

Measuring Social and Political Identities In Social Media Self-Descriptions

Clara Vandeweerd^{1,*}, Gregory Eady^{1,2}, Frederik Hjorth^{1,2}, and Peter Thisted Dinesen²

¹Copenhagen Center for Social Data Science (SODAS), University of Copenhagen

²Department of Political Science, University of Copenhagen

*Corresponding author: clara.vandeweerd@ifs.ku.dk; Øster Farimagsgade 5, 1353 Copenhagen, Denmark

March 12, 2026

Abstract

Identities are fundamental to our understanding of social and political behavior, but are challenging to measure and are rarely observed in real-world settings. We introduce a method for measuring the identity-relevant aspects of brief self-descriptions regularly used online (e.g. on social media). Our approach combines the benefits of word embeddings for finding related identity terms with the ability of clustering algorithms to aggregate terms into discrete categories. To illustrate our approach, we apply it to daily observations of bios from millions of US Twitter/X users. We present three applications of our approach with substantive findings. First, we track users' social and political identities over time and find, among other things, that direct expressions of political affiliations are rare. Second, we map the identities that are most characteristic of each US state. Third, we show that users' political identities are highly predictable based on non-political identity markers. With the growing availability of user self-descriptions on social media platforms and elsewhere, our approach enables researchers to map and analyze expressions of identity at scale.

Keywords: identity, text-as-data, word embeddings, unsupervised learning, social media

1 Introduction

What political and social identities are socially prominent, and how do these identities change over time? Given the centrality of identities to socio-political cognition (Converse, 1964; Tajfel and Turner, 1979; Conover, 1984), this is an essential question for social science. Yet, given the multi-faceted nature and time-varying salience of identities, mapping and measuring identities remains methodologically challenging.

In recent years, a novel data source has emerged as a potential alternative to traditional, non-behavioral survey-based measures of identities: self-descriptions (or ‘bios’) that users enter on social media into open-text fields designed for self-presentation. Social media bios have the advantage of being spontaneously written self-presentations, which people are able to adapt over time. Presenting oneself to others is an explicit goal for many social media users. For instance, besides passive uses of the platform such as consuming news or entertainment, the most-cited motivation for using Twitter/X (indicated by 31% of users) is “to tell others what I am doing and thinking about” (American Press Institute, 2015). Seen in this way, social media profiles provide a naturalistic, *behavioral* measure of identity expression. Moreover, they can be collected and re-collected over time at little or no extra cost—an important feature in light of recent research showing that even self-identifications such as ethnicity, sexual orientation, and religion are more changeable than previously assumed (Agadjanian, 2022; Egan, 2020; Margolis, 2018).

These self-descriptions are an important venue for identity expression, not least because self-description is possible or even required in many aspects of digital life. Some of these, such as Twitter/X on which our running example relies, have tightened researcher API access in recent years (Freelon, 2018), but others remain accessible (Tromble, 2021), and still others have recently expanded researcher API access (Corso et al., 2024). In the section on ‘Collection’ below, we elaborate on opportunities for obtaining data on self-descriptions.

Recent social science research has used self-descriptions from social media bios to measure changes in expressions of group identities that are pre-defined by researchers (Eady

et al., 2022; Jones, 2023; Rogers and Jones, 2021). The most common method for doing so is to manually create a dictionary of terms that reflect a particular social identity. The downside is that this does not necessarily capture all relevant terms, is essentially restricted by researchers’ preconceptions about identities, and can only realistically measure a limited number of identities.

In this paper, we introduce a new method for more expansively measuring expressions of social identities, political identities and other identity-relevant traits from online self-descriptions, with far greater flexibility than dictionary-based approaches. We use word embeddings and clustering to group all of the meaningful words used in self-descriptions into categories such as “Mother” or “Christian”. The method is nearly completely automated, and therefore scalable at little cost, while also being transparent. Importantly, as an unsupervised approach, it is also able to discover new identity categories from the ground up. The method is able to capture identity in a naturalistic way, at massive scale, and over time. We illustrate our approach using a running implementation example, in which we monitor over-time changes in the self-identifications of about 2.6 million US Twitter/X users from mid-2020 to mid-2023, across dozens of identity categories that arise from the data itself.

We proceed as follows. In the next section, we define “identity” and provide an overview of existing methods for measuring identities from self-descriptions. We compare our approach both to earlier methods for measuring identity, and to other potential (computational) approaches, and describe which types of research questions our method is particularly fit to answer. We also motivate our particular focus on how our method can measure *political* identities. We proceed to outline our method, including a detailed walk-through of a specific implementation. We then present results from a validation exercise, demonstrating that our method produces group assignments that approximate those of human annotators. We then briefly present three applications of the method, showcasing its potential: we track identities over time, find the most characteristic identity of each US state, and finally, find the identities that are most predictive of a user’s political affiliation. Lastly, we discuss the method’s

strengths and limitations.

2 Defining and measuring identity

Identity is a notoriously difficult concept to pin down. In this paper, we use the term identity as short-hand for self-identification: the way that we categorize or narrate ourselves, typically with other people as the audience (Brubaker and Cooper, 2000). The content of self-identifications can be a mix of three types of “selves”: the individual self, the relational self, and the collective self (Sedikides et al., 2011). The first category includes characteristics such as traits, goals, and behaviors, which set us apart from other people. The second category consists of our roles in interpersonal relationships. The third category—also known as “social identity” (Ashmore et al., 2004)—contains attributes that we share with in-group members, and that differentiate us from out-groups. We aim to capture identities from all three categories, and generally take an expansive view of what makes for an identity-relevant trait (more on this in section 3.3).

There is some overlap between these types of self-identification. Understanding oneself as having some personal trait or interpersonal role often means acknowledging that one is a member of the social group of people with this same characteristic. For example, parenthood is understood as a social identity by social scientists (e.g. Diamond 2020; Elder and Greene 2012; Klar 2013). Ideological identities such as “liberal” can be considered an individual, value-based identity. However, they are closely related to partisan identities such as “Democrat”, which are typically thought of as social identities (Huddy et al., 2015).

The identities that a person mentions in a social media bio are, of course, not a complete list of his or her self-identifications. Instead, we can think of these social media bios as self-identifications that pass two criteria. First, they must be central enough to be mentioned, given that people have limited space in their profiles (160 characters, for example, in the case of Twitter/X), and limited time and attention to use on crafting a bio. Central identities are

ones that people rate as subjectively important, and that are chronically salient to them—meaning they will be available when people try to give a top-of-the-head self-description (Ashmore et al., 2004; Leach et al., 2008).

Second, the user must decide that a given identity is appropriate (perhaps even advantageous) to mention in a public social media profile that is frequently coupled with at least some personal identifying information (Cappos et al., 2017). Identities that are controversial or taboo are therefore much less likely to be mentioned—especially by the 19% of users who cite “networking” as one of their top reasons for using the platform (American Press Institute, 2015). For that reason, though our aim is to study identities, what we are able to measure is perhaps more accurately labeled as “identity signals”—the part of a person’s identity that they are interested in expressing publicly.

2.1 Earlier approaches to identity measurement

Traditional social science has converged on two measurement approaches for identities. The first, and most common, uses closed-ended survey questions that ask respondents about, for example, their ethnicity, gender, or partisanship. Creative variants of the closed-ended approach include questions that allow respondents to allocate “points” to different identity options within a category—to capture the intensity of various identifications (Lee, 2009). The downside of this approach is that it is difficult to capture rare identities, or identity categories that researchers have not anticipated. Researchers have consistently struggled, for example, to develop categories for closed-ended survey questions on ethnicity (Brown and Langer, 2010).

The second approach to measuring identity is using open-ended survey questions, where respondents are asked to list identities that are important parts of their self-image (e.g. Ethier and Deaux 1994; Reid and Deaux 1996). A typical example is the Twenty Statements Test, where respondents are asked to complete the sentence “I am...” many times over (Kuhn and McPartland, 1954). Human coders may then group answers into overarching categories.

These questions have the benefit of not imposing categories that are less-than-accurate fits for respondents' self-perceptions. At the same time, such survey items can be difficult for respondents to understand, requiring examples to clarify the question. These examples, in turn, can heavily influence responses (Kinder et al., 1989). It is also plausible that answers will be influenced by other elements of the research context, such as previous survey questions or the apparent purpose of the research. In addition, this method is difficult to scale up due to the labor involved. This may explain why open-ended survey questions about identities are rarely used in large-n studies of social or political behavior.

2.2 Mass measurement of identity from self-descriptions

Since these survey methods for measuring identity were invented, potential data sources for doing identity research have expanded. So far, a few studies have tracked the appearance and disappearance of specific identity terms in user self-descriptions on social media. For example, Rogers and Jones (2021) analyze bios from 2015 to 2018 on Twitter/X, and find that mentions of political keywords were on the rise, surpassing mentions of religion in the final year. Tucker and Jones (2023) describe the changing prevalence of pronouns in Twitter/X bios, and Hare and Jones (2023) document the use of the Ukrainian flag in bios and names that spiked after the Russian invasion of Ukraine in 2022. Finally, Eady et al. (2022) find that Twitter/X users deleted Republican identity terms from their bios after the attack on the US Capitol on January 6th, 2021.

All of the studies above use dictionary approaches to tackle this problem. They specify between one and a dozen or so keywords that measure an identity of interest, and track their occurrence in bios. Alternatively, hand-coding approaches are sometimes used (Pew Research Center, 2022). Hand coding has a potential benefit of high validity, but it is also expensive to apply at scale or over time. Recent research has attempted to overcome this problem by using hand-coded content to train models that attempt to classify the identities present in a self-description. These could then be applied to far more data. Hopkins et al. (2024)

achieve reasonable performance with a fine-tuned BERT classifier detecting whether or not a tweet contains a reference to gender identity, political identity, and racial/ethnic identity (but unsatisfactory performance for religious identity). Because this supervised classification approach still relies on a large hand-labeled set of self-descriptions, realistically, it can only be used to detect a handful of identity categories at a time. Moreover, their feasibility for correctly assigning users to identity categories—as opposed to detecting whether or not a dimension of identity is being mentioned—has not been demonstrated yet.

These approaches—dictionary methods, hand-coding, and training a classification model—have been used by researchers looking to measure the prevalence or correlates of some pre-defined identity categories. Also known as rule-based and supervised approaches, they are a good fit for research questions that focus on specific, known identities (Baden et al., 2022). Other questions, on the other hand, are better answered using unsupervised methods that allow for bottom-up discovery of identity categories—such as the method that we introduce in this paper. In particular, it would be challenging for researchers both to come up with all possible identity categories in advance, and to judge which categories are frequent enough in the data set to be worth studying. Our method allows for the automatic discovery of identity categories that are not so frequent as to be obvious from a first inspection of the data, but are in fact sufficiently present to be worth measuring.

As such, we expect the applications of our method to be similar to applications of other unsupervised machine learning methods like topic modeling. Namely, we expect it to enable answers to test theories that are about “identities” without specifying which ones, and where casting a wide net would allow for collection of more data that is relevant to the question. For example, researchers may want to study how identities spread through networks, or whether two people with common identities are more likely to connect. Or they might ask which identities (as described in their profiles) are most likely to be expressed in a person’s social media posts, and which ones are more likely to stay in the background. For example, are someone’s social media posts more likely to give away their politics than their sports

fandom, or vice versa? This type of research could benefit from being able to see patterns across dozens of different kinds of identities, rather than just a few.

Compare this, for instance, to a study by Barberá et al. (2019) investigating whether topics that are brought up by some political actors (e.g. the media) are then picked up by other actors (e.g. Republican audiences). Rather than defining a handful of topics that political tweets might be about, they fit a topic model to extract 100 possible themes. In addition, unsupervised models work well for exploratory analyses. For instance, Yan and Li (2024) ask what topics distinguish censored from uncensored posts on Chinese social media, and start with an unsupervised model to identify 160 topics, among which they then identify the ones that are most and least likely to get a post censored.

2.3 Political identities

Of particular interest in this paper are political self-identifications. Political identities in the US pose something of a puzzle. On the one hand, researchers have argued that partisanship is now a highly salient social identity in the US, which people are keen to express through their stated opinions and behaviors (Bullock et al., 2013; Gift and Gift, 2015; Green et al., 2004; Huddy et al., 2015; Prior et al., 2015; Iyengar and Krupenkin, 2018)—especially online (Mosleh et al., 2021; Osmundsen et al., 2021; Rathje et al., 2021). Indeed, Americans are quite willing to subscribe to a political identity when requested to do so on surveys. 73% call themselves either “conservative” or “liberal” when asked (Gallup, 2022). 52% identify themselves as either Democrats or Republicans; and another 36% are willing to say they lean more towards one party than the other (Gallup, 2023). These so-called “partisan leaners” have been shown to behave similarly to their outright partisan counterparts (Keith et al., 1992; Magleby et al., 2011). Hence, even non-committal expressions of partisan identity are in practice predictive of partisan attitudes and behavior.

On the other hand, studies of Americans’ online behavior suggest that these political identifications are rarely expressed in an active way. A full 70% of social media users say

they never or rarely post or share about social or political issues (Pew Research Center, 2021). Mukerjee et al. (2022) look at follower behavior and show that the average Twitter/X user is far more likely to follow non-political opinion leaders than political ones. As for political self-identification, in a recent analysis by the Pew Research Center (2022) of nearly 600 Twitter/X users with a filled-out bio, only 11% mentioned any kind of political identity. These findings suggest that, at least online, people only rarely consume politics or self-identify as political. By investigating the prevalence of political identities in millions of bios, the substantive findings from our two applications also speak to this debate.

3 Measuring identity-relevant traits

Our proposed method proceeds in four steps, summarized in Figure 1.

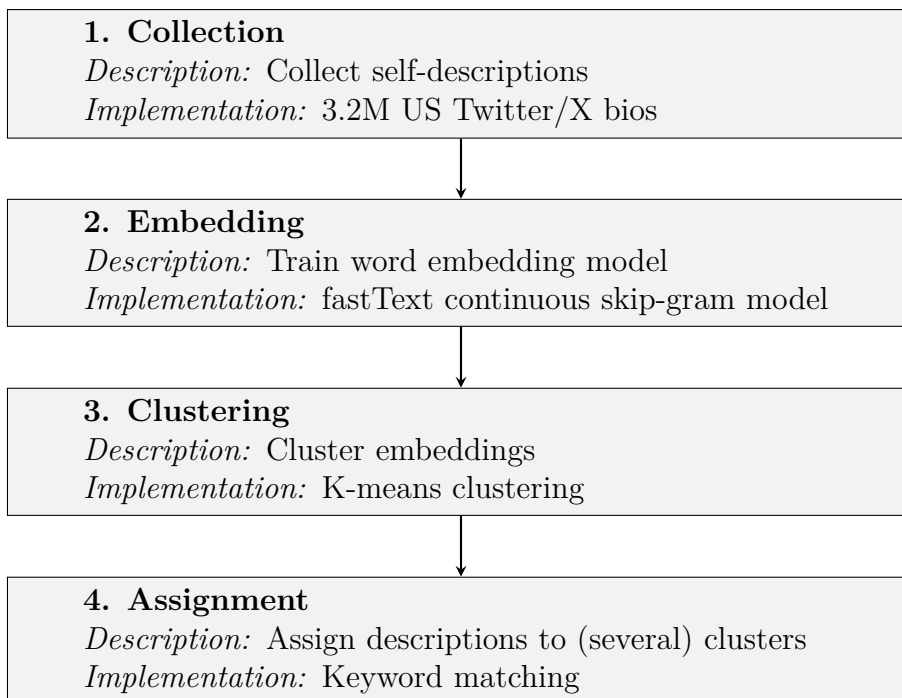


Figure 1: Summary of identity extraction method

In the first step, which we call the *Collection* step, we collect and pre-process self-descriptions (“bios”). In the second step, *Embedding*, we train a word embedding model

on these bios. This means each word that is present in the bios (e.g., “father”) automatically is assigned to a point in a multidimensional space. In this space, words are close to each other if they tend to have similar neighboring words in the bios—in other words, if they tend to be used in the same context. In the third step, *Clustering*, we gather embedded words into clusters—that is, we find groups of words that are close to one another in the embedding space. In the fourth and final step, *Assignment*, we assign each bio to one or more identity-relevant clusters.

Each step in Figure 1 also describes our specific implementation of the method in the running example presented in the following. However, we stress that while the implementation here can serve as a guideline for researchers, features of the research question or data in any given case may dictate another specific implementation. In the remainder of this section, we elaborate on each of the four steps.

An obvious-seeming alternative to our word-embedding-then-clustering pipeline would be to apply a topic model, which attempts to find a number of underlying topics in a text corpus. In our case, many of the topics would hopefully correspond to aspects of identity. It then assigns documents (in our case bios) to one or more topics. The main difference would be that with our method, every meaningful word (or phrase) in the bio is assigned to its own cluster, meaning that a bio can be assigned to as many identities as it has words. With a topic model, an entire bio would be assigned to one or more “topics”, based on all the words it contains. One kind of topic model, a Latent Dirichlet Allocation model, would attempt to guess which topics generated what proportion of each bio (Blei et al., 2003). Other topic models like Top2vec or BERTopic would instead cluster the bios into groups that are close to one another in their general meaning (Angelov, 2020; Grootendorst, 2022), assigning each bio to one group.¹

Compared to our method, topic modeling is a more standard, better-known approach

¹Top2vec and BERTopic were not designed with the goal of outputting more than one topic (identity aspect) per document (bio). Making them do so increases computation time significantly, and we do not know of any attempt to validate the “second”, “third” and so on topic choices by these algorithms.

to the unsupervised grouping of documents by their content. However, in this case, our method has two important advantages over topic modeling. The first is transparency: with our method, we can point to precisely which words in the bio caused it to be assigned to which identity aspects—rather than the whole bio being assigned to one or a few identity aspects for reasons that are not straightforward to explain. This enables analyses like our second application, where we ask whether the presence of political identity words in the bios can be predicted based on only the non-political identity words.

Second, topic models are oriented toward documents that have a consistent theme. They are not geared towards the format of a bio, which is essentially a short list of possibly entirely unrelated phrases. A full-length Twitter/X bio (160 characters) could mention ten or more aspects of a person’s identity, and we would want to capture each one. Methods that treat bios as a whole are not suited to this case. For instance, a user writing “Mom,” followed by a list of ten hobbies, favorite sports teams, and preferred TV shows would almost certainly not be assigned to a Mother topic, because no other words in the bio have any relationship with motherhood.

3.1 Collection

As a first step, our method involves collecting short self-descriptions from a social media platform or another data source where bios are available. In our application, we collect bios from Twitter/X.²

While our running example hence relies on data from Twitter/X, many other online platforms include similar self-description fields. Self-descriptions can come from any online medium where user profiles contain an open text field for self-description. This includes social media platforms such as Mastodon, Threads, Instagram, LinkedIn, Github, Bluesky, and TikTok. Facebook profiles, YouTube channel descriptions, and Pinterest bio sections

²Data were collected via direct calls to the Twitter REST API through implementation of a custom Python function using the `oauth2` package for authentication and `urllib` for HTTP requests. This ensured complete control over request construction, authentication, and rate-limit handling.

all feature self-descriptions. Beyond social media, self-descriptions also feature in online dating and labor markets. Platforms like Tinder, Bumble, and Hinge require users to write self-descriptions, as do professional freelancing sites such as Upwork, Fiverr, and Freelancer. Prior work has incorporated user data from platforms such as GitHub (Terrell et al., 2017), YouTube (Boesinger et al., 2024), and most recently Bluesky (Quelle and Bovet, 2025).

Our data consist of daily records of the Twitter/X bios of 3.2 million American Twitter/X users, from June 2020 to June 2023.³ This is a subset of all users who fit the following criteria (when data collection started): they have some interest in politics or news; are active Twitter/X users; and are located in the US. For the first criterion, a Twitter/X user had to follow at least one of eight major US news outlets: MSNBC, Huffington Post, The New York Times, The Washington Post, CNN, The Wall Street Journal, FOX News, and Breitbart News. For the second criterion, a user had to have tweeted at least once in the past year, made at least 25 tweets since starting their account, and have at least 10 followers. Finally, the user had to be geolocated in the US based on their tweets (using the Carmen tool developed by Dredze et al. 2013). No new users were added to the panel in the course of data collection. Details regarding these data are provided in the online supplement, section A.

We take the following steps to further clean the data. First, we filter out empty bios and bios in another language than English (identified using a pre-trained language identification model from the fastText toolkit, Joulin et al. 2016). The total numbers of empty and non-English bios stay quite stable over time (about 570,000 and 33,000, respectively), leaving us with about 2.6 million English-language bios per day. We keep bios that cannot be assigned to any language with high confidence. They are typically bios that either contain mainly emojis or combine languages—for instance, by quoting a non-English song lyric. We lowercase tokens and remove punctuation, URLs and @ references to other users, but did not apply lemmatization, where morphological variants like “care” and “caring” would be

³Data from June 15th to July 13th 2022 are missing due to issues with a data storage limitation.

reduced to the same lemma (i.e. dictionary entry or vocabulary item). We judged our training data set to be sufficiently large for the embeddings to capture potential nuances in meaning between different forms of the same English lemma.

The next data processing steps—word embedding and clustering—happen at the level of words used in the bios. Before proceeding, however, we first join words that are commonly used together (and less commonly used apart) into short phrases. We use the simple phrase detection algorithm proposed by Mikolov et al. (2013), implemented in the gensim library for python (Rehurek and Sojka, 2011). After this step, common phrases like “black lives matter” or “proud father” are treated as one word. We retain emojis (and phrases using emojis) in the analysis, as they are known to be meaningful parts of users’ self-descriptions (Li et al., 2020).

3.2 Embedding

While the raw content of users’ self-descriptions is a rich source of information on how people identify and present themselves on social media, we need to simplify and classify this content to describe it in the aggregate. The first step in this process is to take the unique words and phrases that were used (at least a handful of times) in the set of bios, and turn them into numerical representations that can be compared and grouped. This is where word embeddings come in.

As noted above, word embedding algorithms are models that translate words into points in a high-dimensional space, so that words with similar meanings are located close to each other in the space. The models start from a large amount of training text (in our case, all collected Twitter/X bios), and use information about which words tend to co-occur. Two words end up with similar representations if they keep the same company—that is, if they tend to have the same neighboring words. Word embeddings have previously been used in political science to code the emotionality of speeches in the US Congress (Gennaro and Ash, 2022) and the UK and Irish parliaments (Osnabrügge et al., 2021); to place legislative

speeches on an ideological scale (Rheault and Cochrane, 2020); and to detect racial or ethnic bias in judicial speech (Rice et al., 2019; Choi et al., 2022), among other applications.

The word embedding algorithm used in our application is a variant of the continuous skip-gram model introduced by Mikolov et al. (2013). This model tries to accurately predict a word’s neighbors, based on the word itself. The weights produced by this model—numbers linking each word to the probabilities of having different neighboring words—are the word embeddings. We use a version of this model included in the fastText toolkit (Bojanowski et al., 2017), where the component letters are also taken into account. In practice, this means the model is much better suited to recognize the similarity between words like “cat” and “cats” or, say, different conjugations of the same verb. This makes the model a good fit for the Twitter/X bio corpus, where users combine words to form hashtags, use alternative spellings of the same word, and so on. Parameter settings (200 dimensions, a context window of six words) were chosen based on tests performed on Congressional records by Rodriguez and Spirling (2022). Our corpus had approximately 70,000 unique lemmas that were used at least 20 times.

A fastText model trained on all of the collected Twitter/X bios from one reference date in the middle of the recorded period (Sept. 6th, 2021) produces satisfying results. In the online supplement, section B we present an overview of the words that were found to be most similar to a number of central identity terms, such as “father” or “Republican”. It shows that the model is especially likely to unearth morphological variants and phrases that include these words, such as “proud father” and “Republican Party”. The model also successfully detects other words that are close in meaning, such as “dad” and “conservative”. In some cases, the model picks up an unexpected degree of nuance: all of the words closest to “Red Sox” refer to other Boston sports teams. Further, results highlight a typical kind of mistake made by this type of model: a word like “ex-Republican” looks similar to “Republican” and is used in similar contexts, and is therefore mistakenly seen as having the same meaning.

Embeddings do not need to be trained on a single date; they can also be dynamically

updated (Rodman, 2020). This way, they will include new words and catch any changes in word usage. We find that 0.4% of the tokens used in bios on the last recorded day are new words (not yet known on the reference date). When we train new embedding models on the first and last recorded bios, the vast majority of (especially the most frequent) words stayed in their current clusters. Only 5% of non-stopword tokens used in the bios (that is, one in every 20 words/phrases in the corpus of bios) end up in a different cluster than they were in on the reference date. In other words, the set of unique lemmas in our data as well as their associations are quite stable over time. Moreover, when we track change in identities over time (in our first application, section 5.1), we prefer to focus on change due to people modifying their bios, as opposed to change due to words changing meaning. As a consequence, while dynamic embeddings will make sense for some applications, we opt for static embeddings in ours.

If this method is applied to a smaller dataset of self-descriptions, then it may be preferable to use pre-trained embeddings rather than training one’s own. This means that the meanings of words are learned from another, larger dataset. The only requirement for this to work well is that words have similar meanings in both data sets. So, for example, if we were working with a dataset of self-descriptions by members of parliament or company employees, then GloVe embeddings trained on Wikipedia (Pennington et al., 2014) will likely be a good fit for the professional language used.

3.3 Clustering

Once all words and phrases in the bios are translated to points in a space, we are able to detect clusters: groups of words that are close to each other, and far from other groups. Before performing clustering, we removed stopwords such as “I” or “am”.⁴ These words are unlikely to contain any identity-relevant information. Further, we recommend weighting

⁴Stopword removal *before* the word embedding step is not necessary (and could in fact harm performance). This is because the embedding algorithm automatically ensures that extremely frequent words do not dominate the training process.

words by their frequency, so that groups of very common words are more likely to be allocated to their own cluster.

A number of different clustering algorithms (HDBSCAN, fuzzy c-means, hierarchical clustering, K-means clustering) could in theory be suitable for the task of clustering bios. Due to the high dimensionality of the word embedding space, and the large number of words (typically tens of thousands) that need clustering, some algorithms are less useful in practice. We found that HDBSCAN and fuzzy c-means were unable to produce meaningful results (allocating all words to the same cluster or no cluster). Hierarchical clustering is computationally expensive in high-dimensional spaces and is also unable to weight words by their frequency. Weighting means that a group of similar words is more likely to get its own cluster if many of the words are common, compared to if most of the words are rare. So, weighting allows the clustering algorithms to avoid creating identity clusters whose combined terms are in fact too rare to make the identity aspect “worth measuring”. On the other hand, it is able to find clusters that consist of just a few different ways to word the same identity, as long as they add up to a reasonably large number of occurrences (e.g. a Pronouns cluster consisting of just a few different variants of “she/her”, “they/them”, etc.).

This leaves K-means clustering as the best option. To measure the distance between words, we used the cosine similarity of their normalized embeddings.⁵ We assessed solutions between 75 and 150 clusters, but chose 100 clusters in our applications. To decide the number of clusters, we opted for a qualitative assessment of the results (later validated against human judgements, see section 4.2). This allowed us to directly validate the output based on the criteria of interest—clusters with identity content meaningful to humans, where key identity concepts (e.g. left-wing and right-wing political affiliation) are neither merged together nor split up into several undesirably small sub-concepts. This was possible due to a manageable number of clusters and the readily interpretable nature of the cluster members (words). Quantitative measures for choosing the ideal number of clusters are also available—

⁵On normalized vectors, ranking vectors by their Euclidean distance (as we do in K-means clustering) is equivalent to ranking them by their cosine distance/similarity.

for example the silhouette score, which assesses whether the resulting clusters are both dense and distinct—but we opted against them here as they are less direct indicators of cluster quality. Clustering results with different numbers of clusters are available on OSF.⁶

As is to be expected, not all clusters of similar words are relevant to describing a user’s identity. In our case, some clusters are artefacts of the platform, such as a Disclaimer cluster capturing variations of statements such as “opinions are my own” or “retweets are not endorsements”. Other clusters capture groups of words that only share a very broad common meaning, or have similar syntactic rather than semantic roles. Examples include the a cluster including variants of the very common word “life” (e.g. “live”, and “life to the fullest”) and a cluster of common filler words (“know”, “think”, “going”). In all, we judged 75 out of the 100 clusters to be identity-relevant. In the online supplement, section C we show the most frequent words for each identity-relevant cluster, as well as a label for the cluster based on those words. We also show the top words for each cluster deemed to be identity-irrelevant.

3.4 Assignment

The fourth and final step in the method involves associating each bio with the set of all identity-relevant clusters that were mentioned in it. A cluster is added to the set of identity-relevant aspects present in a bio if any word from that cluster was present in the bio. For instance, a mention of the word “mom” would be sufficient to assign a bio to the Mother cluster. If the same bio mentioned “Christian”, it would also be assigned to the Christian cluster, and so on.

As for our specific focus, political identities, the identity clustering method results in two main clusters with partisan content: a Conservative/Republican cluster, and a Liberal/Democrat one. Both clusters contain a mix of partisan (“Trump”) and ideological (“patriot”) identifiers. This is perhaps unsurprising given how closely related partisanship

⁶https://osf.io/nxfk3/overview?view_only=3c4d1b412a7d46c3a893218724231788

and ideology are in the American context: 74% of US Republicans identify as conservative and 50% of Democrats as liberal (Gallup, 2022).

Given our specific interest in these political identity clusters, we took a few additional steps to clean their content. We filtered out words without sufficiently clear partisan content (e.g. “vote”). We also flipped the cluster labels of bios that seemed to be using political words as part of a negative phrase. the online supplement, section D describes the process in more detail. Finally, for the application where we predict users’ political identities based their non-political ones (see section 5.3), we also strip the political identity clusters of words referring to issues rather than ideology (covering about 10% of the political cluster tokens in the bios).

4 Validation

4.1 Construct validity

To validate whether the clusters resulting from our procedure represent what we intend to measure—namely identity signals—we can take two steps. First, we can discuss whether the aspects of self-presentation language captured by the bios are indeed “identities”. We judged 75 clusters to be relevant to identity (see Supplemental Materials section C). The identity-relevant clusters describe occupations (Artist, Business Leadership, Higher Education), personality traits and values (Positivity, Personal Growth), interests (Hobby Emoticons, Sports Disciplines), relationships (Family Roles, Mother) and social groups (Sports Teams, Veteran, Conservative/Republican, Liberal/Democrat). As such, they represent all three dimensions of the self: individual, relational, and collective (Sedikides et al., 2011).

We take an expansive approach to identity. For instance, we deemed a cluster of positive emojis (e.g. hearts, fire, positive smileys) as relevant to identity, as a person using those icons may be communicating a distinct personality trait or mindset; others might see it as an unimportant feature of one’s communication style. Our decision to leave in clusters that

others might not recognize as capturing “identity” is informed by the exploratory nature of our applications, where we (1) track the content of self-descriptions on Twitter/X, and situate politics in it; (2) find the most distinctive identity for each US state; and (3) investigate many possible traits as predictors of political identity, showing that some unexpected categories (e.g. Nature Emojis) are surprisingly related to politics. Researchers who want to use our method to test theory could limit the scope and exclude more clusters—for example, limiting their selection to only social identities. Compare this e.g. to Barberá et al. (2019), who set out to study how political discussion topics spread. They use a topic model to find 100 potential “topics”, but ultimately judge only 53 of them to be actual political issues relevant to their research question.

Second, we can ask whether the *words* that are grouped together into identity clusters, at face value, seem to represent coherent concepts—in particular, whether they seem to represent the concepts that we hand-picked as names for the clusters. After all, groups of words that are used similarly in sentences do not necessarily represent meaningful aspects of identity. In Supplemental Materials section C, we provide lists of words belonging to all the clusters, so that the reader may form their own judgment about their construct validity. We deemed most clusters to be quite coherent, but there are exceptions. For example, the Volunteer cluster includes a number of high-frequency words that instead point to a person shifting roles, e.g. “currently”, “past” and “previously”. This problem could partly be tackled by several further rounds of stopword culling.

4.2 Comparison to human raters

To validate our assignment of bios to identity-relevant clusters, we had three coders hand-code 3,000 combinations of a bio and an identity-relevant cluster that may or may not occur in that bio. For each combination, the coders rated whether the bio should belong to the identity-relevant cluster or not. For each cluster, the hand-coding set included 20 positive examples (bios that our algorithm had indeed assigned to the cluster) and 20 negative

examples (bios that our algorithm had not assigned to the cluster). For example, for 40 of the bios in the hand-coding set, the coders were asked to decide whether or not these bios include the identity of Entrepreneur. Of these 40 bios, 20 had been randomly chosen among bios that include a word from the Entrepreneur cluster, according to our method. The other 20 had been randomly chosen among bios that do not include any such words. Our method would have perfect performance if the coders (who could not see the method’s classifications) answered “yes” to all bios from the first set, and “no” to all bios from the second.

Deciding which identity-relevant aspects are present in a bio is a difficult task even for humans; Krippendorff’s α between coders was 0.71. Krippendorff’s α between the algorithm’s ratings and the coders’ mode (majority vote) was .59. Performance varied significantly between clusters; Figure 2 displays the results by cluster. In the online supplement, section E we present the full set of results. Notably, the main type of error made by the model was to assign identities that the coders did not assign (rather than missing ones that coders did assign); 85% of errors were of this type. In other words, the clusters’ keywords cast too broad of a net, catching some false positives. In the rarer cases where our algorithm failed to detect a cluster that the coders did assign, it was often due to the fact that another cluster was an even better fit for a given keyword (e.g. bios containing “mom” were hand-coded as having Family Roles, while the algorithm categorized this word as Mother instead).

The clusters that were most difficult for the algorithm to allocate correctly are, by and large, also the clusters that humans struggled with (those with lowest intercoder reliability). Most of these difficult clusters were also ones that had lower coherence at face value (e.g. the “Volunteer” cluster mentioned above); in other words, potentially problematic clusters appear to be easy to identify. Combined with the fact that most of the algorithm’s mistakes were false positives, this implies that performance could be improved by researchers manually filtering out words that are ambiguous or seem to lie at the edge of a cluster’s meaning—i.e., ones are likely to render false positives. We did not engage in that here, in order to show our results in their raw form, and to make clear that a degree of researcher intervention might

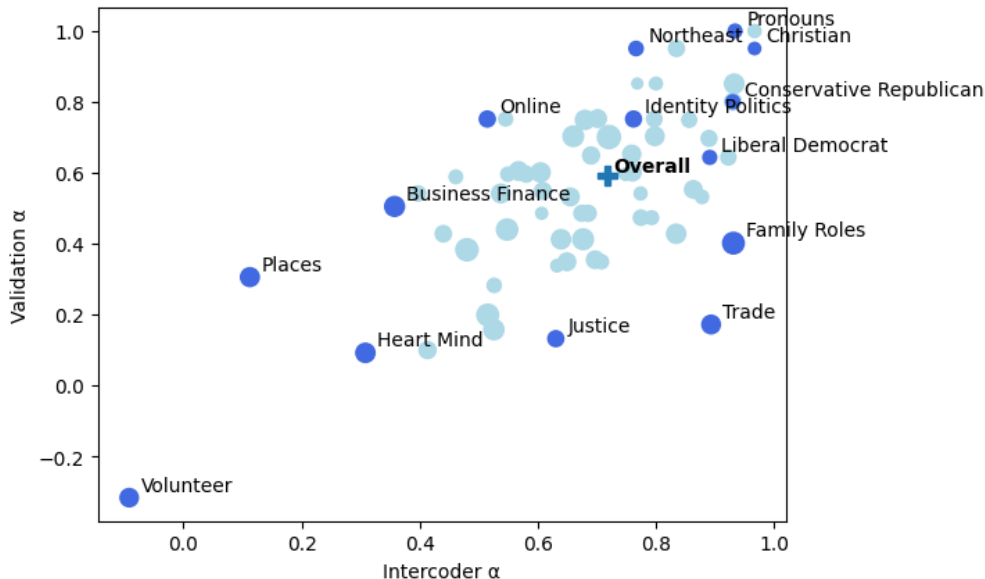


Figure 2: Scatterplot of identity-relevant clusters showing Krippendorff’s alpha between our algorithm and coders’ (modal) identity assignment, plotted against intercoder alpha. Dot sizes reflect popularity of the clusters (no. of bios using them). Cross shows the overall algorithm–coder and intercoder alpha across the whole hand-coding dataset (n=3000).

be needed in order to make sure that clusters match theoretical concepts of interest.⁷

5 Applications

In this section, we present three applications illustrating the utility of our method. In the first application, we track the evolution of political and non-political identities (and identity-relevant traits) for American Twitter/X users. In the second, we map the most distinctive identities for each US state. Lastly, we evaluate how predictable political identities are based on other identity information. Note that throughout these applications, we do not report uncertainty estimates, as our results are affected by neither sampling nor randomization variance.

⁷The exception are the political clusters, which were of particular interest in our applications—as mentioned above, here we did manually screen out words that were likely to lead to false positives.

5.1 Tracking Twitter/X identities

Figure 3 describes the over-time evolution of popular identity-relevant traits on US Twitter/X. It shows the percentage of bios in our data set that mention each of the 15 most frequent identity-relevant clusters throughout the observed period (we limit this figure to the most popular clusters for legibility, but present a version with all clusters in the on-line supplement, section F). A few observations stand out. First, most clusters’ popularity hardly changes over time. This makes sense: self-identifications should be stable within a given context. Moreover, even when identities change, people may or may not make the effort to change their social media profiles accordingly, meaning that our results will be biased towards over-time stability.

Among the most popular identity-relevant clusters, only Hobby Emojis and Higher Education show a steady increase and decline, respectively. The first is part of a general tendency towards growing use of emojis in self-descriptions. The second appears to be partly caused by students in our panel graduating and replacing their school with a job description.

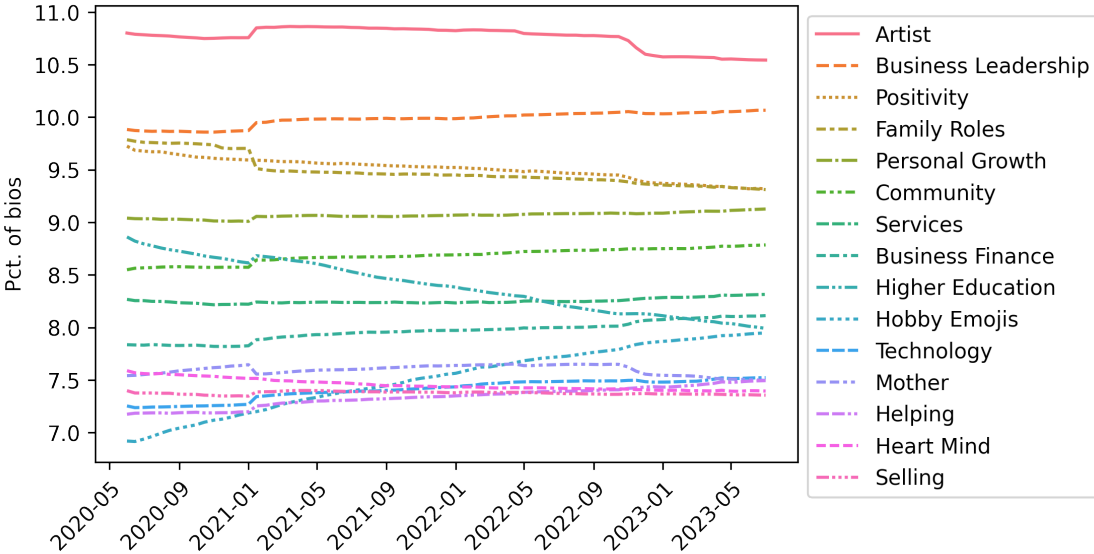


Figure 3: Percentage of bios containing each of the 15 most frequent identity-relevant clusters in our sample of US Twitter/X bios, at biweekly intervals from June 2020 to June 2023.

The small up- and downticks in January 2021 are due to almost 63,000 users dropping

out of our data set after the attack on the US Capitol by supporters of Donald Trump (see Eady et al. 2022). These users either deleted their bio, or left Twitter/X altogether. A disproportionate number (46%) of these dropped-out users had Conservative/Republican identities in their bios (see below). As a result, clusters that tend to co-occur with the Conservative/Republican cluster (such as Family Roles) saw small declines, while the opposite is true for clusters that tend not to co-occur with conservatism (such as Artist or Business Leadership).

Second, the results confirm that even on a news-oriented platform like Twitter/X, and even among politically interested users, politics does not even come close to being a central part of most people’s self-identifications. None of the 15 most frequent clusters have political content. This implies that even though 57% of Twitter/X users identify as partisans when asked in a survey (Pew Research Center, 2019), users are still far more likely to use their bio to describe their hobbies, personalities and professions. In the words of Dahl (1961), “politics is a sideshow in the great circus of life”. This is particularly remarkable given the composition of our user panel, where users had to follow at least one major news outlet to be included. Among users with little or no interest in current affairs, political self-identifications are likely to be even rarer.

The left panel of Figure 4 shows the prevalence and evolution of three clusters that do have clear political content: the Conservative/Republican cluster, the Liberal/Democrat cluster, and the Justice cluster (with words such as peace, justice, freedom, and government). It also shows the proportion of bios that have at least one of the three political clusters represented. In the beginning of the observed period, Conservative/Republican is the largest political cluster, perhaps because some typical left-wing values are instead picked up by the Justice cluster. However, it experiences a sharp drop-off in January 2021 due to the Capitol insurrection, and lands at nearly the same level as the Liberal/Democrat cluster; around 4%.

Throughout the period, only 11%–13% of bios contain any mention of a political identity.

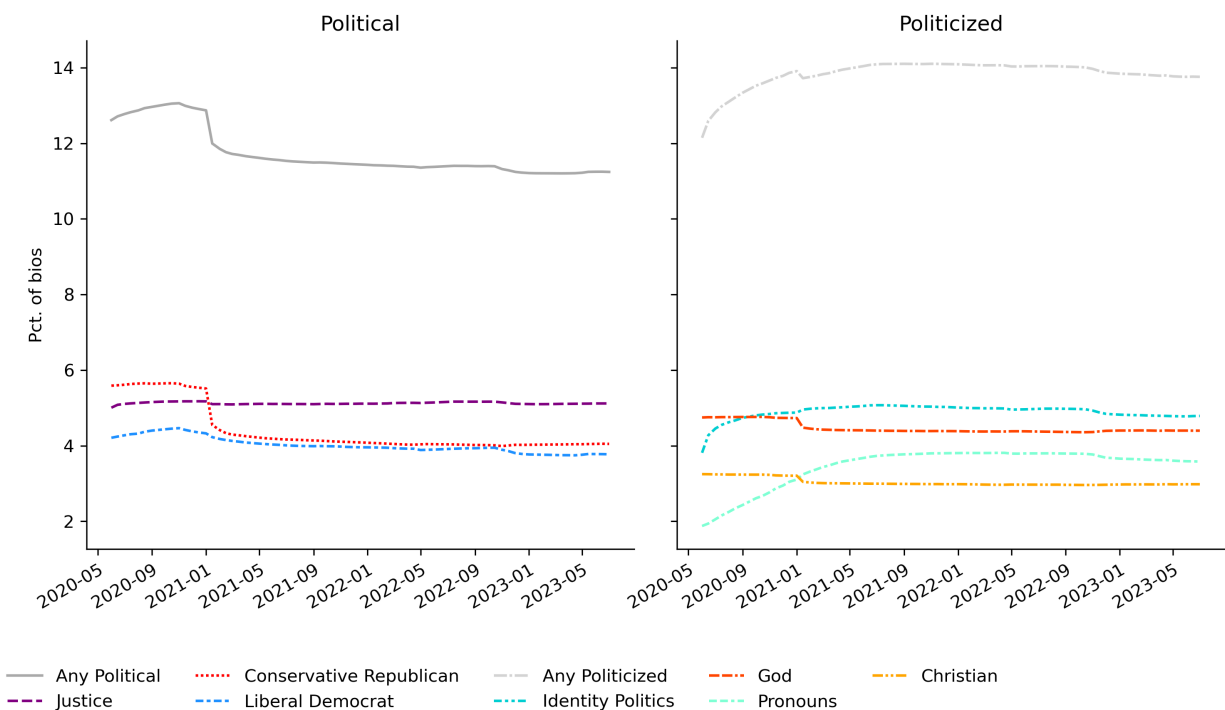


Figure 4: Prevalence of clusters with explicitly political content (left) and with likely politicized content (right).

This matches the findings by Mukerjee et al. (2022), who show that people rarely engage with political content on Twitter/X. It also almost exactly replicates the finding by the Pew Research Center (2022) that 11% of bios mention politics—based on a much smaller, hand-coded sample. These relatively low numbers are somewhat surprising, given that 41% of Twitter/X users (in a multi-country survey sample) say they are “very” or “extremely” interested in politics (Reuters Institute, 2023). This highlights the discrepancy between Twitter/X users’ self-reported interest in politics on the one hand and their willingness to foreground politics in their self-presentations on the other.

At the same time, there are several identity-relevant traits in our set of clusters that are likely to be *politicized* if not explicitly political. The most prominent examples of these are Pronouns (e.g. “she/her”), Identity Politics (a cluster with words such as BLM, Black, feminist, vegan and LGBTQ), God, and Christian. Most readers coming across these in a bio would give these identities some political meaning. In fact, in our second application in

section 5.3, we will see that the presence of these identities is predictive of a user’s politics. The right panel of Figure 4 shows that at the end of the observed period, these identities are more common than the explicitly political ones. In particular, identities associated with Identity Politics and Pronouns saw an increase at the beginning of the period (reflecting findings on pronouns by Tucker and Jones 2023), perhaps indicating an increase in salience of “identity politics” as often argued in the public debate (Fukuyama, 2018; Mounk, 2023).⁸ Combined, these four politicized identities cover 12-14% of all bios. In sum, users are somewhat more likely to communicate a political affiliation through identities that are implicitly rather than explicitly political.

5.2 Mapping identities in the US states

As a second application of our approach, we map expressions of identities across US states in order to understand their local variation. We think of this as a novel approach to characterizing US states in their culture and other relevant traits, following earlier work by (see, e.g., Elazar (1984) and Cohen (1996)). It highlights the strength of being able to process large amount of data sets to provide credible estimates of local identity prevalence, something that would typically be challenging or expensive using traditional survey data.

To conduct this analysis, we first identified a subset of 2.1 million users whose geographical location (via their profile “location” field) could be pinpointed to a particular US state. We then used chi-squared tests to calculate the interdependence between each state and cluster—that is, the degree to which a cluster is uniquely prevalent in a state (as opposed to all other states).⁹ Figure 5 shows the identity cluster with the highest χ^2 statistic for each state. the online supplement, section G reports on the top three clusters per state, their χ^2 statistics, and state sample sizes.

⁸The increase in the Identity Politics cluster appears to be driven primarily by an increase in the use of Black Lives Matter related terms, and to a lesser extent by LGBTQ-related terms.

⁹In rare cases, an identity cluster will receive a high χ^2 because it is *negatively* associated with a state, its absence being predictive of residing in the state. We only report on positive associations. We also leave out the “Northeast” cluster, which is trivially associated with Northeastern states.

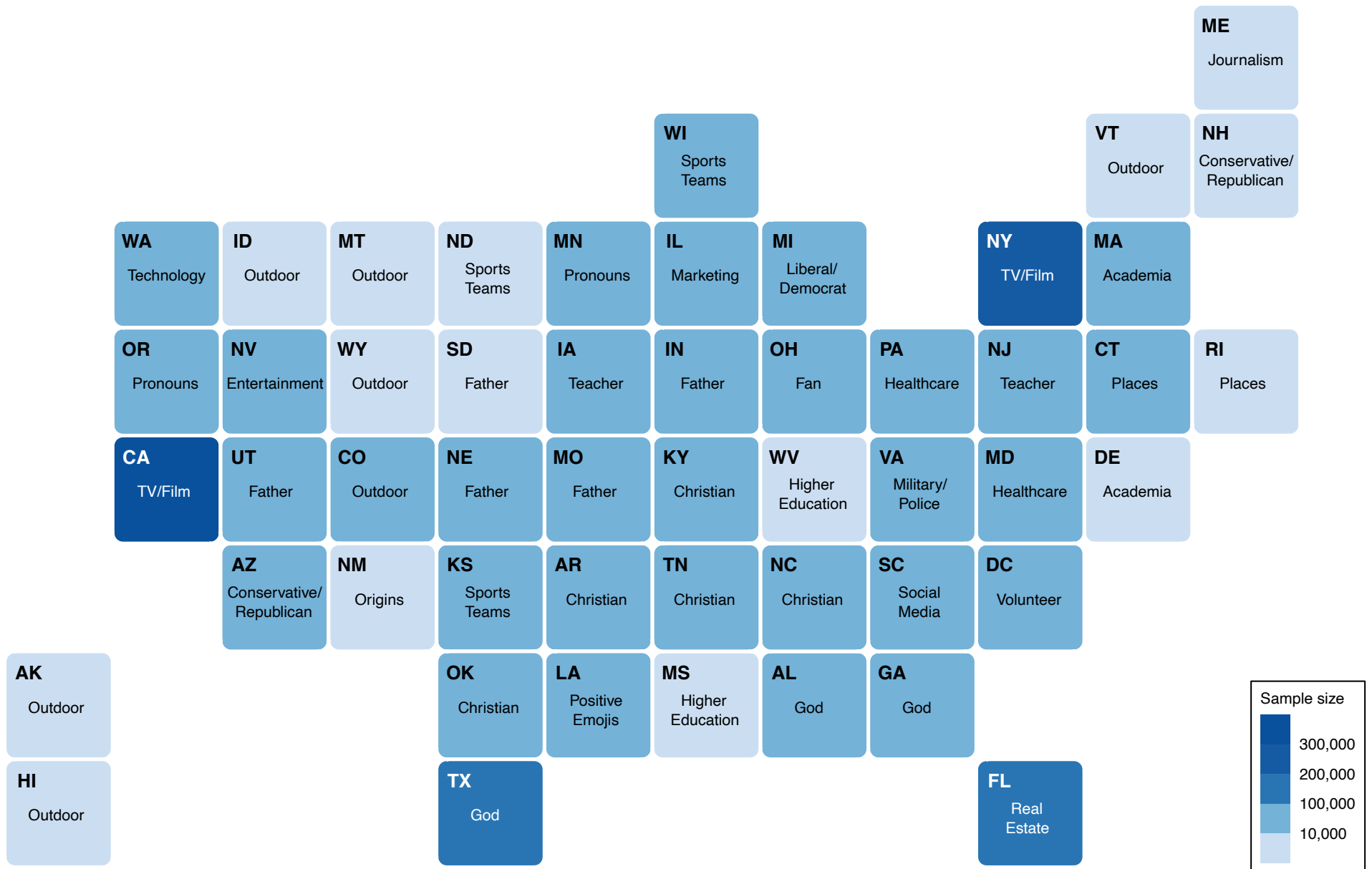


Figure 5: The identity cluster whose presence is most associated with each US state (highest χ^2).

To a large degree, the results are intuitive, confirming the validity of our way of extracting and cutting identities. Several states are associated with industries they are known for: Washington with Technology, California and New York with TV/Film, Massachusetts with Academia. Other states are connected to cultural practices: the God and Christian clusters are prevalent in the South. Father is prevalent in non-South conservative states; we will come back to its association with conservatism below. Outdoor identities are predictive of states that overlap with the Rocky Mountains, as well as Vermont, Hawaii and Alaska—all states known for their natural features. Pronouns are associated with Oregon, where the city of Portland is one of the most left-leaning cities in the US (the state’s next-most associated identities are Identity Politics and Liberal/Democrat). They are also linked with Minnesota, which has two cities in the top-13 of left-leaning cities (Tausanovitch and Warshaw, 2014).

However, a number of surprising patterns arise as well. For example, both New Hampshire and Arizona have Conservative/Republican as their most uniquely prevalent cluster, but both states also feature Liberal/Democrat as the second-highest (see Supplementary Table G.1). This suggests that in some states, people are simply more willing to declare their politics than in others. One explanation may be that both states are split fairly evenly in their politics, mobilizing citizens to express their affiliation. However, other politically “competitive” states (Gallup, Inc., 2017) feature no political clusters. Another curious pattern is the prevalence of Higher Education (typically people naming their alma mater) in West Virginia and Mississippi, the two states with the *lowest* proportions of people with a Bachelor’s degree or higher (U.S. Census Bureau, 2022). Perhaps it is in these states that higher education status is most valuable as a signal.

5.3 Predicting politics from non-political identities

Several studies have indicated that Democrats and Republicans are different from each other with respect to characteristics that are, at first glance, unrelated to politics: they shop at different stores (Wilson et al., 2014), drive different cars (Gebru et al., 2017), watch different

TV shows (Carter, 2012) and use written language differently (Sylwester and Purver, 2015). Liberals and conservatives have also been found to have different personalities and non-verbal communication styles, and put different objects in their homes and offices (Carney et al., 2008). If people’s hobbies, personalities and habits are indeed correlated with their politics, and if they tend to seek out like-minded others on all of those aspects, then they might find themselves surrounded with social contacts of the same political persuasion (DellaPosta et al., 2015). This matters because interacting with others who hold dissenting political views is associated with greater tolerance, more complex opinions, and better understanding of others’ political points of view (Mutz, 2002; Green et al., 2000; Price et al., 2002).

In our third application, then, we ask whether users’ partisan/ideological identities are predictable from their other, non-political identity-relevant traits. Related, Essig and DellaPosta (2024) showed that a model trained on Twitter/X self-descriptions to detect users’ political identities, based only on the *non-political* words in the self-description, was able to correctly guess a user’s political leaning 74% of the time. We perform a similar analysis, but rather than focusing on the performance of the model, our aim is mainly to draw substantive conclusions about what non-political identity categories are most predictive of politics.

We classify users as Liberal/Democrat if they have more Liberal/Democrat words in their bio than Conservative/Republican ones. This step helps us avoid misclassifying users with undetected negative uses of the partisan cluster terms (e.g. “can’t stand Hilary”), as those uses are typically accompanied by one or more positive uses of the other partisan category. This leaves us with about 89,000 users who appear to identify as Republicans/conservatives and 86,000 Democrats/liberals. The fact that there are slightly fewer explicitly self-identified Democrats than Republicans in our data set is surprising, as survey research suggests that 36% of Twitter/X users identify as Democrats, and only 21% as Republicans (Pew Research Center, 2019). Perhaps Conservative/Republican users are more likely to signal their politics more explicitly, while Democrats/liberals either express theirs indirectly (e.g. by specifying pronouns as shown earlier) or not at all.

We fit a regularized logistic regression model¹⁰ to predict bios’ partisan/ideological label, based on all the other clusters present in the bio. There are 73 explanatory variables in this model, one for each identity-relevant cluster other than the partisan clusters. Their value for each bio is the number of unique words in the bio that belong to that cluster. We delete users that have only one identity-relevant cluster (or none) present in their bios besides the partisan ones.

The model has an out-of-sample accuracy of 69%. In other words, we can correctly guess 69% of Twitter/X users’ explicit partisan identities using their other, non-political identity-relevant clusters. In other words, for people who have political self-identifications on Twitter/X, we can predict their affiliation fairly well from less than 160 characters’ worth of other identity-relevant information, suggesting that partisan sorting on social and lifestyle dimensions is significant. The predictive performance of our model is slightly lower than that of Essig and DellaPosta (2024), because we bundle non-political identity words into clusters first. If we add the average word embedding of all non-political-cluster words in the bio, out-of-sample accuracy increases to 76%. A BERT model (Devlin et al., 2018) trained on all the words in the bio except for the political cluster words has an accuracy of 86%. However, these models do not result in easily interpretable conclusions about which identities are most predictive of users’ politics, whereas our clustering method offers precisely that.

Figure 6 shows the 15 clusters that are mentioned proportionally the most by Liberal/Democrat and Conservative/Republican users respectively. the online supplement, section H shows these findings numerically. The results are largely intuitive—for example, Pronouns and Identity Politics clusters are used far more by Liberals/Democrats, while God, Family, and Military/Police are used much more by Conservatives/Republicans. Others would have been less obvious a priori, and point to connections that would have been difficult to uncover in a population sample of perhaps 1,000 or 2,000 adults. For example, there are fairly strong connections between Liberal/Democrat identities and professions like

¹⁰with an L2 penalty whose C value was chosen via cross-validation.

Teacher or Healthcare—the latter perhaps due to the fact that during our research period, Republican rhetoric de-emphasized the risks of the COVID-19 pandemics, and Republican state policies were far less stringent, leading to excess deaths in red states (Halpern, 2020; Güss et al., 2023). On the other hand, mentioning clusters related to private sector employment and business ownership is correlated with being Conservative/Republican.

Another pattern of interest is that Father is associated with being Conservative/Republican, whereas Mother is not (in fact, it is almost 60% more likely to occur in Liberal/Democrat bios than in Conservative/Republican ones). This can be partly explained by the modest gender gap in US political affiliations (Pew Research Center, 2024), but not entirely.¹¹ Perhaps mentioning fatherhood is a way to signal affiliation with the family values underlying conservatism—though it is unclear why motherhood does not take the same function. Qualitative research about conservatism and fatherhood could help shed light on this. Finally, there is a surprising correlation between being Liberal/Democrat and a cluster named Nature Emojis (consisting largely of plants, colored hearts, stars, and zodiac signs). Several of these seem to be connected to liberal identities, by signifying environmentalism, acceptance of diversity, or spiritualism. An interesting extension of this analysis would be to see whether the predictors of politics change over time—e.g. whether Pronouns become more associated with Liberal/Democrat politics.

Some patterns are present in the bios that are misclassified by the model. A few users truly deviate from partisan sorting, mentioning identities that are typically associated with the opposite party or ideology (e.g. military and Democrat). Others mention very few clusters, or clusters that are not very predictive of politics (e.g. hobbies and sports teams). For a handful of users, it is their partisan/ideological identity that is mischaracterized, typically because they are using a political term either in a negative way or with an alternative, non-political meaning. We estimate that 6.5% of bios would have received an incorrect political label (see the online supplement, section D), suggesting that model performance would be

¹¹In fact, married men and women both tilt Republican (59 and 50%), whereas men and women with live-in partners both tilt Democrat (60 and 64%, Pew Research Center 2024).

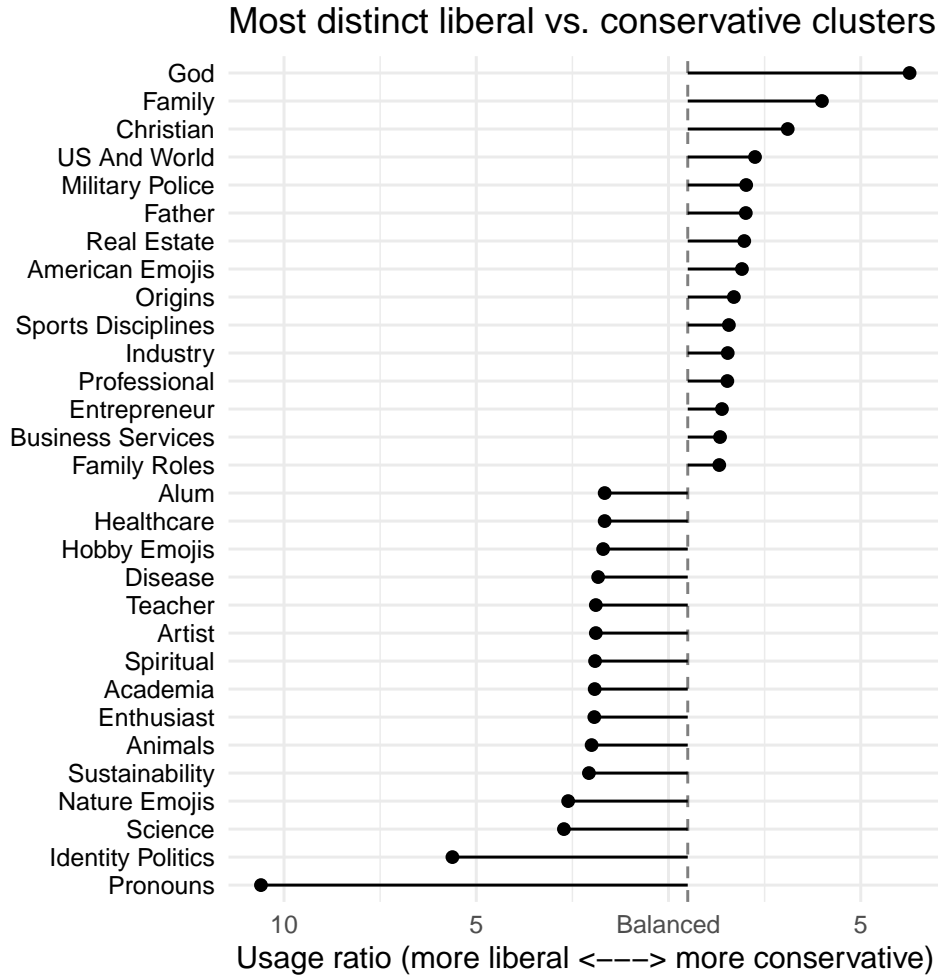


Figure 6: Identity-relevant clusters that are mentioned proportionally the most by Conservative/Republican users and by Liberal/Democrat users. X-axis shows the ratio of liberal to conservative users (left side) and of conservative to liberal users (right side), with respect to the rates at which these groups mention the clusters. E.g., liberal users are eleven times more likely to mention Pronouns; conservative users are six times more likely to mention God.

stronger if the political labels contained no errors.

Since a user had to mention a partisan or ideological term in their bio to qualify, these results speak to the connection between partisan and non-partisan identities for those users who were willing to mention both. It is possible that users whose political identity is atypical given their other identities (e.g. evangelical Democrats) are more likely to stay silent about either their politics or their other affiliations. In other words, expressed political identities

may be more predictable than private political identities.

6 Discussion

In this paper, we present a bottom-up method for measuring mass self-identification—namely, a combination of unsupervised machine learning steps applied to social media self-descriptions. By discovering people’s identity-relevant traits from a naturally occurring behavior, we avoid measuring “doorstep identities”, where survey respondents must come up with answers to self-identification questions that they rarely think about in the course of their daily lives. By building identity categories from the ground up, we also avoid mismatched or outdated coding schemes. By using dynamic embeddings, our method can easily be extended to incorporate new identity terms that appear over time. Because the large data sets generated by social media allow us to apply natural language processing methods such as word embeddings, less common identity terms can be picked up and incorporated. Similarly, because the approach allows us to automatically process a number of users that is several orders of magnitude larger than the average survey, we can find and analyze people with rare (combinations of) identities.

Our approach illustrates how large-n analyzes of social media profiles can go beyond tracking keywords. Compare, for instance, the approximately 700 words and phrases that we identified as having partisan or ideological meaning to the lists of 13 explicitly and 13 implicitly political keywords identified by Rogers and Jones (2021). Manually constructed keyword dictionaries are useful due to their precision: they are fairly unlikely to catch false positives, texts that do not actually fall into the desired categories. At the same time, they are likely to miss many instances of the concept to be measured (e.g., users showing their political affiliation by including a blue wave emoji). Using word embeddings, a manually selected set of keywords can be expanded by supplementing each keyword with additional words that are close to it in the embedding space. This approach can unearth less common

identity terms. Even more importantly, our method allows for the bottom-up discovery of identity categories that the researchers had not anticipated.

While many of the steps in our method are automated, it is clear that the process also leaves significant leeway for researcher decisions. This is both a strength and a potential weakness. For instance, the current clustering algorithm (K-means with 100 clusters) was selected for its interpretable result—a criterion that is difficult to quantify except by applying a costly validation step to several possible parameter settings of the algorithm. Naming the clusters is also a matter of discretion. Moreover, there is no clear boundary between identity-relevant and irrelevant clusters. In sum, while the method is largely automated, it is still subjective.

Finally, as noted in the Introduction, the content of social media bios can be seen as self-identifications that meet two criteria: importance to the user, and (perceived) acceptability to the audience. The second criterion in particular means that our findings will be specific to whichever social norms govern identity expression on a given platform. In the end, though, self-identification is by definition situational: how we characterize ourselves will always depend on the context. And it always requires “at least some degree of explicit discursive articulation” (Brubaker and Cooper, 2000). Whether we express our identity in the context of a survey or a social media website may influence what we say, but there is no true “neutral ground” for identity expression.

Statements and Declarations

Declaration of conflicting interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical considerations

Ethics approval for this study was not required, as it involved only the processing of publicly available information. Data processing for this article was governed by GDPR, and personal data was processed on the grounds of legitimate interest.

Funding statement

This research was funded by the Independent Research Fund Denmark. Grant number: 9038-00123B. Grant project title: "E Pluribus Plures: Is identity politics a new fault line in contemporary societies?"

Preregistration statement

This study was not pre-registered.

Data Availability statement

Documentation and data for replicating results from the paper are available at the OSF repository Measuring Social and Political Identities In Social Media Self-Descriptions. The posted replication code and data reproduce the aggregate-level results in the article. As a consequence of the EU-wide General Data Protection Regulation (GDPR), we are unable to share individual-level data in the replication materials as they constitute personal data which enable identification of individuals. Thus, although we provide the code necessary to implement our method on another data set, as well as scripts to replicate the figures in the article, we cannot make public the data needed to replicate the individual-level results. The individual-level data can potentially be accessed by other researchers for replication if (1) permission is obtained from the Danish Data Protection Agency and (2) a data transfer agreement is signed with the University of Copenhagen. Please contact the corresponding author with data sharing requests.

References

- A. Agadjanian. How many Americans change their racial identification over time? *Socius*, 8, 2022.
- American Press Institute. How people use Twitter in general, 2015. URL <https://www.americanpressinstitute.org/publications/reports/survey-research/how-people-use-twitter-in-general/>. Accessed May 2022.
- D. Angelov. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.
- R. Ashmore, K. Deaux, and T. McLaughlin-Volpe. An organizing framework for collective identity: Articulation and significance of multidimensionality. *Psychological Bulletin*, 130: 80–114, 2004.
- C. Baden, C. Pipal, M. Schoonvelde, and M. A. G. van der Velden. Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication methods and measures*, 16(1):1–18, 2022.
- P. Barberá, A. Casas, J. Nagler, P. J. Egan, R. Bonneau, J. T. Jost, and J. A. Tucker. Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, 113(4):883–901, 2019.
- D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- L. Boesinger, M. Horta Ribeiro, V. Veselovsky, and R. West. Tube2vec: Social and semantic embeddings of YouTube channels. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):2084–2090, 2024. doi: 10.1609/icwsm.v18i1.31450.
- P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.
- G. K. Brown and A. Langer. Conceptualizing and measuring ethnicity. *Oxford Development Studies*, 38(4):411–436, 2010.
- R. Brubaker and F. Cooper. Beyond “identity”. *Theory and society*, 29(1):1–47, 2000.
- J. G. Bullock, A. S. Gerber, S. J. Hill, and G. A. Huber. Partisan bias in factual beliefs about politics. Technical report, National Bureau of Economic Research, 2013.
- J. Cappos, S. T. Peddinti, and K. W. Ross. User anonymity on Twitter, 2017. URL <https://www.infoq.com/articles/user-anonymity-twitter>. Accessed July 2023.
- D. R. Carney, J. T. Jost, S. D. Gosling, and J. Potter. The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. *Political psychology*, 29(6):807–840, 2008.

- B. Carter. Republicans like golf, Democrats prefer cartoons, TV research suggests. *Media Decoder Blog. New York Times*, 2012. URL <https://archive.nytimes.com/mediadecoder.blogs.nytimes.com/2012/10/11/republicans-like-golf-democrats-prefer-cartoons-tv-research-suggests/>.
- D. D. Choi, J. A. Harries, and F. Shen-Bayh. Ethnic bias in judicial decision making: Evidence from criminal appeals in kenya. *American Political Science Review*, pages 1–14, 2022.
- D. Cohen. Law, social policy, and violence: The impact of regional cultures. *Journal of personality and Social Psychology*, 70(5):961, 1996.
- P. J. Conover. The influence of group identifications on political perception and evaluation. *The Journal of Politics*, 46(3):760–785, 1984.
- P. E. Converse. The nature of belief systems in mass publics. In D. E. Apter, editor, *Ideology and Discontent*. Free Press, New York, 1964.
- F. Corso, F. Pierri, and G. D. F. Morales. What we can learn from tiktok through its research api, 2024.
- R. A. Dahl. *Who governs?: Democracy and power in an American city*. Yale University Press, 1961.
- D. DellaPosta, Y. Shi, and M. Macy. Why do liberals drink lattes? *American Journal of Sociology*, 120(5):1473–1511, 2015.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- E. P. Diamond. The influence of identity salience on framing effectiveness: An experiment. *Political Psychology*, 41(6):1133–1150, 2020.
- M. Dredze, M. J. Paul, S. Bergsma, and H. Tran. Carmen: A Twitter geolocation system with applications to public health. In *Workshops at the twenty-seventh AAAI conference on artificial intelligence*, 2013.
- G. Eady, F. Hjorth, and P. T. Dinesen. Do violent protests affect expressions of party identity? Evidence from the capitol insurrection. *Working paper*, 2022.
- P. J. Egan. Identity as dependent variable: How Americans shift their identities to align with their politics. *American Journal of Political Science*, 64(3):699–716, 2020.
- D. J. Elazar. *American federalism: A view from the states (3rd edition)*. Harper & Row, 1984.
- L. Elder and S. Greene. *The politics of parenthood: Causes and consequences of the politicization and polarization of the American family*. SUNY Press, 2012.

- L. Essig and D. DellaPosta. Partisan styles of self-presentation in us twitter bios. *Scientific reports*, 14(1):1077, 2024.
- K. A. Ethier and K. Deaux. Negotiating social identity when contexts change: Maintaining identification and responding to threat. *Journal of personality and social psychology*, 67(2):243, 1994.
- D. Freelon. Computational research in the post-api age. *Political Communication*, 35(4):665–668, 2018.
- F. Fukuyama. *Identity: Contemporary identity politics and the struggle for recognition*. Profile books, 2018.
- Gallup. U.S. political ideology steady; conservatives, moderates tie, 2022. URL <https://news.gallup.com/poll/388988/political-ideology-steady-conservatives-moderates-tie.aspx>. Accessed June 2022.
- Gallup. Party affiliation, 2023. URL <https://news.gallup.com/poll/15370/party-affiliation.aspx>. Accessed July 2023.
- Gallup, Inc. 2017 U.S. party affiliation by state. Gallup News, 2017. URL <https://news.gallup.com/poll/226643/2017-party-affiliation-state.aspx>.
- T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113, 2017.
- G. Gennaro and E. Ash. Emotion and reason in political language. *The Economic Journal*, 132(643):1037–1059, 2022.
- K. Gift and T. Gift. Does politics influence hiring? Evidence from a randomized experiment. *Political Behavior*, 37:653–675, 2015.
- D. P. Green, B. Palmquist, and E. Schickler. *Partisan hearts and minds: Political parties and the social identities of voters*. Yale University Press, 2004.
- M. C. Green, P. S. Visser, and P. E. Tetlock. Coping with accountability cross-pressures: Low-effort evasive tactics and high-effort quests for complex compromises. *Personality and Social Psychology Bulletin*, 26(11):1380–1391, 2000.
- M. Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.
- C. D. Güss, L. Boyd, K. Perniciaro, D. C. Free, J. Free, and M. T. Tuason. The politics of covid-19: Differences between us red and blue states in covid-19 regulations and deaths. *Health Policy OPEN*, 5:100107, 2023.
- L. W. Halpern. The politicization of covid-19. *AJN The American Journal of Nursing*, 120(11):19–20, 2020.

- M. Hare and J. Jones. Slava Ukraini: Exploring identity activism in support of Ukraine via the Ukraine flag emoji on Twitter. *Journal of Quantitative Description: Digital Media*, 3, 2023.
- D. J. Hopkins, Y. Lelkes, and S. Wolken. The rise of and demand for identity-oriented media coverage. *American Journal of Political Science*, 2024. forthcoming.
- L. Huddy, L. Mason, and L. Aarøe. Expressive partisanship: Campaign involvement, political emotion, and partisan identity. *American Political Science Review*, 109(1):1–17, 2015.
- S. Iyengar and M. Krupenkin. Partisanship as social identity; implications for the study of party polarization. In *The Forum*, volume 16, pages 23–45. De Gruyter, 2018.
- J. J. Jones. A dataset for the study of identity at scale: Annual prevalence of American Twitter users with specified token in their profile bio 2015–2020. *PLoS ONE*, 16(11):e0260185, 2023. doi: 10.1371/journal.pone.0260185.
- A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.
- B. E. Keith, D. B. Magleby, C. J. Nelson, E. A. Orr, and M. C. Westlye. *The myth of the independent voter*. Univ of California Press, 1992.
- D. R. Kinder, G. S. Adams, and P. W. Gronke. Economics and politics in the 1984 American presidential election. *American Journal of Political Science*, pages 491–515, 1989.
- S. Klar. The influence of competing identity primes on political preferences. *The Journal of Politics*, 75(4):1108–1124, 2013.
- M. H. Kuhn and T. S. McPartland. An empirical investigation of self-attitudes. *American sociological review*, 19(1):68–76, 1954.
- C. W. Leach, M. Van Zomeren, S. Zebel, M. L. Vliek, S. F. Pennekamp, B. Doosje, J. W. Ouwerkerk, and R. Spears. Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification. *Journal of personality and social psychology*, 95(1):144, 2008.
- T. Lee. Between social theory and social science practice. Toward a new approach to the survey measurement of ‘race’. In R. Abdelal, Y. M. Herrera, A. I. Johnston, and R. McDermott, editors, *Measuring Identity: A Guide for Social Scientists*, pages 113–44. Cambridge University Press Cambridge, UK, 2009.
- J. Li, G. Longinos, S. Wilson, and W. Magdy. Emoji and self-identity in Twitter bios. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 199–211, 2020.
- D. B. Magleby, C. J. Nelson, and M. C. Westlye. The myth of the independent voter revisited. *Facing the challenge of democracy: Explorations in the analysis of public opinion and political participation*, pages 238–263, 2011.

- M. F. Margolis. How politics affects religion: Partisanship, socialization, and religiosity in America. *The Journal of Politics*, 80(1):30–43, 2018.
- T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.
- M. Mosleh, C. Martel, D. Eckles, and D. G. Rand. Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proceedings of the National Academy of Sciences*, 118(7), 2021.
- Y. Mounk. *The Identity Trap. A Story of Ideas and Power in Our Time*. Penguin Press, 2023.
- S. Mukerjee, K. Jaidka, and Y. Lelkes. The political landscape of the US Twittersverse. *Political Communication*, 39(5):565–588, 2022.
- D. C. Mutz. Cross-cutting social networks: Testing democratic theory in practice. *American Political Science Review*, 96(1):111–126, 2002.
- M. Osmundsen, A. Bor, P. B. Vahlstrup, A. Bechmann, and M. B. Petersen. Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*, 115(3):999–1015, 2021.
- M. Osnabrügge, S. B. Hobolt, and T. Rodon. Playing to the gallery: Emotive rhetoric in parliaments. *American Political Science Review*, 115(3):885–899, 2021.
- J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL <http://www.aclweb.org/anthology/D14-1162>.
- Pew Research Center. Sizing up Twitter users, 2019. URL <https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/>. Accessed June 2022.
- Pew Research Center. 70% of U.S. social media users never or rarely post or share about political, social issues, 2021. URL <https://www.pewresearch.org/short-reads/2021/05/04/70-of-u-s-social-media-users-never-or-rarely-post-or-share-about-political-social-issues/>. Accessed June 2023.
- Pew Research Center. Jobs, hobbies top the list of things U.S. adults put in their twitter profiles; references to politics relatively rare, 2022. URL <https://www.pewresearch.org/short-reads/2022/05/05/jobs-hobbies-top-the-list-of-things-u-s-adults-put-in-their-twitter-profiles-references-to-politics-relatively-rare/>. Accessed July 2023.
- Pew Research Center. Partisanship by gender, sexual orientation, marital and parental status, 2024. URL <https://www.pewresearch.org/politics/2024/04/09/partisanship-by-gender-sexual-orientation-marital-and-parental-status/>. Accessed August 2025.

- V. Price, J. N. Cappella, and L. Nir. Does disagreement contribute to more deliberative opinion? *Political communication*, 19(1):95–112, 2002.
- M. Prior, G. Sood, K. Khanna, et al. You cannot be serious: The impact of accuracy incentives on partisan bias in reports of economic perceptions. *Quarterly Journal of Political Science*, 10(4):489–518, 2015.
- D. Quelle and A. Bovet. Bluesky: Network topology, polarization, and algorithmic curation. *PLOS ONE*, 20(2):e0318034, 2025. doi: 10.1371/journal.pone.0318034.
- S. Rathje, J. J. Van Bavel, and S. Van Der Linden. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26):e2024292118, 2021.
- R. Rehurek and P. Sojka. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011.
- A. Reid and K. Deaux. Relationship between social and personal identities: Segregation or integration. *Journal of personality and social psychology*, 71(6):1084, 1996.
- Reuters Institute. Here’s what our research says about news audiences on Twitter, the platform now known as X, 2023. URL <https://reutersinstitute.politics.ox.ac.uk/news/heres-what-our-research-says-about-news-audiences-twitter-platform-now-known-x>. Accessed October 2023.
- L. Rheault and C. Cochrane. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1):112–133, 2020.
- D. Rice, J. H. Rhodes, and T. Nteta. Racial bias in legal language. *Research & Politics*, 6(2):2053168019848930, 2019.
- E. Rodman. A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis*, 28(1):87–111, 2020.
- P. L. Rodriguez and A. Spirling. Word embeddings: What works, what doesn’t, and how to tell the difference for applied research. *The Journal of Politics*, 84(1):000–000, 2022.
- N. Rogers and J. J. Jones. 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, 2021.
- C. Sedikides, L. Gaertner, and E. M. O’Mara. Individual self, relational self, collective self: Hierarchical ordering of the tripartite self. *Psychological Studies*, 56(1):98–107, 2011.
- K. Sylwester and M. Purver. Twitter language use reflects psychological differences between Democrats and Republicans. *PloS one*, 10(9):e0137422, 2015.
- H. Tajfel and J. C. Turner. An integrative theory of intergroup conflict. In W. G. Austin and S. Worchel, editors, *The Social Psychology of Intergroup Relations*, pages 33–47. Brooks/Cole, Monterey, CA, 1979.

- C. Tausanovitch and C. Warshaw. Representation in municipal government. *American Political Science Review*, 108(3):605–641, 2014.
- J. Terrell, A. Kofink, J. Middleton, C. Rainear, E. Murphy-Hill, C. Parnin, and J. Stallings. Gender differences and bias in open source: pull request acceptance of women versus men. *PeerJ Computer Science*, 3:e111, 2017. doi: 10.7717/peerj-cs.111.
- R. Tromble. Where have all the data gone? a critical reflection on academic digital research in the post-api age. *Social Media+ Society*, 7(1):2056305121988929, 2021.
- L. Tucker and J. Jones. Pronoun lists in profile bios display increased prevalence, systematic co-presence with other keywords and network tie clustering among us twitter users 2015-2022. *Journal of Quantitative Description: Digital Media*, 3, 2023.
- U.S. Census Bureau. Educational attainment. U.S. Census Bureau, 2022. URL [https://data.census.gov/table?q=S1501:+Educational+Attainment&g=010XX00US\\$0400000_040XX00US01](https://data.census.gov/table?q=S1501:+Educational+Attainment&g=010XX00US$0400000_040XX00US01). Archived from the original on 19 September 2022. Retrieved 18 September 2022.
- C. Wilson, D. Johnson, and P. Rebal. Are you a J. Crew Democrat or a Pizza Hut Republican?, 2014. URL <http://time.com/3559482/stores-politics/>.
- X. Yan and L. Li. Censoring the intellectual public space in china: What topics are not allowed and who gets blacklisted? *Perspectives on Politics*, 22(3):753–770, 2024.