How Does Stereotype Content Differ across Data Sources?

Kathleen C. Fraser, Svetlana Kiritchenko, and Isar Nejadgholi
National Research Council Canada, Ottawa, Canada
{kathleen.fraser,svetlana.kiritchenko,isar.nejadgholi}@nrc-cnrc.gc.ca

Abstract

For decades, psychologists have been studying stereotypes using specially-designed rating scales to capture people's beliefs and opinions about different social groups. Now, using NLP tools on extensive collections of text, we have the opportunity to study stereotypes "in the wild" and on a large scale. However, are we truly capturing the same information? In this paper we compare measurements along six psychologically-motivated, stereotype-relevant dimensions (Sociability, Morality, Ability, Assertiveness, Beliefs, and Status) for 10 groups, defined by occupation. We compute these measurements on stereotypical English sentences written by crowd-workers, stereotypical sentences generated by ChatGPT, and more general data collected from social media, and contrast the findings with traditional, survey-based results, as well as a spontaneous word-list generation task. We find that while the correlation with the traditional scales varies across dimensions, the free-text data can be used to specify the particular traits associated with each group, and provide context for numerical survey data.

1 Introduction

There is growing interest in the possibility of using NLP and large corpora to augment, complement, or even replace traditional psychological surveys to collect social sciences data (Goldstone and Lupyan, 2016; Argyle et al., 2022; Jackson et al., 2022; Dillion et al., 2023). One area where NLP research has started to contribute is in the study and analysis of *stereotypes*.

Stereotypes are "a set of cognitive generalizations (e.g., beliefs, expectations) about the qualities and characteristics of the members of a group or social category" (VandenBos, 2007). There are a number of properties of stereotypes that motivate the use of NLP tools to better study and understand them. First, stereotypes are often communicated and perpetuated through natural language (Beukeboom and Burgers, 2019). Second, they are by definition widely-held and pervasive, and so should be detectable in large samples of data (Garg et al., 2018). Third, they can lead to far-reaching negative consequences, and so there is practical interest in understanding how stereotypes are expressed "in the wild" in order to develop effective counter-strategies (Fraser et al., 2021). NLP researchers have begun to study methods of uncovering stereotype information in Twitter data (Marzouki et al., 2020; Fokkens et al., 2018), news texts and books (Garg et al., 2018), spoken conversations (Charlesworth et al., 2021), and large language models (Cao et al., 2022).

However, the question remains whether the information we can extract from these natural language datasets can actually replicate the information obtained from more traditional methods in social psychology; namely, rating scales. A common paradigm in stereotype research involves choosing a set of attributes, or dimensions, of interest, and then asking human participants (often college undergraduates) to rate social groups along those dimensions. The dimensions of interest vary according to different theoretical models, but can include, for example, *warmth* and *competence* in the Stereotype Content Model (Fiske et al., 2007), or *agency*, *beliefs*, and *communion* in the ABC Model (Koch et al., 2016). The social groups may be categorized based on gender, race, age, or any other social variable relevant to the research. As a result, for each social group, the researchers obtain annotations along each dimension.

In this work, we investigate the possibility of reproducing the results of such a scale-based study, using low-dimensional vector representations of natural language data to estimate the dimensions of interest. We consider six psychologically-motivated dimensions – Sociability, Morality, Ability, Assertiveness, Status, and Beliefs[1] – and a set

[1] See Appendix A for detailed definitions of each dimen-

*Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 18 - 34
June 20-21, 2024 ©2024 Association for Computational Linguistics

of ten groups defined by occupation. We conduct a detailed comparison of the kind of stereotype data we obtained through (1) direct stereotype elicitation from crowd-workers, (2) direct stereotype elicitation from a generative large language model, and (3) targeted data collection from Twitter (now known as ‘X’). We compare these sources of information to two paradigms in the psychology literature: the traditional method using rating scales, as mentioned above, and a newer method involving spontaneous word list elicitation. We consider three research questions in the current study:

1. Can we reproduce the numerical, scale-based results from the social psychology literature through analysis of natural language? We explore this question using three different sources of text: crowd-workers, social media, and ChatGPT, and by transforming the data to a 6-dimensional representation such that each dimension corresponds to a scale measure.
2. Are all of the six aforementioned dimensions spontaneously mentioned in the free text, or are certain dimensions more frequently discussed than others?
3. Are there certain types of information which are available only from the ratings scales, or only in the natural language data? Or can we treat them equivalently?

Our findings suggest that particular dimensions can be estimated more reliably than others, with Morality and Status measurements being highly correlated with the traditional scales on all of the text datasets. The dimensions of Assertiveness and Beliefs were less accurately estimated; statements relevant to these dimensions were also less frequent in the data. However, the natural language texts were found to contain additional types of information not available in the scale-based dataset, adding detail and specificity to the stereotype descriptions.

2 Background

2.1 Psychological Models of Stereotypes

Stereotyping is an extensive area of research in social psychology. Numerous models have been developed to explain the underlying dimensions of social cognition, including stereotyping (Fiske et al., 2007; Koch et al., 2016; Abele and Wojciszke, 2007). Regardless of the specific dimensions in question, the measurements have almost always been collected using scales or checklists (i.e., a *forced-choice* paradigm).

One recent study has questioned whether the exclusive use of forced-choice methods has limited, or even biased, the resulting information about how different social groups are viewed. Nicolas et al. (2022) propose a Spontaneous Stereotype Content Model, arguing that “free-response, open-ended stereotypes of social groups may best systematically reveal the complex contents that are spontaneously available to perceivers upon encountering a target.” For a given dimension, the authors distinguish between *direction* (e.g., is the group perceived as friendly or unfriendly), which is measured directly by the scales and can be inferred from the open-ended responses, and *representativeness*, which measures how strongly a given dimension is associated with a group (regardless of polarity). In an example from Nicolas et al. (2022), doctors and nurses are both rated as being highly Warm and Competent on rating scales. However, when people spontaneously think about doctors and nurses, they think more about nurses’ Warmth traits, and more about doctors’ Competence traits. Such differences cannot be observed using the traditional, scale-based methods.

Nicolas et al. compare traditional, scaled-based methods against open-ended responses in the form of single words, and sets of words. We use their data as a baseline, and build on this basic premise by extending the types of open-ended responses to include full sentence stereotypes (generated either by humans or ChatGPT), and then further extending the analysis to the case of Twitter data (which is not specifically stereotypical in nature, but represents a large sample of public opinions on various topics).

2.2 NLP Methods for Analyzing Stereotypes

Numerous NLP methods have been used to extract, discover, and track stereotype content in naturally-occurring texts (Marzouki et al., 2020; Fokkens et al., 2018; Garg et al., 2018; Charlesworth et al., 2021; Fast et al., 2016). In some cases, stereotyping has been labelled as a subcategory of hate speech or offensive language, including gender stereotypes (Chiril et al., 2021; Parikh et al., 2019; Fersini et al., 2018) and stereotypes about immigrants (Sanguinetti et al., 2018; Sánchez-Junquera et al., 2021). For example, the EVALITA 2020 Hate Speech Detection Task involved a subtask

sion.

	Stereotypes?	All dimensions?	Human-generated?	Contextual?
Scales	✓	✓	✓	✘
Adjectives	✓	✘	✓	✘
Stereoset	✓	✘	✓	✓
ChatGPT	✓	✘	✘	✓
Tweets	✘	✘	✓	✓

Table 1: Summary of some relevant differences between the various data sources under consideration.

on detecting stereotypes targeting Muslims, Roma, and immigrants (Sanguinetti et al., 2020). Other closely-related work has compared stereotypical biases in large language models with human survey data (Cao et al., 2022). Our work is most similar to that of Fraser et al. (2022), which presents a computational model of Fiske et al.'s Stereotype Content Model (SCM), using the POLAR framework introduced by Mathew et al. (2020). We make use of a similar method to define an interpretable, psychologically-motivated, low-dimensional embedding space.

Other relevant NLP work has examined the verbs and adjectives which are mostly highly associated with certain social groups. Dong et al. (2019) collected words describing various social 'roles' from crowd-workers from different cultures, and also used NLP methods to predict the most likely social role, given a descriptor. Choenni et al. (2021) probed the stereotypes present in pretrained language models with prompts such as "Why are [TARGET GROUP] so [MASK]?" and observed the output attributes.

While similar in spirit to some of these earlier works, our work differs critically in our goal of trying to map *natural language sentences* down to six *numerical dimensions*, for direct comparison against the social psychology rating scales. Furthermore, we compare and contrast these different ways of collecting stereotypical beliefs to explore the types of information available from each source.

3 Methods

In the following section, we describe several different sources of survey and natural language data in English, namely: psychological rating scales (Sec 3.1) as well as lists of spontaneously-produced adjectives, crowd-sourced stereotypes from the Stereoset dataset (Nadeem et al., 2020), stereotypes prompted from ChatGPT, and tweets from Twitter (Sec 3.2). These data sources differ in many relevant aspects, summarized in Table 1. For example, were the writers of the text asked specifically to come up with stereotypes, or are they writing on a more general topic, that may or may not convey implicit stereotypes? Were the annotators required to make a judgement on every dimension, or did they comment only on the dimensions that most easily came to mind? Was the text generated by humans or by a language model? And does the format of the text provide context for the attributes being assigned, or must they be interpreted in isolation? We will discuss these aspects in relation to each dataset in the following.

To make a direct comparison across all the data sources, we first identify the subset of social groups for which data is available in all the existing datasets. The majority of this subset consists of different occupations: Politicians, Teachers, CEOs, Scientists, Bankers, Accountants, Engineers, Farmers, Lawyers, and Nurses. Thus we consider only these 10 target groups in the analysis.

Following our discussion of the datasets, in Section 3.3 we present the dimensionality-reduction method we use to reduce the free-text sentences in the four natural language datasets down to six dimensions, so that they can be compared directly to the 6-dimensional gold standard rating scale data.

3.1 Gold Standard Rating Scales

The gold-standard rating scale values are obtained from the supplemental materials for Experiment 1 in Nicolas et al. (2022). In that experiment, 400 Amazon Turk workers provided annotations for 43 social groups. Each annotator saw a random sample of six groups, and for each group provided six open-ended, free text responses describing "characteristics, traits, or descriptions of the group." Annotators were additionally informed that it was not necessary that they *personally* believe these characteristics to be true, in order to reduce social desirability bias. Most responses are single adjectives.

After annotators provided their free text responses, they were asked to provide a rating from 1 through 5 for "how society views the targets" along various dimensions: Sociability (measured by two subscales, *friendly* and *sociable*), Morality (*trustworthy* and *honest*), Ability (*competent* and *skilled*), Assertiveness (*confident* and *assertive*), Beliefs (*traditional* and *conservative*), and Status (*wealthy* and *high-status*).

In our analysis, we combine the two subscales for each dimension, and normalize the values to lie between -1 and +1, for better comparison with our computational models. We average the annotations for each group over all annotators (on average, 57 per group).

3.2 Alternative Data Sources

Spontaneous Adjectives As the first alternative data source, we consider the adjectives from Experiment 1, described above (Nicolas et al., 2022). The sets of adjectives represent an intermediate step between the rating scales and the spontaneously-produced sentences in the rest of the data sources. Additionally, the adjectives were provided by precisely the same annotators as the scale-based ratings. Thus, the information conveyed by the adjectives likely represents an upper bound for how well we can reproduce the scale ratings via language. Since our NLP analysis (described in Section 3.3) operates on the sentence level, we embed each adjective into a sentence template of the form: `These people are always ADJ`.

StereoSet We also consider data from the StereoSet dataset (Nadeem et al., 2020). This dataset was crowd-sourced on Mechanical Turk. [2] Annotators were asked to generate sentences about a particular group which were (1) stereotypical, (2) anti-stereotypical, and (3) neutral. In this work, we use the stereotypical sentences. There are approximately 55 sentences per target group. These data differ from the adjective sets in that they consist of complete sentences, of varying length and complexity. However, they were still generated in an artificial scenario, with the goal of communicating stereotype information.

ChatGPT As an additional source of data, we generate novel sentences using ChatGTP.[3] Unlike the other data sources, this text does not originate from human authors. However, other researchers have begun exploring the possibility of using large language models as potential sources of information for studying bias and stereotypes (Cao et al., 2022), or even as replacements for human participants in psychological studies (Argyle et al., 2022; Dillion et al., 2023).

We consider three prompts to ChatGPT: (1) `What are some adjectives people in North America use to describe GROUP?` This prompt attempts to directly replicate the open-response portion of Experiment 1 from Nicolas et al. (2) `In North America, what are some commonly held stereotypes about GROUP?` This prompt attempts to directly elicit stereotypes about various groups. (3) `What are some beliefs that many North Americans hold about GROUP?` After observing that many of the generations for the previous prompt focused on negative beliefs about groups, we added this prompt to elicit more neutral/positive characteristics. We re-run each prompt three times for each group, with the default temperature. Each response from ChatGPT contains a list of characteristics, each taken as a separate observation, resulting in an average of 81 sentences for each group.

Twitter Finally, we consider Twitter as a potential source of data about social groups. One significant difference between this dataset and the others is that the writers of the texts were not instructed to generate stereotypes, but rather had other communicative goals in mind. Another factor that may affect the Twitter data is *social desirability bias*. While someone might hold a belief privately, and even report it on an anonymous survey, it doesn't necessarily mean they will state that belief openly on a public forum. However, our hypothesis is that if we have a large data sample, the most common beliefs about different groups should emerge.

We used the Research API[4] to collect data containing the substring 'GROUP are' for the target groups of interest, from 1 January 2022, to 7 October 2022. We ignored re-tweets, duplicates, tweets with more than five hashtags, tweets with URLs, and tweets written by bots (user name or description contains 'bot') and other prolific users. This resulted in a large number of tweets, on average 118,768 per group.

To increase the likelihood of capturing relevant tweets, we then performed the following filtering steps: (1) filter by the user 'location' field to include only those tweets from the US and Canada;

[2] The annotators were all located in the USA, and the stereotypes were validated by an independent set of annotators to ensure that they represented commonly-held views.

[3] `https://chat.openai.com/chat`, GPT-3.5, September 25 2023 version

[4] Prior to the introduction of the data paywall.

(2) parse the sentence and include only those sentences where the target group is not modified by a quantifier or adjective (*Some lawyers are ...*, *Republican politicians are ...*), (3) using the sentence parse, include only those sentences where *are* is followed by an adjective (e.g., keep *Nurses are angry*, but discard *Nurses are going to go on strike*). This last filtering step is based on research that stereotype-consistent information tends to be communicated with abstract terms, like adjectives, while concrete terms like action verbs describe a particular, contextual behaviour that is not necessarily an essential trait that is present across situations (Beukeboom and Burgers, 2019). These filtering steps drastically reduce the amount of data available (to an average of 2,830 tweets per group), but with the goal of increasing the relevance.

3.3 POLAR Model

Here, we describe our methodology for embedding the text sentences into the six-dimensional social space. For each sentence, we begin by masking the target group name with the generic phrase *these people*. This is to avoid any bias in the sentence embeddings related to the group name (e.g., we want *Scientists are smart* and *Nurses are smart* to map to the same point, regardless of any intrinsic bias in the embedding model related to scientists and nurses). We represent each input sentence as a 1024-dimensional RoBERTa sentence embedding, and then reduce the embedding space to the six dimensions of interest using a variation on the method described by Fraser et al. (2022). The mathematical details are given in Appendix B, but essentially the method is as follows: For each dimension, collect a set of examples to define each pole of the axis. Here, since we want to reproduce the scale ratings of Nicolas et al. (2022), we use the same adjectives that were presented to the participants during data collection (e.g., for the dimension Sociability, they were shown *friendly* and *sociable*, for Morality they were shown *trustworthy* and *honest*, and so on). To define the negative pole, we used the direct antonym according to our own judgement (e.g., *unfriendly*, *unsociable*, *untrustworthy*, and *dishonest*). We then inserted those adjectives into the sentence template `These people are always ADJ`, to generate representative stereotypical sentences for the two poles of each dimension.

The positive examples are then averaged to define the positive direction, and the negative examples are averaged to define the negative direction. The difference between the positive and negative vectors, for each dimension, is then used to define a transformation matrix such that sentence embeddings in the high-dimensional embedding space can then be projected onto the interpretable, six-dimensional space. The dimension score for each sentence is simply the scalar projection of the sentence onto that dimension, ranging from -1 to 1. For each group, we then obtain the average dimension ratings over all sentences in the dataset.

The POLAR model has a small number of parameters that should be optimized for best performance. We validate the model on a hand-crafted lexicon of adjectives for each dimension (Nicolas et al., 2021). Our optimized model uses RoBERTa-NLI embeddings[5], Partial Least Squares (Rosipal and Krämer, 2005) to initially reduce the embedding dimensionality from 1024 to 30, and achieves an average accuracy of 95% at correctly predicting whether each word is positively or negatively associated with the relevant dimension. Further information about the validation process is available in Appendix C.

3.4 Word-Counting Baseline

We also consider a word-counting baseline. Although word-counting tends to be less effective in assessing sentence-level meaning due to negation, sarcasm, etc. (Fraser et al., 2022), we can use this as a baseline method in the case of the adjective lists. Nicolas et al. (2021) provides a set of lexicons for various psychologically-motivated dimensions, including the six dimensions studied here. Words in each lexicon are assigned either a positive (+1) or negative (-1) value according to their direction. Thus, the estimated score for each group on a given dimension is simply the average of all the lexicon values for the words associated with each group (ignoring words that are not in the lexicon for that dimension).

4 Results

4.1 Correlation with Rating Scales

To compare the scores from the text data sources with the gold-standard scale ratings, we measure correlation. Because the most important information is the *relative* differences between the groups,

[5] `https://huggingface.co/sentence-transformers/nli-roberta-large`

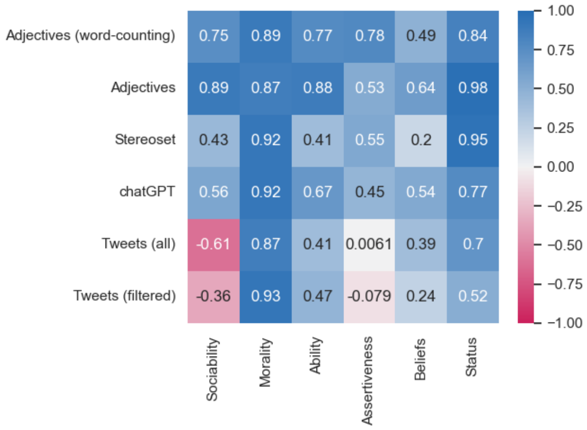


Figure 1: Spearman rank correlation with the scale-based measurement, for each dimension and dataset.

rather than absolute values, we compute Spearman's rank correlation. Correlation values for each dimension and each data source are shown in Figure 1 (full correlation matrices in Appendix D).

We begin by observing that the adjectives, elicited at the same time as the scales, are generally good (though not perfect) at approximating the scale values, and that our POLAR model is, in most cases, more effective than the word-counting approach at associating the adjectives with the scale values (first and second rows of Figure 1). One exception to both of these observations occurs in the case of Assertiveness, where our model achieves a correlation of only 0.53 with the scale values. As an example, we examine the data for farmers, the group ranked lowest on Assertiveness in the scale data, but second-highest in the adjectives data. The main underlying cause of the divergence seems to be that annotators interpreted the "Assertive" trait rather narrowly, as being *pushy* or *demanding*. However, when we look at the adjectives, many people mentioned words like *hard-working* or *strong*, which are also associated with Assertiveness in our model. As a result, farmers are rated higher than most other groups on this dimension.

Moving on to the free-text data sources, we observe that some dimensions are estimated more consistently across data sources. Morality in particular shows very high correlation across all data sources. Whether someone is judged as friend or foe, good or bad, has evolutionary significance and forms the basis of many of our social interactions (Fiske et al., 2007). Therefore it is not surprising that many of the data sources mention morality-related traits (more on this in Sec 4.2) and tend to agree on the direction and relative magnitude of those traits for different groups.

The estimates for Sociability show a somewhat different pattern, with the ChatGPT achieving a moderate correlation of 0.56, and Stereoset somewhat lower at 0.43. In the case of the Twitter data however, the correlation with the scales is actually negative. There are many possible explanations for this, stemming from the heterogeneity and diversity of topics in the Twitter dataset. For example, the scales rate nurses as high-Sociability and accountants as low-Sociability. Many of the tweets expressing low-Sociability traits in nurses are written in the context of the COVID-19 pandemic, such as *Nurses are frustrated and tired* or *Nurses are not ok!*. Conversely, some of the tweets expressing high Sociability for other groups are likely sarcastic, e.g. *Accountants are super fun haha*. In Sec 4.3, we perform topic modelling to disaggregate the different topics so they can be examined separately.

Considering now Assertiveness and Ability, sometimes considered two facets of a single dimension "Competence," we again observe a divergence in the results, with Ability estimates being more highly correlated with the scale ratings for all data sources except Stereoset. This may be an artifact of our particular dataset, as the Ability dimension is particularly relevant in the context of occupations. We also observe that in the Twitter data, groups with high Assertiveness on the traditional scales are often criticized as being ineffectual, e.g. *All politicians are spineless.*

For Beliefs, all data sources have only moderate correlation with the scales. In fact, Nicolas et al. (2022) found that very few of the spontaneously produced adjectives (around 5%) carried information about the Beliefs dimension. The data generated by ChatGPT has the best correlation score of the free-test data sources, specifically labelling accountants, bankers, and farmers as conservative.

Finally, the Status dimension shows reasonably high correlation between the scales and the text data. Again, this may be related to the fact that all of our target groups are based on occupation: in all data sources, we observe statements about CEOs and lawyers being rich, and teachers and nurses being underpaid.

4.2 Prevalence of Each Dimension

We now analyze how many of the texts in each dataset are directly relevant to each dimension. Un-

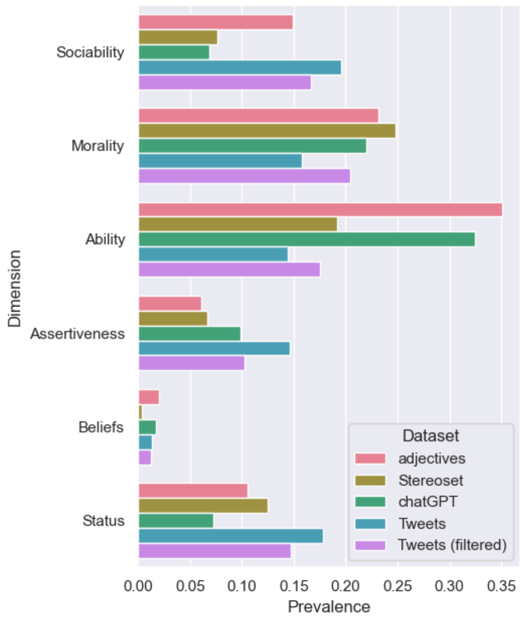


Figure 2: The proportion of text instances assigned an absolute value greater than 0.5. for each dimension.

like in the scale-based paradigm, there may be certain dimensions that simply are not mentioned, leading to difficulties in generating an accurate estimation. This is related to Nicolas et al.'s concept of *representativeness* (Section 2), except that we calculate it over all groups (for the results separated by group, see Appendix E).

Figure 2 shows the proportion of texts in each dataset that are assigned an absolute value greater than (or equal to) 0.5 on each dimension.[6] As hypothesized in the previous section, many of the sentences express Ability judgments, as expected when discussing groups based on occupation. The Morality dimension is mentioned quite often, consistent with the findings of Nicolas et al. (2022). A very small proportion of texts are relevant to the dimension of Beliefs, in all datasets.

However, we note that the trends do look different when considered on a group-by-group basis (Fig D.1). For example, Morality is mentioned in a much higher proportion of texts about politicians. Similarly, the Status dimension is described more frequently in texts about CEOs, bankers, and lawyers. The Ability dimension is the most prevalent dimension when discussing scientists, engineers, and accountants, while for teachers we observe that Ability and Sociability traits are mentioned equally often. The Belief dimension is brought up slightly more in texts about farmers (often described as being conservative).

Group	Mor.	Soc.	Abil.	Ass.	Bel.	Stat.
Politicians	**-0.60**	0.17	-0.11	**0.50**	0.08	0.43
Teachers	**0.56**	**0.53**	0.46	0.30	0.26	-0.30
CEOs	-0.14	0.13	**0.50**	**0.64**	0.23	**0.73**
Scientists	0.48	0.04	**0.81**	0.49	-0.19	0.21
Bankers	-0.20	0.10	0.40	0.43	0.39	**0.59**
Accountants	0.29	0.02	**0.59**	0.32	0.43	0.22
Engineers	0.48	0.15	**0.86**	0.44	0.19	0.49
Farmers	**0.60**	0.36	0.46	0.23	**0.63**	-0.43
Lawyers	-0.47	-0.09	**0.50**	**0.60**	0.20	**0.64**
Nurses	**0.56**	**0.57**	**0.69**	0.40	0.23	-0.16

Table 2: Dimension estimates for each group, from the scale data, with most salient dimensions in boldface.

4.3 Topic Modeling

As we have seen in Section 4.1, our estimates of relevant psychological dimensions from text do *not* perfectly reproduce those obtained through traditional survey-based methods. However, the survey-based methods also have limited interpretability. For example, Nicolas et al. (2022) found in their original study that the limited set of dimensions did not always align well with people's perceptions of groups. When annotators were asked, "Which of the following characteristics fits best what you meant by [response]?" and given a choice of dimensions (Assertive, Friendly, etc.), "No Match" was actually the most common response. So when forced to make a choice, the annotators might rate politicians as being Sociable (because they are charismatic), but it doesn't really mean the same thing as rating nurses as highly Sociable (because they care deeply about other people). Therefore, in this section, we propose to use natural language resources as *complementary* data to explain and differentiate between the ratings obtained on the six-dimensional scales.

Our procedure is as follows: for each group, we defined the most 'salient' dimensions of the group stereotype as those dimensions with an average absolute scale-based estimate of 0.5 or greater (corresponding to an average response on the original survey of less than 2/5, or greater than 4/5). These dimensions are indicated with boldface in Table 2. We then seek to provide evidence, or further elucidation, of those dimensions by examining the topics arising in the free-text data sources.

For the topic modelling, we employ BERTopic

[6] The threshold of 0.5 was chosen based on the validation set data, where it was observed that a score of 0.5 roughly differentiated the words associated with each dimension from words associated with other dimensions.

(Grootendorst, 2022), which uses the HDBSCAN clustering algorithm to remove outliers and concentrate on the most densely populated areas of the embedding space. This aligns with our understanding of stereotypes as being widely-held beliefs, rather than idiosyncratic opinions about a group.

Here, we want to find those topics that help explain the rating scales. Therefore, we then compute the centroid of each topic in the sentence embedding space, and then project the centroid down to the six-dimensional space using the same POLAR model. This allows us to compare the topics along the same dimensions as the rating scales.

We do not expect any single topic to be relevant to all six dimensions simultaneously; rather, we examine one dimension at a time, focusing on the most salient dimensions for each group (as defined above). For a given dimension, we first select all topics where the centroid projection has the same sign as the scale-based score. If there are multiple topics, we rank them according to their centroid projection along that dimension and keep the top three topics (i.e., three most positive or most negative) to analyze. These topics should be the most relevant to understanding why the group would be rated as they were along that dimension. Extended results are given in Appendix F, but we consider several illustrative cases in Table 3:

Differentiating similar groups One way that the text data can be useful is to provide information that differentiates groups that are similarly ranked along a given dimension. For example, scientists, CEOs, and nurses all have high Ability as a salient dimension. However, by examining the text data, we observe qualitative differences in what *aspects* of Ability stereotypically apply to each group (Table 3, Examples 1–3).

Increasing specificity of a stereotype In other cases, even within a particular group, looking at the text data gives a much more specific interpretation of the stereotype. In Example 4 in Table 3, we see that the stereotype of politicians as being low-Morality has a more precise interpretation: i.e., politicians are specifically seen as *corrupt*.

Different responses to stereotypes In other cases, even when there is agreement on the relevance of a dimension in the scale-based data, the text data can reveal different interpretations of that value. In Example 5 (Table 3), we see that teachers are rated as high-Morality. The related topic in the StereoSet data portrays this as *kindness*, while the high-morality topic in the ChatGPT data describes teachers as *strict* and concerned with *discipline*.

Finally, we briefly consider the set of topics not included in the above analysis; that is, those topics which are not strongly associated with one of the salient dimensions. As Nicolas et al. (2022) argue, not all of our social judgements are captured by the dimensions typically studied in social psychology. Aspects of social judgement not directly captured in the six dimensions used here include appearance, gender, and ethnicity, among others.

Table 4 shows examples of some common stereotypes which appear in the text data and are surfaced by the topic modelling, but are not identified with a salient dimension in Table 2. In Example 1 we see the stereotype that nurses are always women, as well as the associated stereotype of the "sexy nurse." In Example 2, in contrast, we see that scientists are stereotyped as being male. In that example, as well as in Example 3, we also see the stereotype that scientists and engineers are "nerdy." Nicolas et al. (2022) identified Appearance as one factor orthogonal to the original scales, and we find some evidence for a stereotype of bankers as being sharply dressed (Ex 4). More concerning, the Stereoset and Twitter data also reveal the offensive stereotype that all bankers are Jewish (Ex 5).

While we have presented only a handful of examples, many of the stereotypes in Table 4, which are not visible in the scale-based data, are arguably more harmful than those summarized in Table 3. Thus the information available in the free text responses provides a valuable and complementary perspective on how certain groups are stereotyped.

5 Conclusion

We conducted a comparison of stereotypes of 10 occupational groups across traditional survey-based data and various text sources. The answers to our research questions are summarized as follows: (1) While our computational analysis of the text data does not perfectly replicate the scale data, certain dimensions (e.g., Morality and Status) could be consistently estimated, with correlation values greater than 0.5 across all text sources. (2) This finding may be partially attributed to the fact that not all dimensions are mentioned spontaneously in the data sources we considered: in particular, adjectives and statements relating to the Beliefs dimension were infrequent. (3) Following from

	Group	Dim.	Scales	Adjectives	StereoSet	ChatGPT	Twitter
1	Scientists	Ability	0.81	thoughtful, observant, careful	the scientist who developed this formula must have been smart	scientists are intelligent and analytical	scientists are usually inherently sane and intelligent
2	CEOs	Ability	0.50	savvy, cunning	he was ruthless and cunning	CEOs are strategic and decisive	CEOs are important to a business
3	Nurses	Ability	0.69	hard-working, tending, caring	a nurse has to be supportive to do their job well	nurses understand the physical and emotional needs of patients	nurses are essential to effective health care
4	Politicians	Morality	-0.60	sneaky, sly, lying	known for being corrupt	politicians are corrupt and dishonest	all politicians are corrupt at some level
5	Teachers	Morality	0.56	loyal, patient, strict	she was kind and maintained order in class	strict disciplinarians who enforced rules	these people are never inappropriate, abusive or violent

Table 3: Example stereotype words and sentences extracted using the topic modelling approach, associated with highly-salient dimensions in the scale-based data.

	Group	Adjectives	StereoSet	ChatGPT	Twitter
1	Nurses	female, women, girls	she had a very short skirt while changing the iv	nurses are primarily female	nurses are the sexiest women alive
2	Scientists	male	he was middle aged man and was socially awkward	historically, there have been stereotypes associated with gender	cause scientists are all male and lonely
3	Engineers	awkward, nerdy, shy	geeks who are not good at social interaction	engineers are “nerds” or “geeks”	engineers are so weird
4	Bankers	shoes, suits, well-dressed	he wears a crisp suit and makes a lot of money	bankers are always wearing suits	*(none)*
5	Bankers	*(none)*	the person was a Jewish man	*(none)*	the left always assumes bankers are exclusively Jewish

Table 4: Example stereotype words and sentences extracted using the topic modelling approach, which are associated with traits not easily mapped to the six dimensions previously considered.

this, the scale data do contain important information not available in the text data, as participants are required to make a judgement for each dimension. However, the natural language data sources also contain useful information not available in the scale-based data, distinguishing between different groups that were rated similarly along a given dimension, and increasing the specificity of the stereotypes. In particular, topic modelling revealed specific beliefs about groups that were not captured by the abstract, high-level, numerical responses. Free text responses were able to distinguish between different groups that were rated similarly along a given dimension, and increase the specificity of the stereotypes.

Deepening our understanding of stereotypes can help in the development of effective counter-strategies. The work presented in this paper can support these goals in a number of ways. For example, if we consider the ratings of scientists and engineers on the scale-based data, it is not entirely clear what an appropriate counter-example should be (a scientist with low morality and low ability?). However, the natural language data helps surface the more specifically harmful stereotype that scientists are all male and anti-social. Challenging that aspect of the stereotype is more likely to be effective at increasing women’s participation in science. At the same time, the scale-based data may provide information that is “hidden” in the social media data, such as the stereotypical idea that most farmers are religious and politically right-wing. This type of information, although essential in gaining a broader understanding of stereotypes, does not tend to be explicitly stated on social media. We also observed that the scale-based data, as well as the ChatGPT data, do not clearly communicate extremely negative or offensive stereotypes – even though these should be the highest priority for mitigation. Therefore, understanding the strengths and weaknesses of the information available in different datasets can have important real-world implications. Furthermore, future work could examine how the data from unconventional sources, such as social media or ChatGPT, may be used to augment more traditional sources, such as lexicons.

Limitations

In this study, we focused on English-language resources only. Further, the collected stereotypes in these resources (survey-based rating scales and word lists, StereoSet) may only be common in the North-American culture. Twitter has a biased demographic representation of users, with most users residing in the U.S. For a fair comparison, we also constrained the ChatGPT responses to the North-American context. Future studies should expand the language and cultural range of stereotype information, although data unavailability may pose a significant barrier.

We examined ten social groups based on occupation since they were common in all the considered data sources. However, stereotypes targeting groups based on other characteristics, such as gender, ethnicity, or socio-economic status, are also prevalent in online and offline communications and may result in severe consequences for the groups and the society at large. Future work should include a wide variety of social groups to investigate how well the results can generalize across the groups.

While social media presents a valuable data source for studying people's opinions and tracking common beliefs, the sheer volume of these data requires computational tools to process the data efficiently. In this study, we applied unsupervised topic modeling, but other unsupervised, semi-supervised, and supervised techniques should be explored and evaluated in this context and may result in different findings. Also, topic modeling and clustering methods tend to be sensitive to parameter settings, and re-running the analysis with different parameters may lead to different results.

Finally, the stereotype information in the different data sources was obtained from different population samples, each of which introducing its own sampling bias. Since for most data sources the information was collected as stereotypical beliefs *common in the society* (as opposed to individuals' beliefs), we expect the effects of sample bias to be small. Still, this may have contributed to the observed differences in findings. Complementary use of several data sources may provide a fuller and less biased view.

Ethics Statement

While collecting stereotype data is a necessary step in studying stereotyping, such resources could inadvertently propagate harmful beliefs or be misused by adversaries to target vulnerable populations. Another open issue is how to counter stereotypical beliefs and mitigate their negative effects. There is a tension between the right to free speech and respect for equality and dignity. Rigid prohibitive mechanisms (e.g., banning any stereotype information from public view) would likely be ineffective. Counter-strategies should work towards weakening stereotypical associations and emphasize the fact that individuals do not neatly fit in boxes prescribed by their demographic characteristics.

References

Andrea E Abele, Nicole Hauke, Kim Peters, Eva Louvet, Aleksandra Szymkow, and Yanping Duan. 2016. Facets of the fundamental content dimensions: Agency with competence and assertiveness—communion with warmth and morality. *Frontiers in Psychology*, 7:219720.

Andrea E Abele and Bogdan Wojciszke. 2007. Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, 93(5):751.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2022. Out of one, many: Using language models to simulate human samples. *arXiv preprint arXiv:2209.06899*.

Camiel J Beukeboom and Christian Burgers. 2019. How stereotypes are shared through language: a review and introduction of the social categories and stereotypes communication (scsc) framework. *Review of Communication Research*, 7:1–37.

Marco Brambilla, Patrice Rusconi, Simona Sacchi, and Paolo Cherubini. 2011. Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology*, 41(2):135–143.

Yang Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Tessa ES Charlesworth, Victor Yang, Thomas C Mann, Benedek Kurdi, and Mahzarin R Banaji. 2021. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, 32(2):218–240.

Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2021. “be nice to your wife! The restaurants are closed”’: Can gender stereotype detection improve sexism classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844.

Rochelle Choenni, Ekaterina Shutova, and Robert van Rooij. 2021. Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1477–1491, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*.

MeiXing Dong, David Jurgens, Carmen Banea, and Rada Mihalcea. 2019. Perceptions of social roles across cultures. In *Social Informatics: 11th International Conference, SocInfo 2019, Doha, Qatar, November 18–21, 2019, Proceedings 11*, pages 157–172. Springer.

Ethan Fast, Tina Vachovsky, and Michael Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 112–120.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2018. Overview of the evalita 2018 task on automatic misogyny identification (AMI). In *CEUR Workshop Proceedings*, volume 2263, pages 1–9. CEUR-WS.

Susan T Fiske, Amy JC Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11(2):77–83.

Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah, and Wouter Van Atteveldt. 2018. Studying Muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kathleen C Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2022. Computational modeling of stereotype content in text. *Frontiers in Artificial Intelligence*, 5.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616, Online. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Felipe L Gewers, Gustavo R Ferreira, Henrique F De Arruda, Filipi N Silva, Cesar H Comin, Diego R Amancio, and Luciano da F Costa. 2021. Principal component analysis: A natural approach to data exploration. *ACM Computing Surveys (CSUR)*, 54(4):1–34.

Robert L Goldstone and Gary Lupyan. 2016. Discovering psychological principles by mining naturally occurring data sets. *Topics in Cognitive Science*, 8(3):548–568.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Joshua Conrad Jackson, Joseph Watts, Johann-Mattis List, Curtis Puryear, Ryan Drabble, and Kristen A Lindquist. 2022. From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, 17(3):805–826.

Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. 2016. The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, 110(5):675.

Yousri Marzouki, Eliza Barach, Vidhushini Srinivasan, Samira Shaikh, and Laurie Beth Feldman. 2020. The dynamics of negative stereotypes as revealed by tweeting behavior in the aftermath of the charlie hebdo terrorist attack. *Heliyon*, 6(8):e04311.

Binny Mathew, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. 2020. The POLAR framework: Polar opposites enable interpretability of pretrained word embeddings. In *Proceedings of the Web Conference 2020*, pages 1548–1558.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Gandalf Nicolas, Xuechunzi Bai, and Susan T Fiske. 2021. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196.

Gandalf Nicolas, Xuechunzi Bai, and Susan T Fiske. 2022. A spontaneous stereotype content model: Taxonomy, properties, and prediction. *Journal of Personality and Social Psychology*.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1642–1652, Hong Kong, China. Association for Computational Linguistics.

Roman Rosipal and Nicole Krämer. 2005. Overview and recent advances in partial least squares. In *Proceedings of the International Statistical and Optimization Perspectives Workshop "Subspace, Latent Structure and Feature Selection"*, pages 34–51. Springer.

Javier Sánchez-Junquera, Berta Chulvi, Paolo Rosso, and Simone Paolo Ponzetto. 2021. How do you speak about immigrants? Taxonomy and StereoImmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences*, 11(8):3610.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. HaSpeeDe 2 evalita 2020: Overview of the evalita 2020 hate speech detection task. *Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Gary R VandenBos. 2007. *APA Dictionary of Psychology*. American Psychological Association.

A Stereotype Dimensions

We consider the same 6 psychological dimensions of stereotyping as Nicolas et al. (2022), to enable comparison against the ratings of the annotators in that study. These dimensions are: *Sociability, Morality, Ability, Assertiveness, Status,* and *Beliefs*. The dimensions are based on previous theories in the social psychology literature. Fiske et al. (2007) present the Stereotype Content Model (“SCM”), which posits that the two primary dimensions of stereotype content are Warmth and Competence. *Sociability* and *Morality* are two facets of Warmth, and *Ability* and *Assertiveness* are two facets of Competence. Koch et al. (2016) present a different, three-dimensional theory of stereotype content known as the “ABC Model,” where A = Agency, B = Beliefs, and C = Communion. While Communion is similar to the concept of Warmth, the other two dimensions diverge from the SCM, with Agency being related to socioeconomic *Status*, and *Beliefs* capturing progressive versus conservative values. To compare the SCM and ABC models, Nicolas et al. (2022) included all 6 distinct dimensions, as did we in the current work.

In the instructions to annotators, Nicolas et al. (2022) define the dimensions with adjectives, as shown in Table A.1. Additional information for each dimension is as follows:

- **Sociability:** friendliness, likability; “pertains to cooperation and to forming connections with others” (Brambilla et al., 2011)
- **Morality:** fairness, honesty, trustworthiness; “ being benevolent to people in ways that facilitate correct and principled relations with them by the adherence to ethics and important social values” (Abele et al., 2016)
- **Ability:** capability, intelligence, competence; relating to the capability to achieve goals (separately from the motivation to actively pursue those goals) (Abele et al., 2016)
- **Assertiveness:** ambition, confidence, activeness; related to the motivation to achieve goals (separately from the ability to do so) (Abele et al., 2016)
- **Beliefs:** measured across a continuum from progressive/liberal/modern to conservative/traditional; can encompass political as well as religious beliefs; “conservative-progressive beliefs are informative of mainstream society’s views about a group’s intention to preserve versus change the status quo” (Koch et al., 2016)
- **Status:** related to power, wealth, dominance, and social standing (Koch et al., 2016)

To give a few examples, society might stereotype a CEO as being intelligent (high-Ability), competitive (high-Assertiveness), right-wing (high-Beliefs), wealthy (high-Status) while at the same time uncaring (low-Sociability) and willing to cheat to get ahead (low-Morality). In contrast, an Asian high-schooler might be stereotyped as very smart (high-Ability) and honest (high-Morality), but passive (low-Assertiveness) and shy (low-Sociability). Some dimensions are more salient for certain social groups, as described in Appendix E below.

B POLAR Model

The following method is adapted from the POLAR framework introduced by Mathew et al. (2020).

Suppose we want to transform from the original sentence embedding space $\mathbb{E}$, $|\mathbb{E}| = D$, to the reduced embedding space $\mathbb{E}'$, $|\mathbb{E}'| = D'$, with $D' < D$.

In general, for each dimension $d \in \{1, 2, ..., D'\}$, we define the set of N_{d+} sentences associated with the positive pole of that dimension as $\mathbb{P}_{d+} = \{p^1_{d+}, p^2_{d+}, ..., p^{N_{d+}}_{d+}\}$, and a set of N_{d-} sentences associated with the negative pole of that dimension as $\mathbb{P}_{d-} = \{p^1_{d-}, p^2_{d-}, ..., p^{N_{d-}}_{d-}\}$. We obtain the POLAR directional vector for that dimension as follows:

$$\overrightarrow{dir}_d = \frac{1}{N_{d+}} \sum_{i=1}^{N_{d+}} \mathbb{V}_{p^i_{d+}} - \frac{1}{N_{d-}} \sum_{i=1}^{N_{d-}} \mathbb{V}_{p^i_{d-}} \tag{1}$$

where $\mathbb{V}_s$ represents the vector representation of the sentence s in the embedding space $\mathbb{E}$.

The set of POLAR direction vectors are then stacked to form $dir \in \mathbb{R}^{D' \times D}$, which represents the change of basis matrix for the new reduced-dimensional embedding subspace $\mathbb{E}'$. In the new subspace, a sentence s is represented by $\mathbb{V}'_s$, which is calculated using the following linear transformation:

$$\mathbb{V}'_s = (dir^T)^{-1} \mathbb{V}_s \tag{2}$$

Each dimension in $\mathbb{E}'$ can now be interpreted in terms of the polar opposites used to define $\overrightarrow{dir_1}$, $\overrightarrow{dir_2}$, ... $\overrightarrow{dir_{D'}}$.

Here, we transform from a high-dimensional RoBERTa sentence embedding space ($D = 1024$),

Dimension	Positive	Negative
Sociability	friendly, sociable	unfriendly, antisocial
Morality	trustworthy, honest	untrustworthy, dishonest
Ability	competent, skilled	incompetent, unskilled
Assertiveness	confident, assertive	meek, submissive
Beliefs	conservative, traditional	liberal, modern
Status	high-status, wealthy	low-status, poor

Table A.1: Adjectives used to define the poles of each dimension. Each adjective was embedded in the sentence template `These people are always <ADJ>`.

to a six-dimensional space, interpretable in terms of six psychologically-defined dimensions ($D' = 6$).

To define our six-dimensional model, we use 12 sets of seed words, each set containing two adjectives ($N_{d+} = N_{d-} = 2$ for $d = 1, 2, 3, 4, 5, 6$). The adjectives representing the positive poles of each dimension are taken from Nicolas et al. (2022). They are the same adjectives that the annotators saw when filling out the rating scales. For the set of adjectives defining the negative poles, we use the direct antonyms of the positive adjectives. See Table A.1 for the full set of adjectives used. Since we want a model that operates on the sentence level, each adjective is inserted in the sentence template `These people are always <ADJ>`. The sentences are then represented as sentence vectors using the 1024-dimensional RoBERTa embedding model, and the change of basis matrix is calculated according to the above.

C Validation Experiments

As a preliminary step to confirm that the POLAR model is capturing the expected information and to select the best parameters, we run a series of small experiments. Briefly, we use lexicons available from Nicolas et al. (2021) to create a validation set of words that should be associated with each dimension. These lexicons were created by hand, based on the existing literature in social psychology.

We then experiment with various parameters relating to the dimensionality reduction. Following Fraser et al. (2022), we consider the options:

- No dimensionality reduction
- Principal Components Analysis (Gewers et al., 2021), optimizing the number of dimensions between 10-100
- Partial Least Squares (Rosipal and Krämer, 2005), optimizing the number of dimensions between 10-100

We considered two evaluation criteria: (1) High accuracy (percentage of times a word was correctly associated with either the positive or negative direction of the salient dimension), (2) Low correlation between dimensions (while we expect some correlation between the dimensions, the POLAR model should represent them as separate, distinct concepts). Fortunately, the setting with the highest accuracy also resulted in the lowest correlation, and so in what follows we use the model with Partial Least Squares applied to reduce the embedding size to 30. This led to an average accuracy of 95% on the validation set, and a mean absolute correlation between the dimensions of 0.13.

We did not optimize the choice of word embeddings, as extensive exploration was previously documented by Fraser et al. (2022), and we use the RoBERTa-NLI embeddings[7] that they found to be optimal across multiple functional test cases.

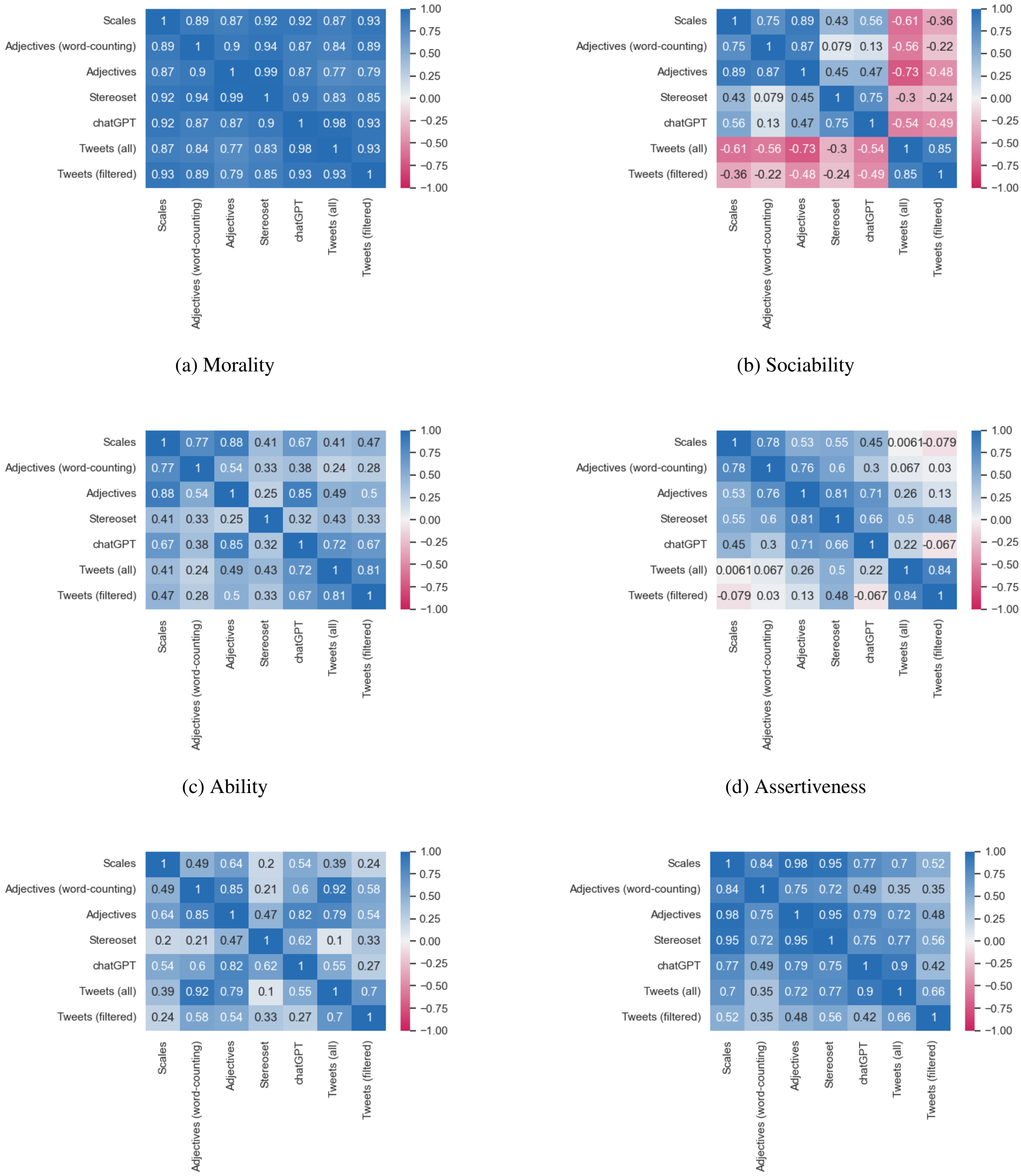
D Correlations between Datasets

Figure C.1 shows the full correlation matrices for each dimension. In general, no unexpected patterns emerge. The two methods of processing the adjectives (our computational method and simple word-counting) tend to be correlated with each other, and the filtered and unfiltered Twitter datasets tend to be correlated with each other. Stereoset and ChatGPT (i.e., human and machine-generated stereotype sentences) are highly correlated ($\rho > 0.5$) for all dimensions except for Ability. The correlations between different datasets are almost always positive, with the notable exception of Sociability estimates based on Twitter, as discussed in the main text.

E Representativeness

In contrast to scale-based measures collected using a forced-choice methodology, when people are

[7] https://huggingface.co/sentence-transformers/nli-roberta-large



(a) Morality

(b) Sociability

(c) Ability

(d) Assertiveness

(e) Beliefs

(f) Status

Figure C.1: Spearman rank correlations between estimates from each of the data sources, for each dimension.

generating spontaneous, free-text responses, they can choose which dimension(s) to focus on for any given group. This choice provides additional information about what stereotype dimensions are seen as being most relevant to each group. Nicolas et al. (2022) defined this as *representativeness*: "the prevalence of a stereotype dimension in perceivers' spontaneous beliefs about a social group." Here, we operationalize this as the proportion of text samples that are assigned an absolute value greater than 0.5 along a given dimension.[8] In the main text, we computed this proportion over all groups, and called it *prevalence*, with the goal of understanding more generally how many text samples make strong statements about the different dimensions. Here, we calculate the proportion per group, and thus call it *representativeness*, as it now captures the information about how representative, or important, any given dimension is perceived as being when describing each target group.

The values are shown in Figure D.1. Briefly, we observe that over 50% of the data in the adjectives dataset, Stereoset, and ChatGPT make statements about politicians' morality. This suggests that when people think about stereotypes of politicians, one of the first things they think about is their (im)morality. From a computational perspective, it also means our estimates of that dimension are based on a much larger dataset than our estimates for the other dimensions.

In contrast, for teachers, we see a more even distribution across the different dimensions. Still, dimensions like Assertiveness and Beliefs are more sparsely represented. CEOs have Morality and Ability as the most representative dimensions, with Status also mentioned 10-20% of the time. Scientists, accountants, engineers, farmers, and nurses all have Ability as the most representative dimension. For nurses, Sociability traits are also mentioned more often than for other groups.

Figure D.1 also shows that some data sources are more extreme in their representativeness values. In particular, the adjectives, Stereoset, and ChatGPT (all of which were collected by explicitly asking for stereotype information) have more extreme values, while the Twitter data is more uniformly distributed across dimensions. This reflects the more general nature of the Twitter data.

F Topic-Modelling Results

BERTopic is available to install at `https://maartengr.github.io/BERTopic/index.html`. We used v0.13.0. For simplicity, we used the default parameters as much as possible.

We use the RoBERTa-NLI pre-trained embedding model, as mentioned in Appendix C. For the vectorizer model, we used the scikit-learn CountVectorizer method, removing English stopwords and ignoring terms that appear in less than 1% of the sentences (`min_df = 0.01`). To ensure reproducibility, we set `random_state = 42` in the UMAP model. For the HDBSCAN clustering algorithm, we specified the `min_samples = 1`, to promote less-conservative clustering.[9] Since we don't know *a priori* how many topics to expect for each group, we set `nr_topics = 'auto'`. For all the other parameters, the default settings of the BERTopic package were used.

G Data Licensing for Existing Datasets

The data associated with Nicolas et al. (2022) is freely available on the Open Science Framework: `https://osf.io/74rax/`. The OSF Terms of Use permit public data to be used for a wide range of non-commercial and commercial uses.

The StereoSet data is available here: `https://huggingface.co/datasets/stereoset` with License CC-BY-SA 4.0.

The Nicolas et al. data was collected with the intention of studying stereotypes. The StereoSet dataset was collected for the purpose of measuring stereotypical biases in language models. We believe our present research is in line with these purposes.

H ChatGPT Dataset

The CSV file containing the pre-processed text is available by contacting the authors.

I Twitter Dataset

The Twitter data was collected in November 2022, under an approved Academic Project on the Twitter developer portal. This was prior to the removal of the Research API and the introduction of a paywall in April 2023. Unfortunately, due to Twitter Terms of Service, we cannot redistribute the Twitter dataset.

[8] The threshold of 0.5 was chosen based on the validation set data, where it was observed that a score of 0.5 roughly differentiated the words associated with each dimension from words associated with other dimensions.

[9] `https://hdbscan.readthedocs.io/en/latest/parameter_selection.html`

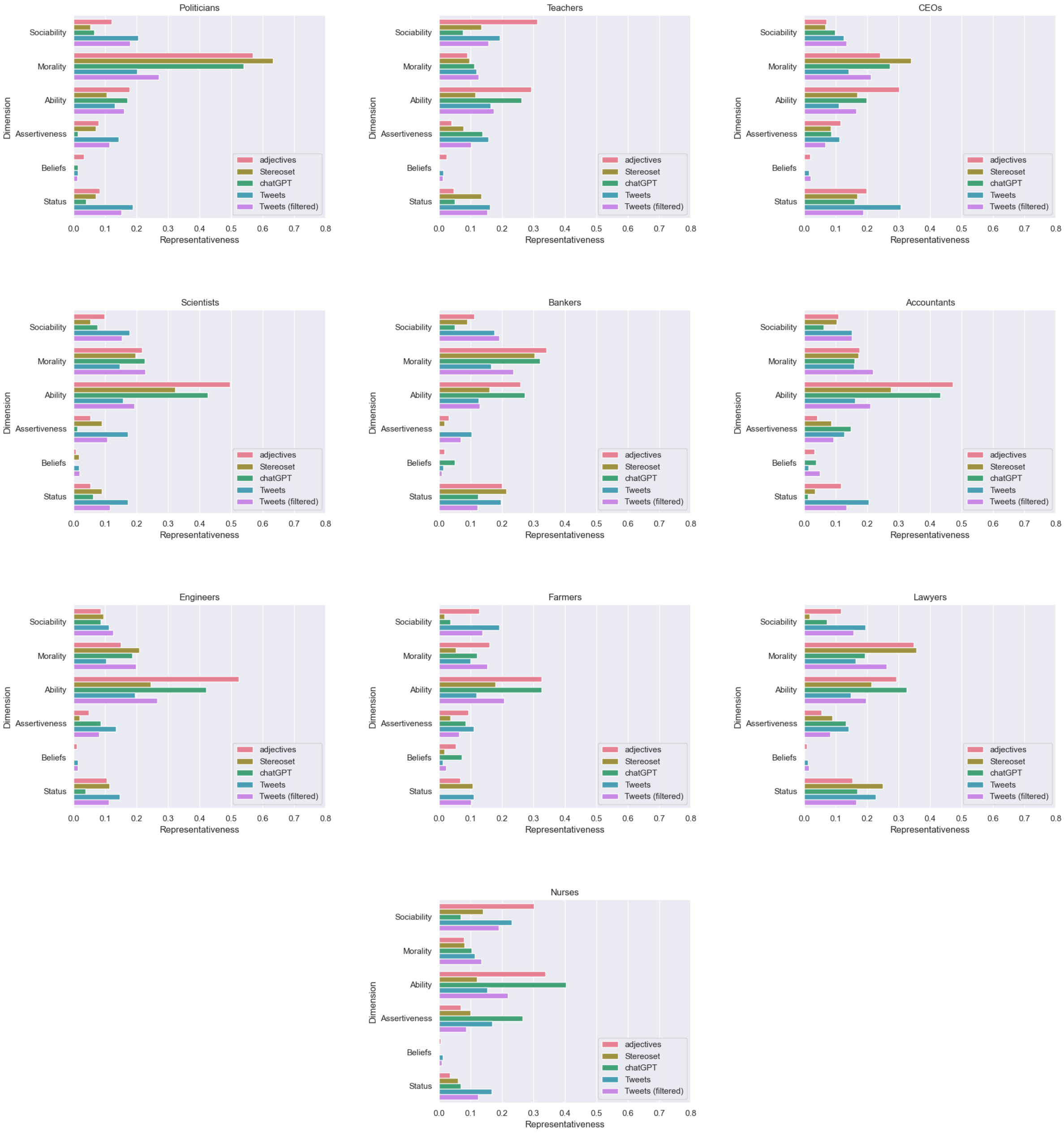


Figure D.1: The proportion of text instances assigned a value greater than 0.5, for each group, dimension, and data source.