, we present 1) new and convincing evidence that on average, presenting as more religious is strongly associated with misinformation sharing.

In sum, the present work provides three contributions:

- We propose, evaluate, and make public² a simple, effective, and theoretically motivated tool to embed English-language bios in socially meaningful latent spaces.
- We show that our method 1) can be used to project both individual social identities and entire social media bios onto salient dimensions of social meaning, such as partisanship, gender, and age in ways that correlate with human judgements in two new survey datasets, and 2) correlates well with other measures of user ideology.
- We use our model to extend our understanding of the relationship between how active news-sharing accounts on Twitter self-present and the proportion of news they share coming from low-quality news sites, particularly with respect to interactions between age and partisanship, and with respect to religiosity.

Background

Measuring text on social dimensions of meaning Social psychologists have developed a host of survey-based methods to measure associations between social identities and dimensions of social meaning. This estimation of identities on such dimensions is a core focus of social psychologists (Schröder, Hoey, and Rogers 2016; Fiske et al. 2002), because connecting self-presentation along specific dimensions of social relevance to social behavior is important for developing and testing new theories (MacKinnon and Heise 2010). Below, we leverage these established approaches to evaluate our methodology. However, survey data do not scale to the myriad ways people identify themselves (MacKinnon and Heise 2010), are usually too small to capture differences across subgroups or contexts (Smith-Lovin

²The model is publicly available for use on https://github.com/navidmdn/identity_embedding

and Douglas 1992), and struggle to account for linguistically complex identities or situations where multiple identities are applied (Joseph and Morgan 2021).

Most of the computational tools developed to address these challenges function by projecting embeddings from distributional semantic models (*DSM*), such as BERT, onto particular dimensions of meaning. The present work is most aligned with efforts that use contextualized embeddings (e.g. Kurita et al. 2019; Lucy, Tadimeti, and Bamman 2022; Field, Bhat, and Tsvetkov 2019) to do so. However, we expand on these methods in that we aim to focus not on *linguistic similarity*, but rather *similarity in the types of people who use particular identities*. More specifically, DSMs are based on the assumption that contextual similarity—similarity in where phrases appear in text—is a strong proxy for semantic similarity (roughly, synonymy). The idea behind this assumption is that phrases with high semantic similarity should have similar cognitive associations to other phrases, and thus high contextual similarity too (Miller and Charles 1991). While deeply intertwined (Caliskan and Lewis 2020), these *linguistic* associations and the *socio-cultural* associations of interest to us differ. Linguistic associations represent phrases with similar associations to *similar other phrases*; what we desire are phrases with similar associations to *similar kinds of people*. While methodologically consistent with prior work, then, our work extends the existing literature by defining a different *socio-theoretic* goal that is consistent with the existing literature on self-presentation. We also provide several new evaluation datasets for future work.

In focusing on sets of identities applied to individuals, our work also relates to *entity-centric* text analysis (Field and Tsvetkov 2019). Entity-centric work focuses on using phrases with known meanings (e.g. from surveys) to understand the portrayal of individuals (Antoniak, Mimno, and Levy 2019; Mendelsohn, Tsvetkov, and Jurafsky 2020). Our method builds on a complementary idea, namely that we can use the fact that all identities in a bio refer to a single entity (a user) to create better embeddings. Similar in this vein is the work of Bamman and Smith (2014), who use this idea to infer character personas in literature. The present work complements these efforts by using entity-centric data to produce embeddings, rather than phrase clusters.

Our work also ties to the literature that explores the language associated with how people express misinformation (Mu and Aletras 2020; Shu et al. 2017; Rashkin et al. 2017). More specifically, we complement these efforts to study how people express specific instances of misinformation by exploring how people who tend to spread misinformation present themselves via particular (sets of) social identities.

The Demographics (and Self-presentations) of Misinformation Sharing The study of misinformation online has exploded in recent years (Lazer et al. 2018). Within this literature, several papers have looked at associations between demographics and rates of misinformation sharing. Typically, they do so by gathering demographics either via survey (e.g. Grinberg et al. 2019) or using voter records (Guess, Nagler, and Tucker 2019; Moseleh, Pennycook, and Rand 2020), and then associating contained demographic information with

rates of misinformation sharing. Across studies, however, prior work has repeatedly found that the best predictors of misinformation exposure and sharing online are 1) old age (Brashier and Schacter 2020), 2) alignment with the political right (Osmundsen et al. 2021; Nikolov, Flammini, and Menczer 2021), and 3) overall levels of online activity (Grinberg et al. 2019).

Our work compliments these existing efforts in a number of ways. First, these prior works tend to use small-N samples of misinformation sharers because of the challenges and biases associated with their recruitment methods (Hughes et al. 2021). For example, Grinberg et al. (2019) and Guess, Nagler, and Tucker (2019), analyze sharing patterns in settings where only 400 Twitter users and 101 Facebook users shared any misinformation, respectively. In contrast, we analyze two different and much larger datasets; in the larger one, 77,190 accounts share at least one low-quality news link.

Second, we focus here on demographics *conveyed through self-presentation*. These self-presented demographics are important in their own right for understanding who *other* Twitter users *perceive* to be sharing misinformation, perceptions that do not always align with a user’s “true demographics” (Nguyen et al. 2014). **Critically, then, we do not claim that our method infers demographics of users, nor do we believe it does so.** Indeed, these self-presentations may vary from demographics in at least two ways. First, individuals may consciously choose not to convey certain demographic information online, such as gender (Buss, Le, and Haimson 2022), and more generally choose which dimensions of the self are most important to present to their Twitter audience (Marwick and boyd 2011). Second, non-human accounts, such as those run by Russia’s Internet Research Agency (IRA), may fabricate self-presentations to shape online discussion (Zhang et al. 2021). Prior studies that link users to surveys or voter records rule out these latter accounts in their sampling approach. In contrast, we aim to focus on which dimensions of social meaning are salient, or cast as salient, by those sharing misinformation unknowingly or for manipulation. We thus study here *how the misinformation shares present their identity*.

These two distinctions shape the two novel research questions we study here. First, the larger size of our dataset lets us explore *interactions between* self-presentations of age and partisanship, as compared to prior work which analyzes only the main effects. Second, our focus on dimensions of self-presentation instead of traditional demographics lets us analyze a novel, although well-theorized (Druckman et al. 2021), empirical question: how is (self-presented) religiosity associated with misinformation sharing?

Methods for Embedding Twitter Bios

We propose three models that leverage existing methods to project phrases onto dimensions of social meaning in different ways. Our first model uses data only from bios, whereas the latter two use fine-tuning to balance between meanings in bios and semantic information in large, pre-trained DSMs. In all cases, our models are trained by using patterns in the multiple identities that appear in many social media bios.

Because of this, it is useful to introduce some limited notation. First, let X denote a dataset of bios where identities have been extracted, e.g. by using a regular expression (Yoder et al. 2020; Pathak, Madani, and Joseph 2021). We assume $X^i = \{x_1^i, x_2^i, \dots, x_k^i\}$ represents a set of k identities extracted from a single bio, and that V is a vocabulary of all unique identities in the training portion of X .

Models

Bio-only model Our *Bio-only model* is constructed by applying `word2vec` (Mikolov et al. 2013) to X . In common terminology for `word2vec`, we treat identities as words and bios as a context. Our intuition is that if the `word2vec` model can leverage contextual similarity on the “word-to-linguistic context” matrix to identify words with shared semantic meanings, it may also be useful to leverage the “identity-to-person context” matrix to identify phrases with shared socio-cultural meanings. We use `word2vec` models with an embedding size of 768 to match the embedding size of the other models used below, and train for 300 epochs with a window size of 8 (only .01% of bios in our training data contain more than 8 identities). Additional minor details are provided in the appendix.

Fine-tuned BERT To fine-tune BERT, we use a masked language modeling (MLM) objective, randomly masking one of the identities in each bio. To prepare our dataset for training, we take each of the instances X^i and concatenate the phrases in it to form a full sentence. We then mask one of the identities and fine-tune a BERT-base model for 5 epochs while monitoring 10% of the training set as validation data. We used a learning rate of $2e-5$ with a batch size of 64. Model training took approximately one day using a single A100 GPU. When bios are shorter than the context window size, padding is added; attention masking is used to ensure padding does not impact the embedding. We keep the embedding size of the BERT model at the default 768.

Fine-tuned SBERT Finally, we construct a final model by fine-tuning Sentence-BERT (Reimers and Gurevych 2019). Sentence-BERT uses *contrastive learning*, where the learning setup must be carefully constructed (Schroff, Kalenichenko, and Philbin 2015). We develop an intuitive but effective approach here based on shared identities. In a contrastive learning framework, each data point is a *triplet* consisting of an *anchor*, a *positive*, and a *negative* sample. Our goal is to reshape the embedding space through fine-tuning such that for each triplet, the distance between anchor and positive samples, which co-occur within a bio in our setup, is minimized while the distance between anchor and negative samples, which do not, is maximized. As above, padding and attention masks are used when bios are shorter than the size of the context window. The embedding size of our selected sentence BERT model is 768 which matches the embedding size of all other proposed approaches for fair comparison.

We can frame this contrastive learning problem as a regression task: given a triplet of anchor (X_a), positive (X_p) and negative (X_n) samples and a similarity measure (here, cosine), our objective is for $cs(X_a, X_p) = 1.0$ and

$cs(X_a, X_n) = 0.0$ for all training points, where cs stands for cosine similarity. We can then optimize this objective using mean squared error. The challenge is to construct an effective set of triplets to train on. To do so, we first take a bio X^i from the training set, and then randomly select an identity from X^i to be the positive sample. We name the remaining identities in X^i the anchor sample. Finally, we randomly select an identity that never co-occurs with the positive sample as the negative sample. As an example, from the bio [*assistant professor, Bernie supporter, #blacklivesmatter*] we set *assistant professor, #blacklivesmatter* as the anchor sample, *Bernie supporter* as the positive sample, and randomly select a negative sample that never co-occurred with *Bernie supporter*. We construct a triplet for each $X^i \in X$ using this method, and use these to fine-tune an *mpnet-base* Sentence-BERT model. Models were trained for 5 epochs for one day on a single A100 GPU.

Training Data

All models are trained on a sample of 3,534,903 bios from users who sent an English-language tweet captured in the Twitter Decahose³ in 2020. We use the method from Pathak, Madani, and Joseph (2021) to extract identities from bios. Their extraction method consists of two steps, one where bios are split into chunks using a manually crafted regular expression, and a second cleaning step. For example, from the Twitter bio “Progressive Christian, wife, I am a proud Canadian,” their method extracts *Progressive Christian, wife* and *proud Canadian*. Given that our models learn from patterns in shared identity, we use only bios that contain at least two identities for training, with $|V|=22,516$. A complete description of the dataset is available in the appendix.

Evaluation

We conduct four evaluations to assess model validity. The first is a prediction task, where we evaluate the full embedding space of each model and its ability to capture information that shapes users’ decisions on which sets of identities to place in their bio. The second connects our work to the existing literature on embedding and projecting individual social identities onto dimensions of social meaning. The third assesses our ability to project entire bios onto meaningful dimensions, with an eye towards validating our approach for our case study. Finally, the fourth task shows that our measures of partisanship correlate with other existing measures, and that our methods are associated with the behavior of interest (misinformation sharing) in the same way. Here, we describe each task and results for it in four separate subsections.

Can we Predict Held-out Identities?

Task Description To perform this evaluation, we randomly sample another 1.5M Twitter bios from the Decahose using the same approach as above, creating an 80/20 train/test split between the two samples. For each observation in the test set, we ensure that *at least one* of the identities

³A sample of approximately 10% of all tweets.

is in V (which is defined using the training data). For the i th sample, we then take one identity, X_t^i , as the hold-out target and call the rest of the bio X_r^i . We ensure X_t^i is in V , i.e. in all cases the target is observed at least once in the training data. This task notably favors the three new models we present here, in that they are each trained in a manner similar to the prediction task. However, the comparison is still useful to 1) ensure that fine-tuning works as expected and 2) to compare between the three new models. With respect to the latter point, there is reason to believe that the word2vec model should outperform the Fine-tuned BERT or fine-tuned SBERT model, because the word2vec model is trained *only* on in-domain data relevant to the task.

To generate predictions, we first generate an embedding for X_r^i , $L_r^i = \text{embedding}(X_r^i)$ ⁴. We then measure the cosine similarity of L_r^i with the embedding of all identities $v \in V$, $\text{Similarity}(L_r^i, L_v^i)$, leaving us with $|V|$ similarity scores to L_r^i . We evaluate similarity scores returned by each model using two evaluation metrics: *average rank* and *log softmax score*. Average rank is computed by, for each test point, finding the ranking of X_t^i in the scores produced by each model, and taking the average over all test points. The log softmax score draws on prior work (Joseph and Morgan 2021) and transforms similarity scores into a probability distribution using the softmax, and then takes the log of the result for X_t^i .

Finally, for evaluating the out-of-domain generalizability of our models, i.e. how well the model generalizes to unseen identities, we split our test data into two sets, a *Main Evaluation* set, where X_r^i also contains at least one identity observed in the training data, and a *Generalizability* set, in which no identities in X_r^i are seen in the training data. This is necessary to fairly evaluate our *Bio-only model*, which has a restricted vocabulary and does not generalize to out-of-domain phrases, to the other models, each of which are capable of handling out-of-domain text.⁵ It is also a useful test of the (in-domain) generalizability of the other models. We evaluate results separately for these two test datasets.

We compare our models to three baseline DSMs used frequently in prior work: BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), and Sentence-BERT (Reimers and Gurevych 2019) (see appendix for details). As in much of the prior work (Lucy, Tadimeti, and Bamman 2022), these approaches are *not* fine-tuned on bios, giving a baseline for how important in-domain training is for our problem.

Results In contrast to our expectations, the Fine-tuned SBERT model consistently outperforms all other models on both evaluation metrics on the Main Test set. Figure 1 shows this, and also reveals that the next best model, in all cases, was the *Bio-only model*, and that the *Fine-tuned BERT* model does not show the same jump in performance relative to the baseline BERT model that the *Fine-tuned SBERT* model does. Finally, we see that the baseline SBERT model outperforms the baseline BERT model.

⁴Extended details on how embeddings are generated for each model are in the Appendix.

⁵Note that X_r^i can still contain phrases that the Bio-only model does not recognize and replaces them with the 0-vector.

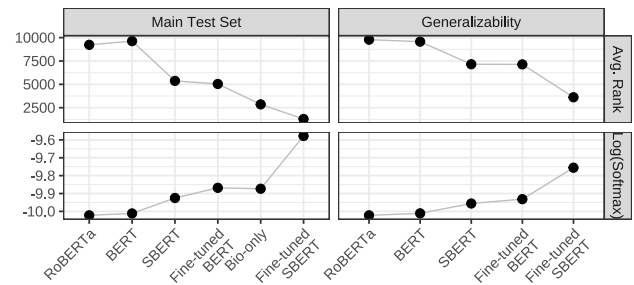


Figure 1: Performance of each model (x-axis) on each of our two outcome metrics (separate plot rows) for the Main Test set and the generalizability test set (separate plot columns). Note that for rankings, lower is better.

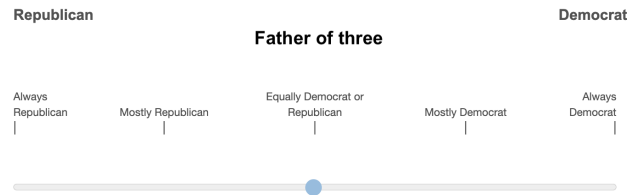


Figure 2: An example of a survey question asked on the single identity projection evaluation for the identity “father of three” on the partisanship dimension

The fact that the fine-tuned SBERT model improves over the Bio-only model, but the Fine-tuned BERT model does not, is evidence that knowledge from the pre-trained SBERT model (but not BERT) is useful in our setting. It also shows that performance gains cannot only be attributed to fine-tuning on in-domain language, but instead that our contrastive learning setup was effective and that Sentence-BERT is indeed the more effective initial model for fine-tuning, at least on this evaluation task. The second column of Figure 1 shows that the Fine-tuned SBERT model also performs best on the Generalizability test set. Even when the Fine-tuned model is not exposed to any of the identities in X_r^i , it improves by nearly 100% over the standard SBERT model in terms of average rank. Figure 1 also makes clear that there is room for improvement. To this end, we conduct an error analysis; see the appendix for details.

Do Projections of Single Identities Correlate with Human Judgements?

Task Description Following prior work (Joseph and Morgan 2020), we assess how well our embeddings can be used to project identities onto dimensions of social meaning in ways that correlate with human judgements. As no prior work focuses on identities that are common in bios, however, we construct a novel survey dataset for this task. We provide brief details on the study here, full details can be found in the Appendix.

We asked 140 respondents on Prolific to rate 250 common identities in Twitter bios on three widely studied dimensions: gender, age, and partisanship. We use the mean

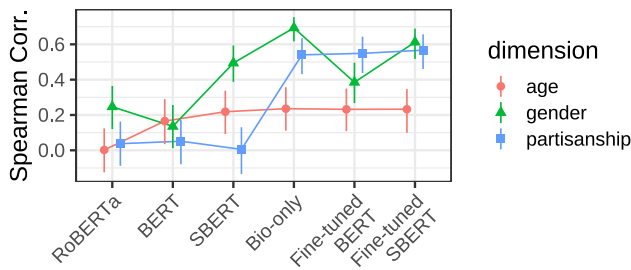


Figure 3: The Spearman correlation (y-axis) between projections and human judgements on 250 social identities for each model (x-axis) and dimension (shape/color). Error bars are 95% bootstrapped confidence intervals.

of ratings for each identity on each dimension in our analysis. To select identities, we ranked identities in the Twitter dataset by frequency, and then manually selected the first 250 phrases that clearly signaled identity. For each identity on each dimension, respondents were asked to move a slider to represent their perception of where people who label themselves as that identity were likely to fall. For partisanship, for example, the slider ranged from “Always [a] Democrat” to “Always [a] Republican.” Figure 2 provides an example question. For gender and age, we followed the approach outlined by Joseph and Morgan (2020) exactly, using the same slider. Specifically, for age, participants were asked to rate identities on perceived age from 0-100, for gender, the question appears as in Figure 2, replacing “Republican” with “Man” and “Democrat” with “Woman.”

To construct projections for each embedding model, for each identity, onto these dimensions, we follow the literature and 1) embed identities as defined above, 2) define a set of words and phrases that denotatively characterize each “end” of the dimension (e.g. “man” vs. “woman” for gender) and then 3) use these to project each identity onto a line in the embedding space defined by those two dimension ends, giving a single number. Several approaches exist to complete steps 2) and 3) (Joseph and Morgan 2020). We follow prior work where possible for 2), and for 3), i.e. to calculate similarity, we use RIPA, the method described by Ethayarajh, Duvenaud, and Hirst (2019). Finally, for each embedding model on each dimension, we then compute the Spearman correlation between the projections and the survey data.

Results Figure 3 shows that our three models all perform at least as well—but not better than—the baselines for projections onto age, all significantly improve over baselines in measuring perceived partisanship, and significantly improve over the baselines, except for Fine-tuned BERT, on perceived gender. Our models’ improvements over baselines are particularly salient in comparison to the RoBERTa and BERT baselines. Notably, we also see that our Bio-only model, based on word2vec performs well compared to the more complex models on this task. This suggests that for phrases that are prominent in bios, this model (which trains only on the bios themselves) may be preferred. However, the Bio-only model cannot extend beyond the vocabulary, and

thus (as we will see) struggles with generalizability. Finally, Figure 3 also shows that the greatest jumps in performance for our models, relative to the baseline, are clearly on the partisan dimension.

Do Projections of Entire Bios Correlate with Human Judgements?

Task Description Our third evaluation assess the question, can our model capture perceptions of self-presentation in entire bios? To address this, we conduct a similar analysis as above, comparing the Spearman correlation of projections of our embedding models to mean ratings by Prolific respondents on a survey task. There are, however, four primary differences between the evaluation of single identities and the one presented here for full bios.

First, of course, is that we ask respondents for their perceptions of entire bios, rather than individual identities within bios. More specifically, we randomly sample 1,300 bios of users in our case study data, described below. Second, because in our case study we are interested in religiosity as well as age, gender, and partisanship, we add a question regarding perceived level of religiosity to the survey. Third, as opposed to selecting only one approach for defining ends of the semantic axis onto which embeddings are projected, we consider two approaches. The first follows our second evaluation and uses a combination of prior work and author intuition to define the terms at each end of the axis. Motivated to ensure accurate measurements for our case study, the second considers whether or not we can improve correlations with human judgement by constructing lists of terms for dimension endpoints that are informed by a qualitative analysis of bios. More specifically, we manually explore bios from the case study data described below that are not included in the survey study and use them to define the list of terms. We compare performance using both approaches here, *but ensure that we compute results only on bios that do not include terms that explicitly define our endpoints*. Fourth, we opt to only compare performance of SBERT and Fine-tuned SBERT, as other models fared poorly on one or both of the prior evaluations.

All other details of our evaluation generally match those in our second evaluation, save for the sample of Prolific users; see the appendix for details on this.

Results Figure 4 shows that the Fine-tuned SBERT model has a significantly higher correlation with human judgements on dimensions of partisanship and religion, regardless of how endpoints of the dimensions are defined. For gender and age, Fine-tuned SBERT performance shows no significant improvement over a non-fine-tuned SBERT model, but is also not significantly impacted by how dimension endpoints are defined. In contrast, for the SBERT model, in one case (age) our use of qualitative exploration to find in-domain sets of phrases to define dimensions significantly increases correlations with human judgement, and in the other (gender) it significantly decreases correlations. These results suggest that the Fine-tuned SBERT model is therefore more robust to the known challenge of defining “good” sets of phrases to define endpoints (Joseph and Morgan 2020).

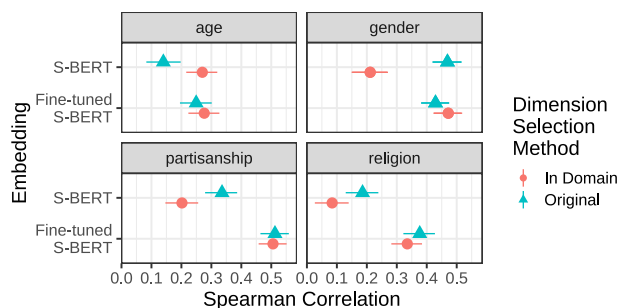


Figure 4: Spearman correlations (x-axis) between projections onto four different social dimensions (separate subplots) for SBERT and Fine-tuned SBERT (y-axis) for two different dimension selection methods (color/shape). Error bars are 95% bootstrapped CIs.

How do our Projections Compare to Other Measures at the User Level?

Data In our final evaluation, we focus specifically on projections of bios onto partisanship and explore 1) how well these projections correlate with a different measure of user partisanship (evaluating convergent validity), and 2) whether or not our measure of partisanship correlates in expected ways—and in a similar way to the other measure of user partisanship—with misinformation sharing (evaluating hypothesis validity Jacobs and Wallach 2021). To conduct these evaluations, we make use of a pre-existing dataset of 374,684 Twitter users 1) for whom prior work (McCabe et al. 2022) has computed a standard measure of user ideology using user follower networks, and 2) for whom we have data on the sharing of low-quality news sites.

With respect to the partisanship measure we compare to, McCabe et al. (2022) apply a method pioneered by Barberá et al. (2015) that uses a combination of information about the partisanship of elite Twitter users and follower relationships to compute an estimate of user partisan ideology. McCabe et al. (2022) adopt the same approach, but update information about elites from the 2012 data used by Barberá et al. (2015) to 2020 data, and apply the approach to a large set of Twitter users that they linked to voter registration records using established methods (Hughes et al. 2021). The work from McCabe et al. (2022) presents full details on implementation and an extensive validation of the measure in comparison to voter registration data; we therefore refer the reader to their work these details and focus here only on a comparison between their updated version of Barberá et al.’s 2015 method and our partisanship measure using bio data.

With respect to the sharing of low-quality news, we focus on the proportion of URLs to low- and high-quality news websites shared on Twitter by these users from July 1st, 2020 through May 31st, 2021. We determine whether a URL comes from a low- or high-quality news website by making use of 1) the list produced by Grinberg et al. (2019) (who categorize sites into a binary high or low-quality distinction) and 2) the NewsGuard domain rating list. NewsGuard is an organization that maintains a widely-used (e.g.

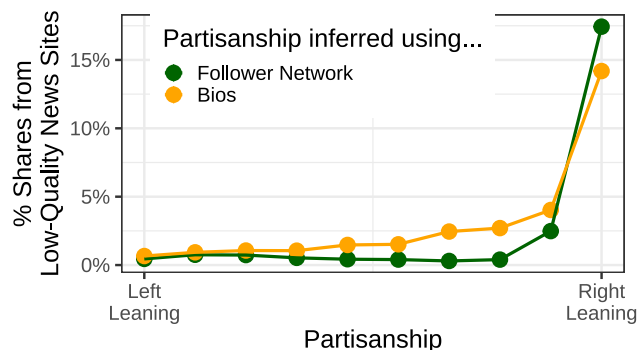


Figure 5: Proportion of low-quality shares (y-axis) across bins (n=10) of projections onto the partisanship (x-axis) dimension, estimated using two different methods (color). Error bars, while small, are present in the figure, and represent 95% normal CIs.

Altay, Nielsen, and Fletcher 2022; Horne, Nørregaard, and Adalı 2019) list of news websites that are rated on a scale of 0-100 for information quality. We follow prior work and dichotomize their ratings into a binary low- or high-quality value for each site. We use a score of 60 as the threshold, following prior work (Lin et al. 2022). Further, while we opt for a particular definition across these two lists, we note that prior work (Lin et al. 2022) has shown that various lists of misinformation domains are highly correlated, and thus that results for the study of misinformation are unlikely to be sensitive to these types of changes. Using these binary domain scores, we define our outcome variable for an individual user as the percentage of all URLs the user shares that come from a low-quality news site, divided by the total number of URLs shared from a domain listed in NewsGuard. We refer to this outcome as the *proportion of low-quality shares*.

Finally, for our bio-based measure, we use user bios collected in July of 2020 (i.e. within the same timeframe of the following data and misinformation sharing data), and project bios onto partisanship using the methods detailed in our evaluation of full bio embeddings (our third evaluation).

Results We find a correlation of 0.39 [0.387,0.393] between the bio-based projection measure of partisanship using our fine-tuned SBERT model and the partisanship estimates from McCabe et al. (2022). This correlation is nearly double the correlation we find between the bio-based projection measure computed using the non-fine-tuned SBERT model and the estimates from McCabe et al. (2022) (0.197 [0.193,0.201]), and even higher than the correlation between the fine-tuned SBERT model and the non-fine-tuned model (0.347 [0.343,0.35]). This strong correlation, even when compared to reasonable baselines, gives us further confidence in the convergent validity of our measure—while we should not expect a measure based on a more private behavior (following) to correlate exactly with one based on a more public behavior (text in bios), we do see, as expected, a reasonably strong relationship.

We also find that these two measures show consistent es-

estimates of the relationship between partisanship and misinformation sharing. These estimates, moreover, conform with what we would expect from prior work (Grinberg et al. 2019; Guess, Nagler, and Tucker 2019). More specifically, Figure 5 shows the estimated proportion of low-quality shares from a binomial regression model where the independent variables are indicators for partisanship after bucketing each variable into 10 separate bins of equal numbers of users, respectively.⁶ All main effects in the model used to estimate the proportions shown in Figure 5 are significant at $p \leq .001$, and trends between the two different measures are consistent in showing that those furthest to the partisan right are most likely to share low quality news.

Misinformation Case Study

The previous evaluations show that our Fine-tuned SBERT model can project bias onto social dimensions of meaning in ways that correlate with how those bios are likely to be perceived by humans, and that the approach has both convergent validity and hypothesis validity in comparison to a different measure of user partisanship. Our case study uses the same dataset, but leverages our methodology to explore the link between self-presentation in bios and the rate of low-versus high-quality news sharing in the context of our two research questions. First, we explore the *interaction between* self-presentations of age and partisanship and their association with low-quality news sharing. Second, we consider the association between self-presenting as religious and misinformation sharing.

We note that while it would be possible for McCabe et al. (2022) to make use of additional voter registration data to, e.g., control for age, our tool is still useful for two reasons. First, most existing social media datasets are not connected to voter registration data, nor are the aligned with follower network data that could be used to apply the method from Barberá et al. (2015). To this end, we are able to use our methods to replicate our analysis on a new dataset from the Twitter Decahose, which we could not do otherwise (see appendix for details). Second, we note that an analysis of religiosity is in any case not possible with existing voter registration data, as religiosity does not exist within versions of these data that have been provided to researchers in the past.

Results

With respect to our first research question on the interaction between age and partisan self-presentation, we find that the direction of effects for age and partisanship generally match prior work, but that interaction effects not observed in prior work on demographics appear in our analyses. Figure 6 more specifically shows that users who self-present as older *and* right leaning is associated with a superlinear increase in proportion of low-quality news shared relative to being old or right-leaning.

⁶Binning is done to avoid assumptions about the precision of these proxy variables and to ease interpretation; bin sizes selected here are for visual clarity. Results are not, however, sensitive to this decision.

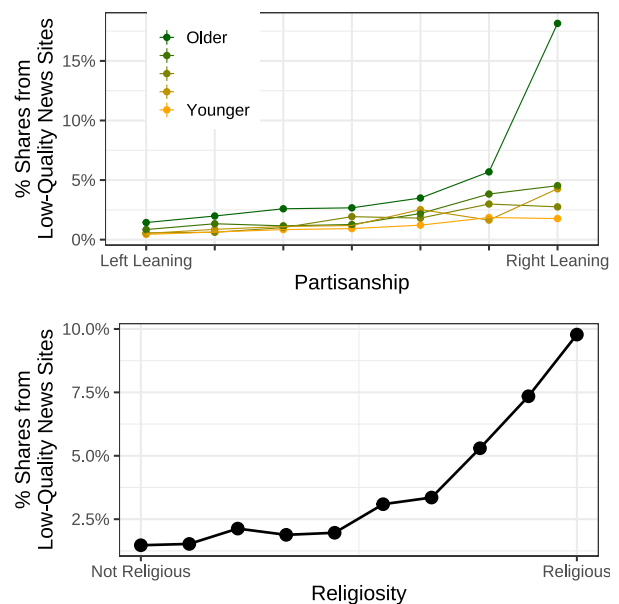


Figure 6: A), Top: Proportion of low-quality shares (y-axis) across bins ($n=10$) of projections onto the partisanship (x-axis) and projections onto the age ($n=5$ bins) (color) dimensions. Error bars, while small, are present in the figure, and represent 95% normal CIs. B), Bottom: Univariate model outcomes for religiosity

Addressing our second research question, and using the same binning approach with univariate statistical models, we observe in Figure 6B) that self-presenting as being highly religious is also associated with statistically significant ($p < .001$) jumps in the proportion of low-quality shares. As in the real world (Druckman et al. 2021), our measures of religiosity and partisanship are highly correlated (*at 0.76*), and thus we do not attempt to tease out whether this effect is particularly salient conditioned on partisanship. More specifically, we avoid interpretation of coefficients in a model with both of these variables because interpretation of regression coefficients with strongly co-linear variables is prone to misidentification of effect size and statistical significance (Gregorich et al. 2021).

We do find, however, that religiosity adds additional predictive power, in that adding religiosity as a predictor to a model with partisanship results in a statistically significant increase ($p < .001$) in predictive power. Descriptive statistics do, however, provide some evidence into a potential interactive relationship between partisanship, religiosity, and sharing content from low-quality sites. Specifically, Figure 9 in the appendix shows that partisanship and religiosity may have a multiplicative association with sharing content from low-quality news sites. Participants who present as both right-leaning and religious have, on average, higher odds of sharing low-quality news links than users who present as only right-leaning or only religious. Future experimental work targeting this specific interaction effect in ways that avoid endogeneity concerns might therefore be

of interest.

Critically, all findings presented here extend beyond this particular dataset to a second large dataset of misinformation sharing drawn from a broader population in the Twitter Decahose. These replication results are presented in the appendix.

Discussion and Conclusion

We present, evaluate, make public, and make use of new methods to project Twitter bios onto meaningful social dimensions. Methodologically, our contribution is not innovation in algorithm development, but in the application of standard algorithms to carefully constructed training datasets and training set-ups to build entity-centric identity embedding spaces (where identities that are commonly presented for the same people are represented in a similar space). Notably, our estimates of self-presentation do not use other user behaviors (e.g. retweets).

After extensively validating our proposed methods, we present a case study on two open questions in the literature on misinformation sharing online. In response to our first question, we find an interaction effect between age and partisanship- that Twitter users presentinand Republican share a much higher proportion of low-quality news. With respect to our second research question, we similarly find that individuals who self-present as religious, perhaps especially on the political right, are much more likely to share a higher proportion of low-quality news. Our results have two important implications. With respect to combating misinformation online, while much work has considered interventions targeted along partisan (Martel et al. 2024) or age-based (Brashier and Schacter 2020), our results provide evidence that we should be considering even more narrow interventions targeted at older *and* right-leaning Americans. With respect to better understanding the misinformation environment online in the U.S., our work provides critical empirical validation of theories that suggest religiosity is an important component of the quality of information consumed and spread online (Druckman et al. 2021).

However, our work contains a number of limitations that should be considered. Specific to our case study, we follow previous work that uses a domain-based approach to identify misinformation sharing (Grinberg et al. 2019; Guess, Nagler, and Tucker 2019; Moore, Dahlke, and Hancock 2023), which includes posts that share a URL with a negative annotation. However, like previous work, we find the number of such posts to be a small proportion of all shares. While other work referenced above suggests our efforts should extend to other social media sites, Twitter, like all sites, has particular elements that limit our ability to generalize claims about self-presentation elsewhere, in particular to “offline” behavior settings (DeVito, Walker, and Birnholtz 2018). In particular, our models learn cultural associations from the majority white, male, younger, and left-leaning users of Twitter (Hughes et al. 2021). Similarly, our survey evaluation data is largely drawn from white, American women, which limits their generalizability.

Our work is also contextualized in a particular period of time, and thus may not be responsive to significant shifts

in social identity. All of these, and results described in the paper, emphasize the contextualized nature of our trained models and datasets. While we hope that future work might expand beyond them, others should consider using our publicly available code to fine-tune our models on their own data. Finally, our case study is necessarily terse, and does not explore a number of additional important and interesting questions. In particular, there may be other interesting dimensions of self-presentation that may be interesting to social scientists interested in misinformation sharing on social media. Our work also has ethical implications, which we discuss in more detail in the Ethics Statement below, as well as in the required paper checklist. Our hope, however, is that the proposed, and public, methods allow future efforts that explore new questions that link identity and behavior on Twitter and other sites with social media bios.

Acknowledgements

NM and KJ were supported by an ONR MURI N00014-20-S-F003 and by NSF IIS2145051. SM was supported by the John S. and James L. Knight Foundation through a grant to the Institute for Data, Democracy, & Politics at the George Washington University. RB was supported by the MSCA Grant Agreement No. 101073351 at GESIS Leibniz Institute for the Social Sciences.

References

- Ali, M.; Sapiezynski, P.; Bogen, M.; Korolova, A.; Mislove, A.; and Rieke, A. 2019. Discrimination through optimization: How Facebook’s Ad delivery can lead to biased outcomes. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–30.
- Altay, S.; Nielsen, R. K.; and Fletcher, R. 2022. Quantifying the “Infodemic”: People Turned to Trustworthy News Outlets during the 2020 Coronavirus Pandemic. *Journal of Quantitative Description: Digital Media*, 2.
- Antoniak, M.; Mimno, D.; and Levy, K. 2019. Narrative Paths and Negotiation of Power in Birth Stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 88:1–88:27.
- Assenmacher, D.; Sen, I.; Fröhling, L.; and Wagner, C. 2020. The End of the Rehydration Era The Problem of Sharing Harmful Twitter Research Data.
- Bailey, A. H.; Williams, A.; and Cimpian, A. 2022. Based on billions of words on the internet, people= men. *Science Advances*, 8(13): eabm2463.
- Bamman, D.; and Smith, N. A. 2014. Unsupervised Discovery of Biographical Structure from Text. *Transactions of the Association for Computational Linguistics*, 2: 363–376.
- Bär, D.; Pröllochs, N.; and Feuerriegel, S. 2023a. Finding Qs: Profiling QAnon Supporters on Parler. *Proceedings of the International AAAI Conference on Web and Social Media*, 17: 34–46.
- Bär, D.; Pröllochs, N.; and Feuerriegel, S. 2023b. Finding qs: profiling Qanon supporters on Parler. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 34–46.

- Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, 26(10): 1531–1542.
- Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) Is Power: A Critical Survey of “Bias” in NLP. *arXiv preprint arXiv:2005.14050*.
- Blodgett, S. L.; Green, L.; and O’Connor, B. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. *EMNLP’16*.
- Brashier, N. M.; and Schacter, D. L. 2020. Aging in an Era of Fake News. *Current Directions in Psychological Science*, 29(3): 316–323.
- Buss, J.; Le, H.; and Haimson, O. L. 2022. Transgender Identity Management across Social Media Platforms. *Media, Culture & Society*, 44(1): 22–38.
- Caliskan, A.; and Lewis, M. 2020. Social Biases in Word Embeddings and Their Relation to Human Cognition.
- Coppock, A.; Leeper, T. J.; and Mullinix, K. J. 2018. Generalizability of Heterogeneous Treatment Effect Estimates across Samples. *Proceedings of the National Academy of Sciences*, 115(49): 12441–12446.
- Darvin, R. 2022. Design, Resistance and the Performance of Identity on TikTok. *Discourse, Context & Media*, 46: 100591.
- DeVito, M. A.; Walker, A. M.; and Birnholtz, J. 2018. ‘Too Gay for Facebook’: Presenting LGBTQ+ Identity Throughout the Personal Social Media Ecosystem. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 44.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*, abs/1810.04805.
- Druckman, J. N.; Ognyanova, K.; Baum, M. A.; Lazer, D.; Perlis, R. H.; Volpe, J. D.; Santillana, M.; Chwe, H.; Quintana, A.; and Simonson, M. 2021. The Role of Race, Religion, and Partisanship in Misperceptions about COVID-19. *Group Processes & Intergroup Relations*, 24(4): 638–657.
- Essig, L.; and DellaPosta, D. 2024. Partisan styles of self-presentation in U.S. Twitter bios. *Scientific Reports*, 14(1): 1077.
- Ethayarajah, K.; Duvenaud, D.; and Hirst, G. 2019. Understanding Undesirable Word Embedding Associations. In *Proceedings of ACL’19*, 1696–1705. Florence, Italy: Association for Computational Linguistics.
- Field, A.; Bhat, G.; and Tsvetkov, Y. 2019. Contextual Affective Analysis: A Case Study of People Portrayals in Online #MeToo Stories. *Proceedings of the International AAAI Conference on Web and Social Media*, 13: 158–169.
- Field, A.; Blodgett, S. L.; Waseem, Z.; and Tsvetkov, Y. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of ACL’21*, 1905–1925. Online: Association for Computational Linguistics.
- Field, A.; and Tsvetkov, Y. 2019. Entity-Centric Contextual Affective Analysis. In *Proceedings of ACL’19*, 2550–2560. Florence, Italy: Association for Computational Linguistics.
- Fiske, S. T.; Cuddy, A. J.; Glick, P.; and Xu, J. 2002. A Model of (Often Mixed) Stereotype Content: Competence and Warmth Respectively Follow from Perceived Status and Competition. *Journal of personality and social psychology*, 82(6): 878.
- Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644.
- Gregorich, M.; Strohmaier, S.; Dunkler, D.; and Heinze, G. 2021. Regression with highly correlated predictors: variable omission is not the solution. *International journal of environmental research and public health*, 18(8): 4259.
- Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake News on Twitter during the 2016 US Presidential Election. *Science (New York, N.Y.)*, 363(6425): 374–378.
- Guess, A.; Nagler, J.; and Tucker, J. 2019. Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook. *Science Advances*, 5(1): eaau4586.
- Hare, M.; and Jones, J. 2023. Slava Ukraini: Exploring Identity Activism in Support of Ukraine via the Ukraine Flag Emoji on Twitter. *Journal of Quantitative Description: Digital Media*, 3.
- Heise, D. R. 1987. Affect Control Theory: Concepts and Model. *The Journal of Mathematical Sociology*, 13(1-2): 1–33.
- Horne, B. D.; Nørregaard, J.; and Adalı, S. 2019. Different Spirals of Sameness: A Study of Content Sharing in Mainstream and Alternative Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 13: 257–266.
- Hughes, A. G.; McCabe, S. D.; Hobbs, W. R.; Remy, E.; Shah, S.; and Lazer, D. M. J. 2021. Using Administrative Records and Survey Data to Construct Samples of Tweeters and Tweets. *Public Opinion Quarterly*, 85(S1): 323–346.
- Jacobs, A. Z.; and Wallach, H. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 375–385.
- Jiang, J.; Ren, X.; and Ferrara, E. 2023. Retweet-bert: political leaning detection using language features and information diffusion on social networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 459–469.
- Johnstone, B. 1996. *The Linguistic Individual: Self-expression in Language and Linguistics*. Oxford University Press.
- Joseph, K.; and Morgan, J. H. 2020. When Do Word Embeddings Accurately Reflect Surveys on Our Beliefs about People? In *Proceedings of ACL’20*, 4392–4415.
- Joseph, K.; and Morgan, J. H. 2021. Friend or Foe: A Review and Synthesis of Computational Models of the Identity Labeling Problem. *The Journal of Mathematical Sociology*, 1–35.
- Kozłowski, A. C.; Taddy, M.; and Evans, J. A. 2019. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, 84(5): 905–949.

- Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring Bias in Contextualized Word Representations. *arXiv preprint arXiv:1906.07337*.
- Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. 2018. The science of fake news. *Science*, 359(6380): 1094–1096.
- Li, J.; Longinos, G.; Wilson, S.; and Magdy, W. 2020. Emoji and Self-Identity in Twitter Bios. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 199–211.
- Lin, H.; Lasser, J.; Lewandowsky, S.; Cole, R.; Gully, A.; Rand, D.; and Pennycook, G. 2022. High level of agreement across different news domain quality ratings.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Lucy, L.; Tadimetri, D.; and Bamman, D. 2022. Discovering Differences in the Representation of People Using Contextualized Semantic Axes. In *Proceedings of EMNLP'22*.
- MacKinnon, N. J.; and Heise, D. R. 2010. *Self, Identity, and Social Institutions*. Palgrave Macmillan.
- Martel, C.; Rathje, S.; Clark, C. J.; Pennycook, G.; Van Bavel, J. J.; Rand, D. G.; and van der Linden, S. 2024. On the Efficacy of Accuracy Prompts Across Partisan Lines: An Adversarial Collaboration. *Psychological Science*, 35(4): 435–450.
- Marwick, A. E.; and boyd, d. 2011. I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. *New Media & Society*, 13(1): 114–133.
- McCabe, S.; Green, J.; Wan, A.; and Lazer, D. 2022. New TweetScores: Or, Did Donald Trump Break tweetScores?
- Mendelsohn, J.; Tsvetkov, Y.; and Jurafsky, D. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3: 55.
- Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.
- Miller, G. A.; and Charles, W. G. 1991. Contextual Correlates of Semantic Similarity. *Language and cognitive processes*, 6(1): 1–28.
- Moore, R. C.; Dahlke, R.; and Hancock, J. T. 2023. Exposure to Untrustworthy Websites in the 2020 US Election. *Nature Human Behaviour*, 1–10.
- Mosleh, M.; Pennycook, G.; and Rand, D. G. 2020. Self-Reported Willingness to Share Political News Articles in Online Surveys Correlates with Actual Sharing on Twitter. *PLOS ONE*, 15(2): e0228882.
- Mu, Y.; and Aletras, N. 2020. Identifying Twitter users who repost unreliable news sources with linguistic information. *PeerJ Computer Science*, 6: e325.
- Ng, L. H. X.; and Cruickshank, I. J. 2023. Recruitment Promotion via Twitter: A Network-centric Approach of Analyzing Community Engagement Using Social Identity. *Digital Government: Research and Practice*, 4(4): 22:1–22:17.
- Nguyen, D.; Trieschnigg, D.; Dođruöz, A. S.; Gravel, R.; Theune, M.; Meder, T.; and de Jong, F. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *COLING 2014*, 1950–1961.
- Nguyen, D. Q.; Vu, T.; and Tuan Nguyen, A. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the EMNLP'20*, 9–14.
- Nikolov, D.; Flammini, A.; and Menczer, F. 2021. Right and Left, Partisanship Predicts (Asymmetric) Vulnerability to Misinformation. *Harvard Kennedy School Misinformation Review*.
- Osmundsen, M.; Bor, A.; Vahlstrup, P. B.; Bechmann, A.; and Petersen, M. B. 2021. Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *American Political Science Review*, 115(3): 999–1015.
- Pathak, A.; Madani, N.; and Joseph, K. 2021. A Method to Analyze Multiple Social Identities in Twitter Bios. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–35.
- Phillips, S. C.; Carley, K. M.; and Joseph, K. 2024. Why Do People Think Liberals Drink Lattes? How Social Media Afforded Self-Presentation Can Shape Subjective Social Sorting. *arxiv:2404.02338*.
- Rashkin, H.; Choi, E.; Jang, J. Y.; Volkova, S.; and Choi, Y. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2921–2927.
- Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *ArXiv*, abs/1908.10084.
- Rogers, N.; and Jones, J. J. 2021. Using Twitter Bios to Measure Changes in Self-Identity: Are Americans Defining Themselves More Politically Over Time? *Journal of Social Computing*, 2(1): 1–13.
- Schröder, T.; Hoey, J.; and Rogers, K. B. 2016. Modeling dynamic identities and uncertainty in social interactions: Bayesian affect control theory. *American Sociological Review*, 81(4): 828–855.
- Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823.
- Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1): 22–36.
- Smith-Lovin, L.; and Douglas, W. 1992. An Affect Control Analysis of Two Religious Subcultures. *Social perspectives on emotion*, 1: 217–47.
- Tajfel, H.; and Turner, J. C. 1979. An Integrative Theory of Intergroup Conflict. In *The Social Psychology of Intergroup Relations*, 33–47. Monterey, CA: Brooks/Cole, w austin & s. worche edition.
- Wilkerson, J.; and Casas, A. 2017. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20: 529–544.

Yoder, M. M.; Shen, Q.; Wang, Y.; Coda, A.; Jang, Y.; Song, Y.; Thadani, K.; and Rosé, C. P. 2020. Phans, Stans and Cisheds: Self-Presentation Effects on Content Propagation in Tumblr. In *12th ACM Conference on Web Science*, 39–48.

Zeng, L.; Dailey, D.; Mohamed, O.; Starbird, K.; and Spiro, E. S. 2019. Detecting journalism in the age of social media: three experiments in classifying journalists on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 548–559.

Zhang, Y.; Lukito, J.; Su, M.-H.; Suk, J.; Xia, Y.; Kim, S. J.; Doroshenko, L.; and Wells, C. 2021. Assembling the Networks and Audiences of Disinformation: How Successful Russian IRA Twitter Accounts Built Their Followings, 2015–2017. *Journal of Communication*, 71(2): 305–331.

Paper Checklist

One can crudely separate the way that members of the NLP community acknowledge sociocultural associations embedded in language into three camps. In the first, acknowledgement is largely non-existent, with scholars either choosing to ignore the social implications of language or arguing that it is not within their purview to address them. In the second, scholars concerned with the potential that NLP tools might reify biases in language have sought to develop tools to reduce or remove the use of biased tools in production. Example efforts include the literature on “debiasing” embedding spaces—where scholars aim to remove meaning on “unwanted dimensions” (i.e. on social dimensions of meaning) and keep it only on “objective” dimensions (i.e. on more traditional semantic dimensions)— and work to carefully elaborate how problems can arise from NLP tools that do not critically engage with the social implications of language (Blodgett et al. 2020; Field et al. 2021).

While we strongly support the latter line of critical scholarship, we fall into a third camp of NLP researchers, who see the potential for NLP methods that capture social meaning in language to help inform our understanding of society and its behaviors (Kozlowski, Taddy, and Evans 2019; Bailey, Williams, and Cimpian 2022; Garg et al. 2018). As individuals who are supportive of existent critical literature, and who are still building methods that are inherently engaging in stereotyping, we must accept that our methods can be used not only to measure perceptions and self-presentations but also to apply them for nefarious purposes, such as racially-biased targeted advertising (Ali et al. 2019). The ultimate ethical question that we must address, then, is whether we believe that our measurement strategies will ultimately do more harm than good.

Our decision to submit the present work for consideration reflects our belief, after deliberation, that the potential benefits of our work outweigh the potential dangers. From a benefits perspective, we show that our work can help us understand the (mis)information environment on Twitter, and hope in the future to use these methods to better understand 1) how hegemonic voices are often given outsized attention on social media platforms, and 2) how attacks on marginalized communities changed along with the political climate in the United States. From a misuse perspective, we do not

believe that the methods we have developed here are more effective at targeting individuals based on stereotyped demographics than the tools already in existence elsewhere, tools which use orders of magnitude more compute data and power. As such, while there are real dangers with being misclassified in a particular way online, we 1) emphasize throughout the paper that we focus only on how someone is likely to be perceived based on what is in their bio, and do not claim to be able to classify some “true self” of users, and 2) expect that if one wanted to engage in such a prediction activity, it could be done better with methods that focus specifically on this task.

A similar value judgement was applied to the other ethical question in our work: whether or not to publicly release parts of our data. Ultimately, restrictions on the Decahose have made it challenging to share data. However, given recent literature in the ICWSM community (Assenmacher et al. 2020), we believe the appropriate course of action is to release training data on a per-case basis, as potential dataset users contact the last author of this paper. However, we believe that the release of our (fully deidentified) survey data and basic case study data does not serve any inherent risks for users in our dataset, and thus have opted to do so in the Github repository for this work. We believe that the utility of these data to the scientific community, both for replicability and extension, outweigh the dangers of data release. Of course, in all cases, data and methods usage is a continually evolving process, and we will re-evaluate this value judgement as often as is needed.

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? [As noted above, there are privacy concerns in our work that we have carefully considered and weighed relative to the benefits of our work.](#)
- (b) Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes](#)
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? [Yes](#)
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? [Yes](#)
- (e) Did you describe the limitations of your work? [Yes](#)
- (f) Did you discuss any potential negative societal impacts of your work? [Yes, see the Ethics Statement](#)
- (g) Did you discuss any potential misuse of your work? [Yes, see the Ethics Statement](#)
- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? [Yes, see the Ethics Statement](#)
- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes](#)

2. Additionally, if your study involves hypotheses testing...
 - (a) Did you clearly state the assumptions underlying all theoretical results? *NA*
 - (b) Have you provided justifications for all theoretical results? *NA*
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? *NA*
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? *NA*
 - (e) Did you address potential biases or limitations in your theoretical framework? *NA*
 - (f) Have you related your theoretical results to the existing literature in social science? *NA*
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? *NA*
3. Additionally, if you are including theoretical proofs...
 - (a) Did you state the full set of assumptions of all theoretical results? *NA*
 - (b) Did you include complete proofs of all theoretical results? *NA*
4. Additionally, if you ran machine learning experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [See the Github link in the first footnote](#)
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [See the main text as well as the appendix below](#)
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes](#)
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes](#)
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? [Yes. Specifically, we use a number of clear, competitive, and comparable baselines](#)
 - (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? [Yes, see the Ethics statement.](#)
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
 - (a) If your work uses existing assets, did you cite the creators? [Yes.](#)
 - (b) Did you mention the license of the assets? [To the best of our ability, we believe that nothing we have provided requires this, but we will update as necessary.](#)
 - (c) Did you include any new assets in the supplemental material or as a URL? [Yes.](#)
 - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [We provide details on the dataset, acknowledging that consent was approved in surveys as IRB requires. Social media users are non-consenting, which we acknowledge in our Ethics statement.](#)
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes, see the ethics statement](#)
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? [In progress for the camera ready version](#)
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? [In progress for the camera ready version.](#)
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
 - (a) Did you include the full text of instructions given to participants and screenshots? [We use a protocol from a prior work, which we do reference.](#)
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? [Yes, our studies are IRB approved.](#)
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes.](#)
 - (d) Did you discuss how data is stored, shared, and de-identified? [Yes.](#)

Appendix

Bios from the Twitter Decahose

The center column of Table 1 provides summary statistics for the Twitter bio data we construct. We begin with a sample of 15,459,872 distinct Twitter bios from users who posted a tweet in 2020 that was found in the Decahose, and who are specified as English-language users by the Twitter API. In order to maintain a focus on culturally-salient identities, we limit the size of the vocabulary to identities used in at least 100 unique Twitter bios in the training set. Further, because we are interested in associations between identities, we further remove Twitter bios that contain less than 2 identities. After these cleaning steps, our training and test data consists of 3,534,903 and 1,546,001 distinct bios respectively with 22,516 unique identities in the vocabulary. We then follow the approach outlined in the main text to produce the main test dataset and the generalizability test set. Note that the size of each of these splits can be larger than the size of cleaned test dataset, because we can generate multiple instances from a given bio by randomly selecting different targets; i.e. we can generate multiple test instances out of each of profile description by selecting multiple pairs of X_r^i and X_t^i .

Finally, to provide further insight into the data, Table 2 showcases the top 7 identities in terms of overall frequency in the training data and 7 of the least frequent identities to show that the tail still contains meaningful phrases.

Dataset Portion	Number of distinct records in Twitter
All raw records	15,459,872
Initial training cut	12,367,897
Training dataset after cleaning	3,534,903
Initial test cut	3,091,975
Test dataset after cleaning	1,546,001
<i>main</i> test dataset	3,044,093
<i>generalizability</i> test dataset	395,583
Vocabulary (distinct phrases)	22,516

Table 1: Summary statistics for the identity-centric datasets we develop

identity	Number of times appeared
she	352,655
her	308,829
he	144,845
him	144,845
they	353,4903
writer	67,824
blm	63,388
mixer streamer freak	100
published photographer	100
sophomore	100
micah 6:8	100
public health specialist	100
britishindependence	100
vikings fan	100

Table 2: Examples of the most and (some of the) least frequent identities in the Twitter dataset

Single Identity Survey Data

A complete, aggregated copy of this data is provided in the Github for this paper. Our survey study was ruled exempt by the IRB at [REMOVED]. Each respondent rated between four and seven identities, and each identity was given to at least 3 respondents. Respondents were paid an average of \$12/hour. Respondents are from a convenience sample, as recent work has suggested that the cost efficiency of convenience samples does not necessarily impact data quality (Coppock, Leeper, and Mullinix 2018). The median age of our sample is 32. Of the 140 respondents, 88 reported their sex as female, 49 as male, and 4 noted other/did not provide. Our sample, like Twitter, was overwhelmingly White; 105 (75%) of the sample reported as White.

We here provide two minor additional details on our survey data. First, it is of note that in contrast to prior work, we focus explicitly on priming respondents to think of social media users, asking, for example, “If you saw [identity] in a social media biography, would you expect that individual to be” and then provided a Likert scale ranging from (e.g.) “Always [a] Democrat” to “Always [a] Republican.” Second, we emphasize that other procedures, including tutorial materials, task details, and attention checks, follow the publicly

available materials from Joseph and Morgan (2020).

Entire Bio Survey Data

A complete, aggregated copy of this data is provided in the Github for this paper.

We use responses from approximately 730 Prolific respondents. Because of a temporary issue with the survey, some respondents were approved for pay without taking the study, we report demographics here on all respondents here because of this issue. The median age of respondents was 37. Reported sex was more balanced than the sample for the single identity study, 51% and 44% of respondents reported a sex of female or male, respectively. As above, however, a majority of the sample (65%) reported as White. Respondents were paid an average of \$12/hour. Full details are provided in the Appendix for the interested reader.

Two final notes are in order. First, with respect to the difference between the two surveys is that because we focus on individual user bios, we in this study ask respondents to assess the likelihood that the individual user who has this bio is, e.g., a Democrat or Republican, rather than asking (as above) about the probability that an individual who holds a given identity is, e.g., a Democrat or Republican. Second, we note that we drop the 5% of respondents whose responses were furthest from all other respondents, on average, across all identities and dimensions (after standardizing measures across dimensions). This results in a final sample of 1,273 bios analyzed here and provided in the Github repository.

Case Study

Replication with Data from the Decahose We replicate findings using the case study dataset analyzed in the main text with a different dataset, broader in scope, drawn from the Decahose. **A complete, anonymized copy of this replication data is provided in the Github for this paper.** For each user, it contains the number of URLs they shared that link to low- and high-quality domains, as identified by NewsGuard, and projections of their bios onto the four dimensions of interest to the case study.

We begin with a sample of roughly 290,000 users who tweeted at least one NewsGuard URL in the decahose in 2020. Of this sample, 143,883 users were 1) still active when we recollected tweets in 2022, and 2) had a non-empty English-language bio. It is possible that our use of

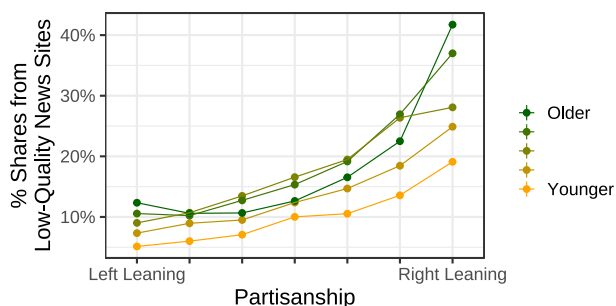


Figure 7: Replication of Figure 6A) with a dataset drawn from the Decahose

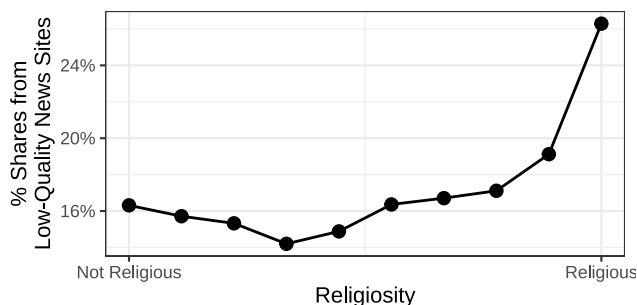


Figure 8: Replication of Figure 6B) with a dataset drawn from the Decahose

langdetect could bias our sampling of what is considered English language (Blodgett, Green, and O’Connor 2016); as such, we manually evaluate it. To do so, we had three research assistants label 900 bios as English, not English, or Vague. The annotators had a Krippendorff’s alpha agreement score of 0.81 and agreed on 90% of the bios with the library. More specific to our case study, 93% of the bios that langdetect measured as English were also annotated as English by all the annotator. The main reason for mis-classification of the bio’s language was that some bios consisted of phrases from multiple languages. We therefore believe that this step did not bias our results in any obvious way.

Of these remaining 143,883 users who matched our initial sampling criteria, a remaining 108,554, or 75.4%, had at least 5 shares of URLs in NewsGuard, which we considered a minimum for estimating proportions. The median user had 49 shares of high quality news website URLs and 3 shares of low quality news website URLs. In total, these users shared 11,735,521 links to NewsGuard domains we assess in our study.

Figure 7 replicates findings from Figure 6A) in the main text, namely, observe a similar increase in the association between self-presented (older) age and misinformation sharing as we see an increase in right-leaning self-presentation. Figure 8 replicates Figure 6B).

Descriptive Statistics Partisanship and Religiosity Figure 9 shows a descriptive statistic that provides some ev-

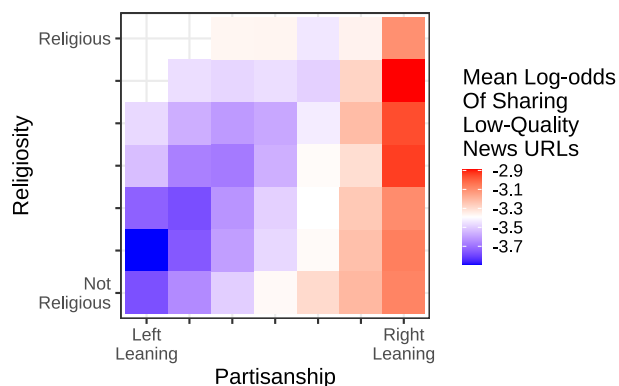


Figure 9: The mean log-odds of sharing a link from a low quality news site, relative to a high quality news site (color of each cell) for panel members with given levels of partisanship (x-axis) and religiosity (y-axis).

idence that religiosity and partisanship have an interactive effect on misinformation sharing. However, as noted in the main text, correlations between these variables make it challenging to assess the statistical significance of these variables relative to each other; in particular, there are few left-leaning accounts that also present as religious.

Modeling

Distributional Semantic Models (Baselines)

For all DSM baseline models except Sentence Bert, including BERT-base, RoBERTa-base and BertTweet-base we experimented using open-source implementations on Hugging Face transformers library⁷. For the Sentence-Bert baseline, we used the *mpnet-base* pre-trained model and the implementation given by open-source Sentence Transformers library.⁸

Bio-only Model To select hyperparameters, we use 10% of the training data as a validation dataset. The primary hyperparameter we tuned was whether to use a Skip-Gram or C-BOW model. We ultimately chose a Skip-Gram model for Twitter and a C-BOW model for Wikipedia, with the other hyperparameters as specified in the main text. Model training took under an hour on a personal laptop. We used the open-source implementation of word2vec in gensim⁹ for our experiments.

Fine-tuned SBERT We here briefly provide additional intuition for our Fine-tuned SBERT model. Given a dataset of pairs of identity phrases with a label denoting the anchor-positive pair or anchor-negative pairs, we input the pair into the pipeline shown in Figure 10 and extract the latent embeddings of each of the identities. Then we calculate the cosine similarity of the embeddings and backpropagate the mean squared error loss with the label through the network. In

⁷<https://huggingface.co/docs/transformers/>

⁸<https://github.com/UKPLab/sentence-transformers>

⁹<https://radimrehurek.com/gensim/>

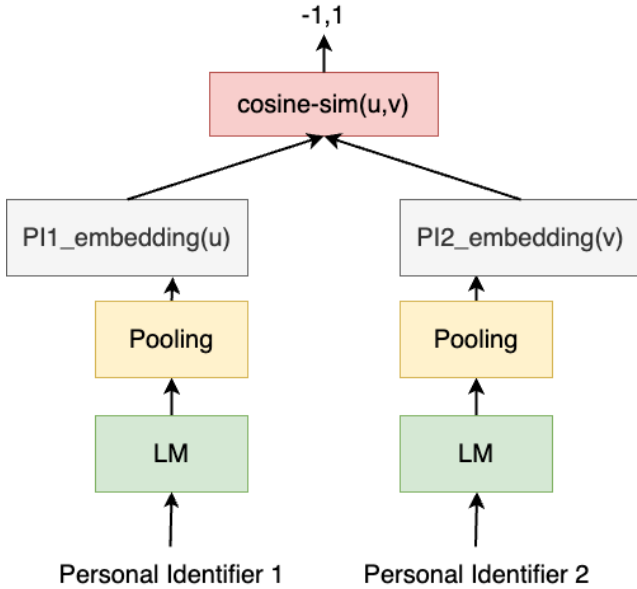


Figure 10: Training procedure for contrastive learning with regression objective function.

this work we fine-tuned a *mpnet-base* Sentence Bert model since it had the same model size as a BERT-base and the pre-trained model was available online.

Further details on Baseline Models For all DSM baseline models except Sentence Bert, including BERT-base, RoBERTa-base and Bertweet-base we experimented using open-source implementations on Hugging Face transformers library¹⁰. For the Sentence-Bert baseline, we used the *mpnet-base* pre-trained model and the implementation given by open-source Sentence Transformers library.¹¹

Other Baseline Models Considered In addition to the three baseline models discussed in the text, we also experimented with a pair of other sensible options.

First, we expected that a DSM pretrained on Twitter would be a strong baseline to compare to, and thus experimented with additional models pre-trained specifically on Twitter data (Nguyen, Vu, and Tuan Nguyen 2020). We use the fine-tuned BERT model on Twitter data proposed by (Nguyen, Vu, and Tuan Nguyen 2020). They propose a BERT-base model fine-tuned using a corpus of 850M English Tweets. However, we find that model performance was no better than the other, more widely used baseline DSMs we proposed in the main experiments.

Second, it seemed reasonable that by first restricting a baseline DSM to known dimensions of social meaning, we could improve their performance. Consequently, we considered baselines where we first projected down all baseline models into the core dimensions of meaning noted by Joseph and Morgan (2020) before the evaluation tasks. In both cases, however, our intuitions did not match empirical

¹⁰<https://huggingface.co/docs/transformers/>

¹¹<https://github.com/UKPLab/sentence-transformers>

reality. These models failed to outperform the baselines used in the main text, and thus we restrict our analysis to the baselines discussed in the main text.

Generating Embeddings for the Predictive Experiment

In order to build inputs to the network, since X_r^i is a list of personal identifiers, to calculate the latent embedding L_r^i for it, depending on the model, we follow different procedures. For the *Bio-only* model, we simply measure the average latent vector of all phrases in X_r^i according to (1). For the Fine-tuned models, as well as the baseline contextualized language models discussed below, we stitch the words in X_r^i together with comma and create a sentence S_r^i . We then measure L_r^i according to Equation (2). Equivalently, this means that for the BERT based models we take the embedding of [CLS] token for pooling and for the Sentence Bert based models we follow the original work and take the average of all token embeddings.

$$L_r^i = \sum_{v \in X_r^i} (v) / |X_r^i| \quad (1)$$

$$L_r^i = \text{Pooling}(LM(S_r^i)) \quad (2)$$

Projection to Social Dimensions

In order to project a piece of text (either a full bio or a single identity) onto a specific dimension, we have to first define the end-points of that dimension. Table 3 and 4 outline the original and in-domain dimension end-points that we talk about in this paper, in particular in our Entire bio projection evaluation. To generate embeddings for an end-point, we assume each end-point is an instance in X and follow the approach outlined above to generate embeddings for each end-point of the dimension d and call them E_d^1 and E_d^2 respectively. Having the embedding of both poles of the dimension, we calculate the difference vector according to Equation 3, and calculate the embedding of the target text using the same approach to a vector L^i . Then we follow the projection approach outlined in (Ethayarajh, Duvenaud, and Hirst 2019) to normalize all vectors and then calculate the projection value of L^i onto dimension d according to equation 4.

$$V^d = E_2^d - E_1^d \quad (3)$$

$$P_d^i = \frac{V^d \cdot L^i}{\|L^i\|} \quad (4)$$

Error Analysis of The Predictive Task

Our understanding of the proposed models is improved by studying where errors occur. Here, we briefly present both quantitative and qualitative reflections on the major sources of error for the Bio-Only and Fine-tuned SBERT models. At a high level, we find that pre-training helps the Fine-tuned SBERT model on predictions requiring knowledge of phrase composition (e.g. that “mother” and “mother of two” convey similar meanings), but appear to cause it to struggle on infrequent identities, which the Bio-only model is better able to capture meanings of from the bio data alone.

Dimension	End point 1	End point 2
Age	young, new, youthful	old, elderly, aged
Partisanship	democratic party supporter, left-leaning, democrat	republican party supporter, right-leaning, republican
Religion	atheistic, agnostic, non-believing, skeptical	religious, faithful, christian, believe in lord
Politics	music, sports, culture, tech	politics, political, democrat, republican
Gender	mother of, grand mother	father of, grand father

Table 3: Dimension endpoints for “Original” dimension endpoint selection method

Dimension	End point 1	End point 2
Age	15 years old, 18 years old, sophomore in college, student at, umich22, 18, 21	retired person, I’m old, 50 years old, 65 years old, 61yr old, grandparent of, old man, old woman, grandma to, grandpa to, tenured, long career
Partisanship	pro socialism, liberal democrat, never trump, proud democrat, vote blue no matter who, #resist, #voteblue, #nevertrump, left leaning, #democratidownballot, #notvotablenuomatterwho, #bidenharris, #resist, #bluewave, #democraticsocialist	right leaning, trump won, never biden, fuck biden, #maga, #kag, Trump conservative, conservative and America First, proud Trump supporter, trump fan, #MAGA Republican, constitutional conservative patriot, #trump2024
Religion	atheist, nonbeliever, proud atheist, totally secular, #cancelreligion	Catholic, jesus christ, follower of christ, priest, lover of jesus, christian episcopalian, jesus loving christian, john 3:16, gospel of the lord jesus christ, minister at united church, christ-follower, god first, isaiah 55:6, woman of faith, man of faith
Gender	sister, wife, mother, Proud Mama and Wife, grandmother of, mother of one, mama of one, wife of, Loving Wife, she, her, hers	husband to, brother, husband, father, grandfather of one, father of one, Loving husband, he, him, his, son, brother, brother-in-law, uncle, nephew

Table 4: Dimension endpoints for “In-Domain” dimension endpoint selection method

Quantitatively, Figure 11 shows that both models performed best, and roughly equally well, on the most frequent identities, but that differences appeared in how the models fared elsewhere. The Bio-only model’s ranking distribution (the marginal density plot on the right-hand side of Figure 11) was bimodal, with a large number of high (poor performance) and low (strong performance) ranks for test points. Perhaps unsurprisingly, we find qualitatively that the poor performance of the Bio-only model relative to the Fine-tuned SBERT model largely came from an inability 1) to learn from compositional identities or 2) to leverage relevant external knowledge. These issues seemed to impact the model most for moderately frequent target identities, those appearing between 300-10,000 times in the training data. With respect to 1), for example, when provided the Twitter bio “mother of two, restaurant owner, partly retired, hockey coach”¹², the Bio-only model ranks the correct held-out identity, “wife,” among the least likely. In contrast, the Fine-tuned SBERT model correctly ranks “wife” in the Top 1%. The core difference is that the Fine-tuned SBERT model, but not the Bio-only model, leverages the gender stereotype implied by the “mother” portion of the phrase “mother of two.” With respect to 2), there were several cases where external knowledge from the pre-trained model benefited the Fine-

tuned models. For example, the Fine-tuned models, but not the Bio-only models, were able to recognize the similarity between the identities “follower of ISKSON” (a Hindu religious organization) and “proud Hindu.” Both of these were relatively infrequently used.

In contrast, relative to the Bio-only model, Fine-tuned models struggled with the *most infrequent* identities, in particular the roughly 18% of identifiers in the test set that occurred fewer than 300 times in the training data. In these cases, as in prior work entity-centric domain adaptive work (Field and Tsvetkov 2019), the Fine-tuned models seemed to rely too heavily on knowledge from the pre-trained model and not enough to domain-relevant context. In contrast, the identity-centric model seemed to benefit on the prediction task from overfitting to stereotypical knowledge for these rarer phrases. The Fine-tuned models also struggled when presented with identities, such as Twitter-specific acronyms, that were likely rare in the DSM data, but more frequent on Twitter. Here, pre-training seemed to induce noise, leading the Fine-tuned models to predict somewhat randomly.

¹²This bio has been modified to protect user privacy

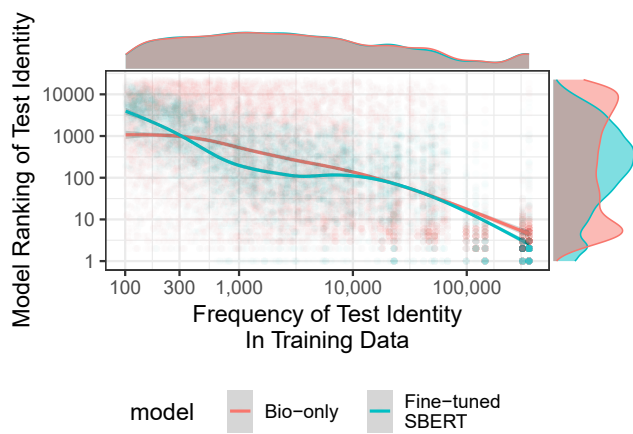


Figure 11: Average rank (y-axis) versus the frequency of that phrase in training dataset (x-axis) for 10K sample test points (each overplotted point) for the Bio-only and Fine-tuned SBERT models. Fit line is a generalized additive model; marginal densities are also displayed.