Understanding and Countering Stereotypes: A Computational Approach to the Stereotype Content Model

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko
National Research Council Canada
Ottawa, Canada
{Kathleen.Fraser,Isar.Nejadgholi,Svetlana.Kiritchenko}@nrc-cnrc.gc.ca

Abstract

Stereotypical language expresses widely-held beliefs about different social categories. Many stereotypes are overtly negative, while others may appear positive on the surface, but still lead to negative consequences. In this work, we present a computational approach to interpreting stereotypes in text through the Stereotype Content Model (SCM), a comprehensive causal theory from social psychology. The SCM proposes that stereotypes can be understood along two primary dimensions: warmth and competence. We present a method for defining warmth and competence axes in semantic embedding space, and show that the four quadrants defined by this subspace accurately represent the warmth and competence concepts, according to annotated lexicons. We then apply our computational SCM model to textual stereotype data and show that it compares favourably with survey-based studies in the psychological literature. Furthermore, we explore various strategies to counter stereotypical beliefs with anti-stereotypes. It is known that countering stereotypes with anti-stereotypical examples is one of the most effective ways to reduce biased thinking, yet the problem of generating anti-stereotypes has not been previously studied. Thus, a better understanding of how to generate realistic and effective anti-stereotypes can contribute to addressing pressing societal concerns of stereotyping, prejudice, and discrimination.

1 Introduction

Stereotypes are widely-held beliefs about traits or characteristics of groups of people. While we tend to think of stereotypes as expressing negative views of groups, some stereotypes actually express positive views (e.g. *all women are nurturing*). However, even so-called 'positive' stereotypes can be harmful, as they dictate particular roles that individuals are expected to fulfill, regardless of whether they have the ability or desire to do so (Kay et al., 2013).

The existence of stereotypes in our society – including in entertainment, the workplace, public discourse, and even legal policy – can lead to a number of harms. Timmer (2011) organizes these harms into three main categories: (1) Misrecognition effects: harms caused by denying members of particular groups an equal place in society, diminishing their human dignity, or other forms of marginalization. (2) Distribution effects: harms resulting from unfair allocation of resources, either by increasing the burden placed on a group, or decreasing a group's access to a benefit. (3) Psychological effects: the distress and unhappiness caused by an awareness and internalization of the stereotyped biases against one's identity group. Additionally, the internalization of these negative stereotypes can lead to anxiety and underachievement. To reduce these harms and promote a more egalitarian society, we must identify and counter stereotypical language when it occurs.

Evidence from the psychological literature suggests that one of the most effective methods for reducing stereotypical thinking is through exposure to counter-stereotypes, or anti-stereotypes. Finnegan et al. (2015) showed participants stereotypical and anti-stereotypical images of highly socially-gendered professions (e.g., a surgeon is stereotypically male, and a nurse is stereotypically female; the genders were reversed in the anti-stereotypical images), and then measured their gender bias in a judgement task. Exposure to anti-stereotypical images significantly reduced gender bias on the task. Blair et al. (2001) used a mental imagery task and reported that participants in the anti-stereotypical condition subsequently showed significantly weaker effects on the Implicit Association Test (IAT). Dasgupta and Greenwald (2001) showed a similar effect by exposing participants to

anti-stereotypical exemplars (e.g. admired Black celebrities, and disliked white individuals). When Lai et al. (2014) compared 17 interventions aimed at reducing stereotypical thinking, methods involving anti-stereotypes were most successful overall.

Thus, creating technology that enables users to identify stereotypical language when it occurs, and then counter it with anti-stereotypes, could help to reduce biased thinking. However, the idea of what constitutes an anti-stereotype remains ill-defined. Is an anti-stereotype simply the semantic opposite of a stereotype? Or can anything that is not a stereotype serve as an anti-stereotype? If two groups are stereotyped similarly, do they have an identical anti-stereotype? Can an anti-stereotype actually reflect an equally harmful view of a target group (e.g. *the cold-hearted career woman* as an anti-stereotype to *the nurturing housewife)*?

Here, we begin to untangle some of these questions using the StereoSet dataset (Nadeem et al., 2020). We begin by analyzing the stereotypes expressed in this dataset. One widely-accepted model of stereotypes, prejudice, and inter-group relationships from social psychology is the "Stereotype Content Model" or SCM (Fiske et al., 2002). The SCM proposes two fundamental and universal dimensions of social stereotypes: *warmth* and *competence*. By defining the warm–cold, competent–incompetent axes in the semantic embedding space, we are able to cluster and interpret stereotypes with respect to those axes. We can then examine the associated anti-stereotypes and their relation to both the stereotyped description and the target group. Thus, our contributions are as follows:

- To develop a computational method for automatically mapping textual information to the warmth–competence plane as proposed in the Stereotype Content Model.
- To validate the computational method and optimize the choice of word embedding model using a lexicon of words known to be associated with positive and negative warmth and competence.
- To compare the stereotypes in StereoSet with those reported in the survey-based social psychology literature.
- To analyze human-generated anti-stereotypes as a first step towards automatically generating anti-stereotypes, as a method of countering stereotypes in text with constructive, alternative perspectives.

2 Related Work

We provide more details on the Stereotype Content Model and its practical implications, and then briefly review the NLP research on computational analysis of stereotypical and abusive content.

Stereotype Content Model: Stereotypes, and the related concepts of prejudice and discrimination, have been extensively studied by psychologists for over a century (Dovidio et al., 2010). Conceptual frameworks have emerged which emphasize two principle dimensions of social cognition. The Stereotype Content Model (SCM) refers to these two dimensions as *warmth* (encompassing sociability and morality) and *competence* (encompassing ability and agency) (Fiske et al., 2002). When forming a cognitive representation of a social group to anticipate probable behaviors and traits, people are predominantly concerned with the others' intent—are they friends or foes? This intent is captured in the primary dimension of warmth. The competence dimension determines if the others are capable to enact that intent. A key finding of the SCM has been that, in contrast to previous views of prejudice as a uniformly negative attitude towards a group, many stereotypes are actually *ambivalent*; that is, they are high on one dimension and low on the other.

Further, the SCM proposes a comprehensive causal theory, linking stereotypes with social structure, emotions, and discrimination (Fiske, 2015). According to this theory, stereotypes are affected by a perceived social structure of *interdependence* (cooperation versus competition), corresponding to the warmth dimension, and *status* (prestige and power), determining competence. Stereotypes then predict emotional response or prejudices. For example, groups perceived as unfriendly and incompetent (e.g., homeless people, drug addicts) evoke disgust and contempt, groups allegedly high in warmth but low in competence (e.g., older people, people with disabilities) evoke pity, and groups perceived as cold and capable (e.g., rich people, businesspeople) elicit envy.

Finally, the emotions regulate the actions (active or passive help or harm). Thus, low warmth–low competence groups often elicit active harm and passive neglect, whereas low warmth–high competence groups may include envied out-groups who are subjects of passive help in peace times but can become targets of attack during social unrest (Cuddy et al., 2007).

The SCM has been supported by extensive quantitative and qualitative analyses across cultures and time (Fiske, 2015; Fiske and Durante, 2016). To our knowledge, the current work presents the first computational model of the SCM.

Stereotypes in Language Models: An active line of NLP research is dedicated to quantifying and mitigating stereotypical biases in language models. Early works focused on gender and racial bias and revealed stereotypical associations and common prejudices present in word embeddings through association tests (Bolukbasi et al., 2016; Caliskan et al., 2017; Manzini et al., 2019). To discover stereotypical associations in contextualized word embeddings, May et al. (2019) and Kurita et al. (2019) used pre-defined sentence templates. Similarly, Bartl et al. (2020) built a template-based corpus to quantify bias in neural language models, whereas Nadeem et al. (2020) and Nangia et al. (2020) used crowd-sourced stereotypical and anti-stereotypical sentences for the same purpose. In contrast to these studies, while we do use word embeddings to represent our data, we aim to identify and categorize stereotypical views expressed in text, not in word embeddings or language models.

Abusive Content Detection: Stereotyping, explicitly or implicitly expressed in communication, can have a detrimental effect on its target, and can be considered a form of abusive behavior. Online abuse, including hate speech, cyber-bullying, online harassment, and other types of offensive and toxic behaviors, has been a focus of substantial research effort in the NLP community in the past decade (e.g. see surveys by Schmidt and Wiegand (2017); Fortuna and Nunes (2018); Vidgen et al. (2019)). Most of the successes in identifying abusive content have been reported on text containing explicitly obscene expressions; only recently has work started on identifying more subtly expressed abuse, such as stereotyping and micro-aggressions (Breitfeller et al., 2019). For example, Fersini et al. (2018) and Chiril et al. (2020) examined gender-related stereotypes as a sub-category of sexist language, and Price et al. (2020) annotated 'unfair generalizations' as one attribute of unhealthy online conversation. Sap et al. (2020) employed large-scale language models in an attempt to automatically reconstruct stereotypes implicitly expressed in abusive social media posts. Their work showed that while the current models can accurately predict whether the online post is offensive or not, they struggle to effectively reproduce human-written statements for implied meaning.

Counter-narrative: Counter-narrative (or counterspeech) has been shown to be effective in confronting online abuse (Benesch et al., 2016). Counter-narrative is a non-aggressive response to abusive content that aims to deconstruct and delegitimize the harmful beliefs and misinformation with thoughtful reasoning and fact-bound arguments. Several datasets of counter narratives, spontaneously written by regular users or carefully crafted by experts, have been collected and analyzed to discover common intervention strategies (Mathew et al., 2018; Chung et al., 2019). Preliminary experiments in automatic generation of counter-narrative demonstrated the inadequacy of current large-scale language models for generating effective responses and the need for a human-in-the-loop approach (Qian et al., 2019; Tekiroğlu et al., 2020). Countering stereotypes through exposure to anti-stereotypical exemplars is based on a similar idea of deconstructing harmful beliefs with counter-facts.

3 Data and Methods

We develop our computational SCM using labelled data from Nicolas et al. (2020) and the POLAR framework for interpretable word embeddings (Mathew et al., 2020), and then apply it to stereotype and anti-stereotype data from StereoSet (Nadeem et al., 2020). Details are provided in the following sections.

3.1 Warmth-Competence Lexicons

To construct and validate our model, we make use of the supplementary data from Nicolas et al. (2020) (`https://osf.io/yx45f/`). They provide a list of English seed words, captured from the psychological literature, associated with the warmth and competence dimensions; specifically, associated with sociability and morality (warmth), and ability and agency (competence). They then use WordNet to generate an extended lexicon of English words either positively or negatively associated with aspects of warmth and competence. Some examples from the seed data and extended lexicon are given in Table 1.

3.2 StereoSet

For human-generated stereotype and anti-stereotype data, we use the publicly-available

Dimension	Component	Sign	Seed word examples	n_{seed}	Extended lexicon examples	n_{extended}
Warmth	Sociability	pos	friendly, warm, pleasant	34	amusing, brother, fun	482
		neg	cold, repellent, disliked	32	detached, grim, surly	423
	Morality	pos	trustworthy, sincere, honest	40	donor, justice, modest	460
		neg	dishonest, selfish, unfair	49	cheat, dreadful, henchman	1750
Competence	Agency	pos	confident, assertive, secure	35	bravery, decisive, stubborn	444
		neg	fearful, lazy, inactive	31	follow, minion, quitter	265
	Ability	pos	smart, intelligent, able	33	analytic, fluency, thorough	579
		neg	stupid, ignorant, incapable	29	forgetful, silly, unfit	301

Table 1: Examples of words from the training data (seed words) and validation data (extended lexicon), for each of the components comprising the warmth and competence dimensions.

portion of the StereoSet dataset (Nadeem et al., 2020). This English-language dataset was constructed to test language model bias, and part of the data is kept hidden as the test set for a leaderboard on language model fairness (https://stereoset.mit.edu/). Instead, we use the development set, which contains stereotype data for 79 target groups across four broad demographic domains: gender, race or nationality, profession, and religion.

In StereoSet, there are two experimental conditions: intra-sentence and inter-sentence. Here, we focus on the intra-sentence data only. The data was collected from crowd-workers as follows (see Nadeem et al. (2020) for more detail): Given a target group label, the annotator is asked to generate a stereotypical word associated with that group, as well as an anti-stereotypical word and an unrelated word. They then construct a context sentence containing the target group label, and a blank which can be filled with the stereotypical or anti-stereotypical word. For example, if the target group was women, the annotator might come up with *emotional* and *rational* as the stereotype and anti-stereotype words respectively, and then construct a sentence like *Women are known for being overly* ⟨*BLANK*⟩. For our current analysis, we consider only the stereotype and anti-stereotype words, and discard the context sentence. We also exclude any targets that do not directly refer to groups of people (e.g., we discard *Norway* but keep *Norwegian*). This results in 58 target groups with an average of 25 stereotype and anti-stereotype word pairs each.

3.3 Constructing Warmth and Competence Dimensions

We consider several possible representations for the words in our dataset, including GloVe (Pennington et al., 2014), word2vec (Mikolov et al., 2013), and FastText (Mikolov et al., 2018).[1] In all cases, the key question is how to project the higher-dimensional word embedding onto the warmth–competence plane.

Rather than using an unsupervised approach such as PCA, we choose the POLAR framework introduced by Mathew et al. (2020). This framework seeks to improve the interpretability of word embeddings by leveraging the concept of 'semantic differentials,' a psychological rating scale which contrasts bipolar adjectives, e.g. *hot–cold*, or *good–bad*. Given word embeddings that define these polar opposites for a set of concepts, all other word embeddings in the space are projected onto the 'polar embedding space,' where each dimension is clearly associated with a concept.

For our purposes, the polar opposites are warmth–coldness and competence–incompetence, as defined by the sets of seed words from Nicolas et al. (2020). To reduce the dimensionality of the space to 2D, we average the word vectors for all seed words associated with each dimension and polarity. That is, to define the warmth direction, we take the mean of all words in the seed dictionary which are positively associated with warmth. Given vector definitions for warmth, coldness, competence, and incompetence, we can then use a simple matrix transformation to project any word embedding to the 2D subspace defined by these basis vectors (mathematical details are given in Appendix A).

We could also to polarity at this dimension level then we do not need antonym pairs

4 Model Validation

We first evaluate the model's ability to accurately place individual words from the lexicons along the

[1] We consider here only noncontextual word embeddings, in line with Mathew et al. (2020). Because the POLAR framework is based on linear algebraic computations, it is not immediately obvious whether it will extend directly to contextualized embeddings, which are notably anisotropic (Ethayarajh, 2019).

Embedding model	Warmth	Comp.
FastText-crawl-subword-300	**85.0**	**85.8**
FastText-wiki-news-subword-300	84.9	84.8
Word2vec-GoogleNews-300	80.2	72.6
GloVe-twitter-200	72.8	74.2
GloVe-wiki-gigaword-300	78.7	77.9

Table 2: Accuracy of the word embedding models on predicting the correct labels for the extended lexicon.

warmth and competence dimensions. We then explore whether we can reproduce findings describing where certain target groups are typically located in the warmth–competence plane, based on the previous survey-based social psychology literature.

4.1 Comparison with Existing Lexicons

As described above, we use the extended lexicon from Nicolas et al. (2020) to validate our model. We remove any words in the lexicon which appear in the seed dictionary and any words which do not have representations in all the pretrained embedding models, leaving a total of 3,159 words for validation.

In the extended lexicon, the words are annotated with either +1 or -1 to indicate a positive or negative association with the given dimension. We pass the same words through our system, and observe whether the model labels the word as being positively or negatively associated with the relevant dimension. Our evaluation metric is accuracy; i.e. the proportion of times our system agrees with the lexicon. Note that all words are associated with *either* warmth or competence, and therefore we can only evaluate one dimension at a time.

We evaluate a number of pre-trained word embeddings in the gensim library (Řehůřek and Sojka, 2010), with the results given in Table 2. The FastText embeddings generally outperform the other embeddings on this task, with the 2M word model trained on 600B tokens in the Common Crawl leading to the highest accuracy. Therefore, we use this embedding model in the analysis that follows.

4.2 Comparison with Psychological Surveys

We now address the question of whether our model, in conjunction with the StereoSet data, is able to reproduce findings from psychological surveys. We project stereotypes from the StereoSet data onto the warmth–competence space for the 24 target groups that meet both of the following criteria: (1) they are included in the publicly available portion of the StereoSet data, and (2) they have been previously studied for stereotyping in the psychological literature. Based on the findings from psychological surveys, we expect these target groups will be mapped to the following quadrants:[2]

- **Warm-Competent:** nurse, psychologist ('healthcare professions') (Brambilla et al., 2010), researcher ('professor') (Eckes, 2002).
- **Warm-Incompetent:** grandfather ('elderly'), mommy, mother ('traditional women') (Cuddy et al., 2008), schoolboy, schoolgirl ('children') (Fiske, 2018).
- **Cold-Competent:** male, gentleman ('man') (Glick et al., 2004), Japanese (Lee and Fiske, 2006), commander (Cuddy et al., 2011), manager, entrepreneur (Fiske, 2010), mathematician, physicist, chemist, engineer ('scientist') (Losh et al., 2008), software developer ('technical expert') (Fiske, 2018).
- **Cold-Incompetent:** African, Ethiopian, Ghanian, Eritrean, Hispanic (Lee and Fiske, 2006), Arab (Fiske et al., 2006).

To locate each target group on the plane, we generate word embeddings for each of the stereotype words associated with the target group, find the mean, and project the mean to the polar embedding space. As we aim to identify commonly-held stereotypes, we use a simple cosine distance filter to remove outliers, heuristically defined here as any words which are greater than a distance of 0.6 from the mean of the set of words. We also remove words which directly reference a demographic group (e.g., black, white) as these words are vulnerable to racial bias in the embedding model and complicate the interpretation. A complete list of the words in each stereotype cluster can be found in the Appendix B.

Figure 1 confirms many of the findings predicted by the literature. Most (67%) of the stereotypes lie in the predicted quadrant, including *grandfather* and *schoolgirl* in the paternalistic warm–incompetent quadrant; *nurse* and *psychologist* in the admired warm–competent quadrant, *manager* and *male* in the envied cold–competent quadrant, and *African* and *Hispanic* in the cold–cold quadrant.

Other stereotypes lie in locations which seem

[2]Note that these research findings simply report stereotypical beliefs which are prevalent in North American society; we in no way aim to perpetuate, confirm, or promote these views.

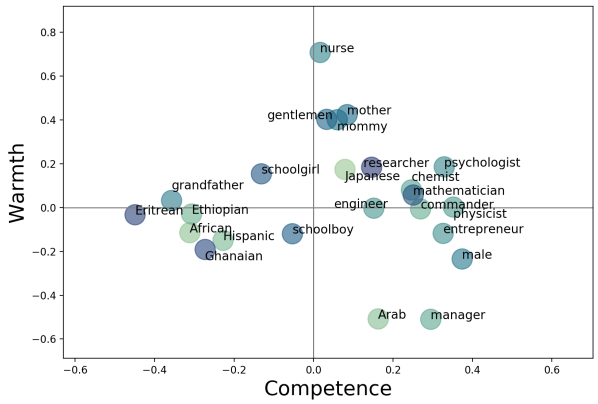

Figure 1: Validating known stereotypes.

reasonable on examination of the underlying data. For example, while men are typically stereotyped as being competent yet cold in the psychological literature, the specific keyword *gentlemen* evokes a certain subset of men (described with words such as *polite*, *respectful*, and *considerate*), which ranks higher on the warmth dimension than the target word *male* (example words: *dominant*, *aggressive*).

We also observe that while children have generally been labelled as warm–incompetent in previous work (Fiske, 2018), this dataset distinguishes between male and female schoolchildren, and, as expected based on studies of gender, schoolboys are ranked as lower warmth than schoolgirls. The words used to describe schoolboys include references to the 'naughty' schoolboy stereotype, while the words describing schoolgirls focus on their innocence and naivety.

It is also notable that *Arab*, predicted to lie in the cold–incompetent quadrant, is here mapped to the cold–competent quadrant instead. We hypothesize that this is due to the use of stereotype words like *dangerous* and *violent*, which suggest a certain degree of agency and the ability to carry out goals. In contrast, the target group *African* as well as those associated with African countries are stereotyped as *poor* and *uneducated*, and thus low on the competence dimension.

In general, we conclude that in most cases the computational approach is successful in mapping stereotyped groups onto the predicted areas of the warmth–competence plane, and that the cases which diverge from findings in the previous literature do appear to be reasonable, based on an examination of the text data. Having validated the model, we can now apply it to the rest of the stereotype data in StereoSet, as well as the anti-stereotypes.

5 Stereotypes and Anti-Stereotypes

The SCM presents a concise theory to explain stereotypes and resulting prejudiced behaviour; however, it does not generate any predictions about anti-stereotypes. Here, we explore the anti-stereotypes in StereoSet within the context of the SCM, first at the level of individual annotators, and then at the level of target groups (combining words from multiple annotators). We then discuss how we might use information about warmth and competence to generate anti-stereotypes with the specific goal of reducing biased thinking.

5.1 Anti-Stereotypes in StereoSet

In this section, we investigate the question: What do human annotators come up with when asked to produce an anti-stereotype? One possibility is that they simply produce the antonym of their stereotype word. To test this hypothesis, for all 58 groups and each pair of stereotype and anti-stereotype words, we obtain a list of antonyms for the stereotype word using the Python library PyDictionary. We additionally search all the synonyms for the stereotype word, and add all of their antonyms to the list of antonyms as well. Then, if the lemma of the anti-stereotype matches the lemma of any of the retrieved antonyms, we consider it a match.

However, as seen in Table 3, the strategy of simply producing a direct antonym is only used 23% of the time. We consider four other broad possibilities: (1) that the annotator generates an anti-stereotype word that lies in the opposite quadrant from the stereotype word, e.g., if the stereotype word is low-competence, low-warmth (LC-LW), then the anti-stereotype word should be high-competence, high-warmth (HC-HW); (2) that the annotator chooses a word with the opposite warmth polarity (i.e. flips warmth), while keeping the competence polarity the same; (3) that the annotator chooses a word with the opposite competence polarity (i.e. flips competence), while keeping the warmth polarity the same; (4) that the annotator chooses a word that lies in the same quadrant as the stereotype word. We report the proportion of times that each strategy is observed; first overall, then for each quadrant individually. The choice of whether to modify warmth or competence might also depend on which of those dimensions is most salient for a given word, and so we consider separately words for which the absolute value of competence is greater than the absolute value of warmth, and

Strategy	Overall $n = 895$	HC-HW $n = 192$	LC-HW $n = 183$	LC-LW $n = 176$	HC-LW $n = 344$	$\|C\| > \|W\|$ $n = 428$	$\|W\| > \|C\|$ $n = 467$
Direct antonym	23.4	26.0	**32.6**	27.8	15.0	27.2	19.2
Opposite quadrant	**29.6**	**30.2**	15.5	26.1	**38.3**	**28.1**	**31.2**
Flip warmth	20.6	14.6	26.5	**29.5**	16.4	12.3	29.8
Flip competence	16.7	24.0	12.7	13.1	16.7	22.8	10.1
Same quadrant	9.6	5.2	12.7	3.4	13.5	9.6	9.6

Table 3: The percentage of times each of the hypothesized strategies of anti-stereotype generation is used for stereotypes, overall and in each quadrant. Quadrants are labelled as HC-HW, LC-HW, LC-LW, and HC-LW, where HC/LC denotes high/low competence, and HW/LW denotes high/low warmth. We also consider separately those stereotypes which have competence as the most salient dimension ($|C| > |W|$), and those which have warmth as the most salient dimension ($|W| > |C|$).

vice versa. The results are given in Table 3.

While no single strategy dominates, we can make a few observations. In general, it is more likely that people select an anti-stereotype which is not a direct antonym, but which lies in the opposite quadrant in the warmth-competence plane. Flipping only one axis is less frequent, although we see in the last two columns that it is more likely that the competence will be flipped when competence is the salient dimension for a word, and similarly for warmth. Finally, choosing another word in the same quadrant is rare, but more common in the ambivalent quadrants.

While it is not possible to know what thought process the annotators followed to produce anti-stereotypes, we consider the following possible explanation. Just as we have here conceptualized a stereotype as being defined not by a single word, but by a set of words, perhaps each annotator also mentally represents each stereotype as a set of words or ideas. Then, the anti-stereotype word they produce sometimes reflects a different component of their mental image than the initial stereotype word. To give a concrete example from the data, one annotator stereotypes Hispanic people as *aggressive*, but then comes up with *hygienic* as an anti-stereotype, suggesting that *unhygienic* is also part of their multi-dimensional stereotype concept. The choice of whether to select a direct antonym, or whether to negate some other component of the stereotype, may depend on the availability of a familiar lexical antonym, the context sentence, or any number of other factors. In short, it appears that the process by which human annotators generate pairs of stereotype and anti-stereotype words is complex and not easily predicted by the SCM.

We then examine how these pairs of stereotype and anti-stereotype words combine to produce an overall anti-stereotype for the target group in question. Taking the same approach as in the previous section, we average the anti-stereotype word vectors to determine the location of the anti-stereotype in the warmth–competence plane. For each target group, we then select the word closest to the mean for both the stereotype and anti-stereotype clusters. Similarly to when we look at individual word pairs, in 22% of cases, the mean of the anti-stereotype is the direct antonym of the stereotype mean. In the other cases, 45% of the anti-stereotype means lie in the opposite quadrant to the stereotypes, in 16% of cases the warmth polarity is flipped, in 10% of cases the competence polarity is flipped, and in only 7% cases (4 target groups), the anti-stereotype lies in the same quadrant as the stereotype.

Target	Stereotype	Antonym	Anti-stereotype
African	poor	rich	rich
Hispanic	poor	rich	hardworking
mother	caring	uncaring	hateful
nurse	caring	uncaring	rude
commander	strong	weak	stupid
mover	strong	weak	weak
football player	dumb	smart	weak

Table 4: Examples comparing stereotypes with their direct antonym and the anti-stereotype from StereoSet.

In Table 4, we offer a few examples of cases where the anti-stereotype means agree and disagree with the direct antonyms of the stereotypes. As in the pairwise analysis, in many cases the anti-stereotypes appear to be emphasizing a supposed characteristic of the target group which is not captured by the stereotype mean; for example, the anti-stereotype for 'dumb football player' is not *smart*, but *weak* – demonstrating that *strength* is also part of the football player stereotype. This is also seen clearly in the fact that two target groups with the same stereotype mean are not always assigned the same anti-stereotype: for example, both Africans and Hispanics are stereotyped as *poor*, but Africans are assigned the straightforward anti-stereotype *rich*, while Hispanics are assigned *hard-*

working (perhaps implying that their poverty is due to laziness rather than circumstance).

The general conclusion from these experiments is that stereotypes are indeed multi-dimensional, and the anti-stereotypes must be, also. Hence it is not enough to generate an anti-stereotype simply by taking the antonym of the most representative word, nor is it sufficient to identify the most salient dimension of the stereotype and only adjust that. When generating anti-stereotypes, annotators (individually, in the pairwise comparison, and on average) tend to invert both the warmth and competence dimensions, taking into account multiple stereotypical characteristics of the target group.

5.2 Anti-Stereotypes for Social Good

The anti-stereotypes in StereoSet were generated with the goal of evaluating language model bias. Ultimately, our goal is quite different: to reduce biased thinking in humans. In particular, we want to generate anti-stereotypes that emphasize the positive aspects of the target groups.

As underscored by Cuddy et al. (2008), many stereotypes are ambivalent: they take the form 'X but Y'. Women are *nurturing but weak*, scientists are *intelligent but anti-social*. When we simply take the antonym of the mean, we focus on the single most-representative word; i.e., the X. However, in the examples we can observe that it's actually what comes after the "but ..." that is the problem. Therefore, in generating anti-stereotypes for these ambivalent stereotypes, we hypothesize that a better approach is not to take the antonym of the primary stereotype (i.e., women are *uncaring*, scientists are *stupid*), but rather to challenge the secondary stereotype (women can be nurturing *and strong*, scientists can be intelligent *and social*).

As a first step towards generating anti-stereotypes for such ambivalent stereotypes, we propose the following approach: first identify the most positive aspect of the stereotype (e.g., if the stereotype mean lies in the incompetent–warm quadrant, the word expressing the highest warmth), then identify the most negative aspect of the stereotype in the other dimension (in this example, the word expressing the lowest competence). Then the stereotype can be phrased in the X but Y construction, where X is the positive aspect and Y is the negative aspect.[3] To generate a positive anti-stereotype which challenges stereotypical thinking while not promoting a negative view of the target group, take the antonym only of the negative aspect. Some examples are given in Table 5. A formal evaluation of these anti-stereotypes would involve carrying out a controlled psychological study in which the anti-stereotypes were embedded in an implicit bias task to see which formulations are most effective at reducing bias; for now, we simply present them as a possible way forward.

As shown in the table, taking into account the ambivalent aspects of stereotypes can result in more realistic anti-stereotypes than either taking the mean of the crowd-sourced anti-stereotypes, or simply generating the semantic opposite of the stereotype. For example, the group *grandfather* is mostly stereotyped as *old*, and then counter-intuitively anti-stereotyped as *young*. It is more useful in terms of countering ageism to combat the underlying stereotype that grandfathers are *feeble* rather than denying that they are often old. Similarly, it does not seem helpful to oppose biased thinking by insisting that entrepreneurs can be *lazy*, engineers and developers can be *dumb*, and mothers can be *uncaring*. Rather, by countering only the negative dimension of ambivalent stereotypes, we can create realistic and positive anti-stereotypes.

6 Discussion and Future Work

Despite their prevalence, stereotypes can be hard to recognize and understand. We tend to think about other people on a group level rather than on an individual level because social categorization, although harmful, simplifies the world for us and leads to cognitive ease. However, psychologists have shown that we can overcome such ways of thinking with exposure to information that contradicts those biases. In this exploratory study, we present a computational implementation of the Stereotype Content Model to better understand and counter stereotypes in text.

A computational SCM-based framework can be a promising tool for large-scale analysis of stereotypes, by mapping a disparate set of stereotypes to the 2D semantic space of warmth and competence. We described here our first steps towards developing and validating this framework, on a highly constrained dataset: in StereoSet, the annotators were explicitly instructed to produce stereotypical ideas, the target groups and stereotypical words

[3] A similar method can be used for warm–competent and cold–incompetent stereotypes, although if all words are positive, an anti-stereotype may not be needed, and if all words are negative, then an antonym may be more appropriate.

Target	Stereotype	Anti-stereotype	*X but Y* construction	*X and* $\neg Y$ anti-stereotype
Grandfather	old	young	kind but feeble	kind and strong
Entrepreneur	savvy	lazy	inventive but ruthless	inventive and compassionate
Engineer	smart	dumb	intelligent but egotistical	intelligent and altruistic
Mommy	loving	uncaring	caring but childish	caring and mature
Software developer	nerdy	dumb	intelligent but unhealthy	intelligent and healthy

Table 5: Examples of positive anti-stereotypes created by identifying positive and negative words along each of the dimensions, and taking the antonym only of the negative words.

are clearly specified, and every stereotype has an associated anti-stereotype generated by the same annotator. In future work, this method should be further assessed by using different datasets and scenarios. For example, it may be possible to collect stereotypical descriptions of target groups 'in the wild' by searching large corpora from social media or other sources. We plan to extend this framework to analyze stereotypes on the sentence-level and consider the larger context of the conversations. Working with real social media texts will introduce a number of challenges, but will offer the possibility of exploring a wider range of marginalized groups and cultural viewpoints.

Related to this, we reiterate that only a portion of the StereoSet dataset is publicly available. Therefore, the data does not include the full set of common stereotypical beliefs for social groups frequently targeted by stereotyping. In fact, some of the most affected communities (e.g., North American Indigenous people, LGBTQ+ community, people with disabilities, etc.) are completely missing from the dataset. In this work, we use this dataset only for illustration purposes and preliminary evaluation of the proposed methodology. Future work should examine data from a wide variety of subpopulations differing in language, ethnicity, cultural background, geographical location, and other characteristics.

From a technical perspective, with larger datasets it will be possible to implement a cluster analysis *within* each target group to reveal the different ways in which a given group can be stereotyped. A classification model may additionally improve the accuracy of the warmth–competence categorization, although we have chosen the POLAR framework here for its interpretability and ease of visualization.

We also examined how we might leverage the developed computational model to challenge stereotypical thinking. Our analysis did not reveal a simple, intuitive explanation for the anti-stereotypes produced by the annotators, suggesting they exploited additional information beyond what was stated in the stereotype word. This extra information may not be captured in a single pair of stereotype–anti-stereotype words, but by considering *sets* of words, we can better characterize stereotypes as multi-dimensional and often ambivalent concepts, consistent with the established view in psychology. This also allows us to suggest anti-stereotypes which maintain positive beliefs about a group, while challenging negative beliefs.

We propose that this methodology may potentially contribute to technology that assists human professionals, such as psychologists, educators, human rights activists, etc., in identifying, tracking, analyzing, and countering stereotypes at large scale in various communication channels. There are a number of ways in which counter-stereotypes can be introduced to users (e.g., through mentions of counter-stereotypical members of the group or facts countering the common beliefs) with the goal of priming users to look at others as individuals and not as stereotypical group representatives. An SCM-based approach can provide the psychological basis and the interpretation of automatic suggestions to users.

Since our methodology is intended to be part of a technology-in-the-loop approach, where the final decision on which anti-stereotypes to use and in what way will be made by human professionals, we anticipate few instances where incorrect (i.e., not related, unrealistic, or ineffective) automatically generated anti-stereotypes would be disseminated. In most such cases, since anti-stereotypes are designed to be positive, no harm is expected to be incurred on the affected group. However, it is possible that a positive, seemingly harmless anti-stereotypical description can have a detrimental effect on the target group, or possibly even introduce previously absent biases into the discourse. Further work should investigate the efficiency and potential harms of such approaches in real-life social settings.

Ethical Considerations

Data: We present a method for mapping a set of words that represent a stereotypical view of a social category held by a given subpopulation onto the two-dimensional space of warmth and competence. The Stereotype Content Model, on which the methodology is based, has been shown to be applicable across cultures, sub-populations, and time (Fiske, 2015; Fiske and Durante, 2016). Therefore, the methodology is not specific to any subpopulation or any target social group.

In the current work, we employ the publicly available portion of the StereoSet dataset (Nadeem et al., 2020). This English-only dataset has been created through crowd-sourcing US workers on Amazon Mechanical Turk. Since Mechanical Turk US workers tend to be younger and have on average lower household income than the general US population (Difallah et al., 2018), the collected data may not represent the stereotypical views of the wider population. Populations from other parts of the world, and even sub-populations in the US, may have different stereotypical views of the same social groups. Furthermore, as discussed in Section 6, the StereoSet dataset does not include stereotype data for a large number of historically marginalized groups. Future work should examine data both referring to, and produced by, a wider range of social and cultural groups.

Potential Applications: As discussed previously, the automatically proposed anti-stereotypes can be utilized by human professionals in a variety of ways, e.g., searching for or creating anti-stereotypical images, writing counter-narratives, creating educational resources, etc. One potential concern which has not received attention in the related literature is the possibility that the process of generating counter-stereotypes may itself introduce new biases into the discourse, particularly if these counter-stereotypes are generated automatically, perhaps even in response to adversarial data. We emphasize the importance of using counter-stereotypes *not* to define new, prescriptive boxes into which groups of people must fit (e.g., from Table 3, that all software developers should be intelligent and healthy, or that all entrepreneurs must be inventive and compassionate). Rather, counter-stereotypes should weaken common stereotypical associations by emphasizing that any social group is not actually homogenous, but a group of individuals with distinct traits and characteristics. In most cases, the algorithm-in-the-loop approach (with automatic suggestions assisting human users) should be adopted to reduce the risk of algorithmic biases being introduced into the public discourse.

Often, harmful stereotyping is applied to minority groups. Work on identifying and analyzing stereotypes might propagate the harmful beliefs further, and it is possible that collections of stereotypical descriptions could be misused as information sources for targeted campaigns against vulnerable populations. However, this same information is needed to understand and counter stereotypical views of society. We also note that although we take advantage of word embedding models in our approach, we do not use the representations of target group names. Previous work has shown that biased thinking is encoded in these models, and using them to represent groups can be harmful to specific demographics.

Identifying Demographic Characteristics: The proposed methodology deals with societal-level stereotypical and anti-stereotypical representations of groups of people and does not attempt to identify individual user/writer demographic characteristics. However, work on stereotyping and anti-stereotyping entails, by definition, naming and defining social categories of people. Labeling groups not only defines the category boundaries, but also positions them in a hierarchical social-category taxonomy (Beukeboom and Burgers, 2019). We emphasize that our goal is not to maintain and reproduce existing social hierarchies, as cautioned by Blodgett et al. (2020), but rather to help dismantle this kind of categorical thinking through the use of anti-stereotypes.

Energy Resources: The proposed SCM-based method is computationally low-cost, and all experiments were performed on a single CPU. Once the pretrained vectors are loaded, the projection and analysis is completed in less than a minute.

References

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counter-

speech on Twitter: A field study. *A report for Public Safety Canada under the Kanishka Project*.

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.

Irene V Blair, Jennifer E Ma, and Alison P Lenton. 2001. Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, 81(5):828.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Marco Brambilla, Simona Sacchi, Federica Castellini, and Paola Riva. 2010. The effects of status on perceived warmth and competence. *Social Psychology*, 41(2):82–87.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674, Hong Kong, China. Association for Computational Linguistics.

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

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in French tweets. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1397–1403.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Amy J. C. Cuddy, Susan T. Fiske, and Peter Glick. 2007. The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, 92(4):631.

Amy J. C. Cuddy, Susan T. Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology*, 40:61–149.

Amy J. C. Cuddy, Peter Glick, and Anna Beninger. 2011. The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior*, 31:73–98.

Nilanjana Dasgupta and Anthony G Greenwald. 2001. On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, 81(5):800–814.

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of Mechanical Turk workers. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 135–143.

John F. Dovidio, Miles Hewstone, Peter Glick, and Victoria M. Esses. 2010. *The SAGE handbook of prejudice, stereotyping and discrimination*. Sage Publications.

Thomas Eckes. 2002. Paternalistic and envious gender stereotypes: Testing predictions from the stereotype content model. *Sex Roles*, 47(3):99–114.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 task on automatic misogyny identification (AMI). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:59.

Eimear Finnegan, Jane Oakhill, and Alan Garnham. 2015. Counter-stereotypical pictures as a strategy for overcoming spontaneous gender stereotypes. *Frontiers in Psychology*, 6:1291.

Susan T. Fiske. 2010. Venus and Mars or down to Earth: Stereotypes and realities of gender differences. *Perspectives on Psychological Science*, 5(6):688–692.

Susan T. Fiske. 2015. Intergroup biases: A focus on stereotype content. *Current Opinion in Behavioral Sciences*, 3:45–50.

Susan T. Fiske. 2018. Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, 27(2):67–73.

Susan T. Fiske, Amy J. C. Cuddy, and Peter Glick. 2006. Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11(2):77–83.

Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6):878–902.

Susan T. Fiske and Federica Durante. 2016. Stereotype content across cultures: Variations on a few themes. In M. J. Gelfand, C.-Y. Chiu, and Y.-Y. Hong, editors, *Handbook of Advances in Culture and Psychology*, pages 209–258. Oxford University Press, New York, NY.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Peter Glick, Maria Lameiras, Susan T. Fiske, Thomas Eckes, Barbara Masser, Chiara Volpato, Anna Maria Manganelli, Jolynn C. X. Pek, Li-li Huang, Nuray Sakalli-Uğurlu, et al. 2004. Bad but bold: Ambivalent attitudes toward men predict gender inequality in 16 nations. *Journal of Personality and Social Psychology*, 86(5):713.

Aaron C Kay, Martin V Day, Mark P Zanna, and A David Nussbaum. 2013. The insidious (and ironic) effects of positive stereotypes. *Journal of Experimental Social Psychology*, 49(2):287–291.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Calvin K Lai, Maddalena Marini, Steven A Lehr, Carlo Cerruti, Jiyun-Elizabeth L Shin, Jennifer A Joy-Gaba, Arnold K Ho, Bethany A Teachman, Sean P Wojcik, Spassena P Koleva, et al. 2014. Reducing implicit racial preferences: A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, 143(4):1765.

Tiane L. Lee and Susan T. Fiske. 2006. Not an outgroup, not yet an ingroup: Immigrants in the stereotype content model. *International Journal of Intercultural Relations*, 30(6):751–768.

Susan C Losh, Ryan Wilke, and Margareta Pop. 2008. Some methodological issues with "Draw a Scientist Tests" among young children. *International Journal of Science Education*, 30(6):773–792.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on Twitter. *arXiv preprint arXiv:1812.02712*.

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.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gandalf Nicolas, Xuechunzi Bai, and Susan T. Fiske. 2020. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversations.

In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124, Online. Association for Computational Linguistics.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. 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 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Alexandra Timmer. 2011. Toward an anti-stereotyping approach for the European Court of Human Rights. *Human Rights Law Review*, 11(4):707–738.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

A Constructing POLAR dimensions

In contrast to the standard POLAR framework introduced by Mathew et al. (2020), we do not have a set of polar opposite word pairs, each representing a different interpretable dimension, but rather a set of words for each of the concepts warmth, coldness, competence, and incompetence from Nicolas et al. (2020). Therefore, we use a slightly different formulation to obtain the polar directions associated with warmth and competence.[4]

Let $\mathbb{D} = [\overrightarrow{\mathbb{W}^a_1}, \overrightarrow{\mathbb{W}^a_2}, \overrightarrow{\mathbb{W}^a_3}, ..., \overrightarrow{\mathbb{W}^a_V}] \in \mathbb{R}^{V \times d}$ denote the set of pretrained d-dimensional embedding vectors, trained with algorithm a, where V is the size of the vocabulary and $\overrightarrow{\mathbb{W}^a_i}$ is a unit vector representing the i_{th} word in the vocabulary.

In this work, we use four sets of seed words; a set of N_1 words associated with positive warmth $\mathbb{P}_{w+} = \{p^1_{w+}, p^2_{w+}, ..., p^{N_1}_{w+}\}$, a set of N_2 words associated with negative warmth, $\mathbb{P}_{w-} = \{p^1_{w-}, p^2_{w-}, ..., p^{N_2}_{w-}\}$, a set of N_3 words associated with positive competence, $\mathbb{P}_{c+} = \{p^1_{c+}, p^2_{c+}, ..., p^{N_3}_{c+}\}$, and a set of N_4 words associated with negative competence, $\mathbb{P}_{c-} = \{p^1_{c-}, p^2_{c-}, ..., p^{N_4}_{c-}\}$. In order to find the two polar opposites, we obtain the following directions:

$$\begin{aligned} \overrightarrow{dir_1} &= \frac{1}{N_1}\sum_{i=1}^{N_1} \mathbb{W}^a_{p^i_{w+}} - \frac{1}{N_2}\sum_{i=1}^{N_2} \mathbb{W}^a_{p^i_{w-}} \\ \overrightarrow{dir_2} &= \frac{1}{N_3}\sum_{i=1}^{N_3} \mathbb{W}^a_{p^i_{c+}} - \frac{1}{N_4}\sum_{i=1}^{N_4} \mathbb{W}^a_{p^i_{c-}} \end{aligned} \tag{A.1}$$

where $\mathbb{W}^a_\upsilon$ represents the vector of the word υ. The two direction vectors are stacked to form $dir \in \mathbb{R}^{2\times d}$, which represents the change of basis matrix for the new two-dimensional embedding subspace $\mathbb{E}$. In the new subspace, a word υ is represented by $\overrightarrow{\mathbb{E}}_\upsilon$, which is calculated using the following linear transformation:

$$\overrightarrow{\mathbb{E}}_\upsilon = (dir^T)^{-1}\mathbb{W}^a_\upsilon \tag{A.2}$$

In our experiments, we showed that, as expected, the two dimensions of $\mathbb{E}$ are associated with warmth and competence.

[4] We use the same notation as Mathew et al. (2020) to explain our method.

B Stereotype Data

Here we present all of the words contributing to each stereotype for each target group. In addition to 37 tokens which did not have vector representations in the pre-trained embeddings, there were two reasons why words were discarded from the analysis, as described in the paper. First, if a word directly referenced a demographic category, it was discarded. This was to avoid, as much as possible, including effects of language model bias in our model. For example, a number of annotators used the word *black* to describe prisoners; however, if the language model has some racial bias involving the word *black*, then it would affect the placement of the word *prisoner* on the warmth-competence plane. While acknowledging that stereotypical associations between groups are problematic and worth of study in their own right (including this disturbing example involving race and incarceration), it is beyond the scope of the current analysis.

Additional words were discarded in a filtering step, where words greater than a cosine distance of 0.6 from the mean or centroid of the group were discarded. As people's views towards different groups naturally vary, this was done to prevent outlier words from impacting the analysis, which is focused here on the most widespread or prevalent stereotypes of a given group. While heuristically chosen, the threshold value appears to be acceptable in many cases (see Table B.1). However, other times a large number of words are discarded, which appear in some cases to represent a second, coherent cluster of stereotype words for a given group (see, for example, the words for *policeman*, which split into two clusters alternately characterizing the group as corrupt and racist, or strong and heroic). As mentioned in the discussion, future work will examine how we can identify clusters of stereotypes in larger datasets.

In the following table, for each target group we present three lists of words:

- **Included:** the words included in the analysis. This list is ranked according to distance to the mean, and thus most 'representative' words occur first. Words which occur more than once have the frequency given in parentheses.
- **Discarded:** the words discarded as outliers by the thresholding step.
- **Demographic:** the words discarded as referring primarily to a demographic characteristic.

Target group	Words
Target groups from the 'gender' category	
mother	**Included:** *caring(6), loving(4), nurturing(5), comforting, protective, kind(2), love, strong, feminine, overbearing(2), children, housekeeper* **Discarded :** *old* **Demographic:** []
mommy	**Included:** *loving(4), caring(4), nurturing(5), doting, sweet, protective, understanding, child, sexy, overbearing, mama, childish, busy* **Discarded:** *kindly, nagging* **Demographic:** *female*
schoolgirl	**Included:** *smart(2), naive, immature(2), young, innocent(2), sexy, studious, artsy, girly, small, hopeful, thin, hardworker* **Discarded:** *innnocent, short, cellphone, does* **Demographic:** []
schoolboy	**Included:** *naughty(2), mischevious(3), immature(2), innocent, cute, energetic, horny, young, troublesome, brainy, troublemaker, pranks* **Discarded:** *nerd, weak, athletic, pretentious, strong, student, touched, mean, football, scholar, tag, uniform* **Demographic:** []
male	**Included:** *strong(8), dominant(3), aggressive, muscular, masculine(2), violent, strength, brute, abusive, stupid, dumb* **Discarded:** *blue, messy, ignorant, demanding, alpha, manager, sports, mansplaining* **Demographic:** []
gentlemen	**Included:** *polite(3), respectful, classy(2), considerate, gentle, kind, cordial, honorable, mannerable, handsome, proper(2), haughty, nicest, respect, bossy* **Discarded:** *nerdy, open, wealthy, wealthy, understanding, were, flask* **Demographic:** []
herself	**Included:** *emotional(2), kind, dumb, negative, upbeat, beautiful, unsure, incapable, mean girl* **Discarded:** *lonely, uptight, created, dress, pink, slap, cake, beating, shopping, alone, needlepoint, walked* **Demographic:** []
himself	**Included:** *athletic, progressive, brawny, strong, runner, workout, egotistical, gaming male, boy* **Discarded:** *funny, quiet, doctors, safe, job, aggression, admire, yardwork, enjoy, push, opinions, alone* **Demographic:** []
grandfather	**Included:** *old(8), antiquated, feeble, boring, stubborn, kind* **Discarded:** *wise(4), smart, sexist, slow(2), hero, sleeping, protected, watch, spoil, fight* **Demographic:** []
sister	**Included:** *caring(2), loving, nurturing, talkative, sassy, bossy, empathy, fragile, friend* **Discarded:** *annoying(2), close, small, bonded, protect, argumentative, naggy, noisy, mean(2), college, catty, rival* **Demographic:** *girl*
Target groups from the 'religion' category	
Muslim	**Included:** *violent(2), terrorist(6), dangerous(2), evil(2), radical(2), hateful, religious, aggressive, crazy, misogynist, intollerant, oppression, legalistic, prejudice, unprecedented* **Discarded:** *religion, quaran, hijab* **Demographic:** *men*
Brahmin	**Included:** *religious(3), sacred, ancient, foreign, temples, nobleman, loving* **Discarded:** *prideful, snobby, arrogant, haughty, wise, unknown, despised, dirty(2), prestigeous, creator(2)* **Demographic:** *Muslim, Hindu*
Target groups from the 'race or nationality' category	
Eritrean	**Included:** *poor(6), impoverished, lazy(2), needy, dirty(2), rich, starving, weird, aggressive, unsustainable* **Discarded:** *proud, gang, crime, submissive, stole, huts, artwork, rug* **Demographic:** *black(2)*
Norweigan	**Included:***blond(7), pale, tall, attractive* **Discarded:** *beautiful, wealthy(2), cold(2), strong, serious, nice, sincere, quiet* **Demographic:** *white(4)*
Saudi Arabian	**Included:** *rich(4), greedy(2), poor, spoiled, oppressed, conservative* **Discarded:** *cruel, oil, docile, pretentious, sandy, racist, hateful, prince, mecca, camel, extremist, turbans, terrorist(2), isis* **Demographic:** *Muslim(4)*
Japanese	**Included:** *friendly, smart, quiet, polite, playful, peaceful, wise, clean(2), sneaky, loyal, passive, educated, shy, exotic, proud, xenophobic* **Discarded:** *sexist, short, academic, work, kamikaze, sushi, seafood, workaholics* **Demographic:** *black*
Russian	**Included:** *drunk(4), alcoholic, drinking, angry, stupid, vodka, cold, homophobic* **Discarded:** *corrupt, suspicious, scary, outspoken, communist, brutal, spy, deceitful, spies, slavic, shady, strong, tough, stern, mean(2), scammer, control, communism* **Demographic:** []
Ethiopian	**Included:** *skinny(7), starving(5), hungry(2), emaciated, thin Christian* **Discarded:** *empty, unfortunate, corrupt, somber, stews, save* **Demographic:** []
Columbian	**Included:** *drugs(8), cocaine(2), druggie, coffee(2), criminal, druglord* **Discarded:** *football, hot, sexy(2), local, educated, poor, voluptuous, undeveloped, fiery* **Demographic:** []
Crimean	**Included:** *poor, impoverished, weak, nonexistent, corrupt, harsh, indigenous, boring, wartorn, native, abundance, drunk* **Discarded:** radical, communist, exotic, independent, controversial, stolen, combative, crime, war, lawlessness, dumber, Ukraine **Demographic:** Islam, Russian, black, white
Iranian	**Included:** *dangerous(4), violent, scary, hateful, bloodthirsty, evil, foolish, dumb, terrorist(2), religious, pretentious, conservative, rogue, greasy* **Discarded:** *terrorists(3), panicked, war, fragrant, traditional, hookah* **Demographic:** []

Persian people	**Included:** *greedy(2), rich(2), uneducated(2), poor, extravagant, disgusting, intelligent, mysterious, flashy* **Discarded:** *hairy, smelly, violent, religious, smoke, strict, sand, naan* **Demographic:** *black, Muslim*
Arab	**Included:** *violent, dangerous(2), evil, terrorist(2), radical, crazy, religious(2), extremism, extreme, bad, cruel, terrorists(2), dirty(2), conservative, misogynist, war, diverse* **Discarded:** *rich(2), oil, turban* **Demographic:** *Muslim*
African	**Included:** *poor(3), impoverished, starving(2), dirty, stupid, harsh, dark, fast(2), violent* **Discarded:** *savages, criminal, crackhead, stealing, run, atheltic* **Demographic:** *black(6)*
Ghanaian	**Included:** *poor(4), dangerous(2), dark(2), lazy, strange, reckless, brutish, fun, starving* **Discarded:** *friendly, athletic, criminals, tribal, less, gold, rapists, tall, dated* **Demographic:** *black(2), Indian*
Ukrainian	**Included:** *stupid, dumb, poor(3), dirty, hateful, cynical, evil, unintelligent, pretty, cheap, powerless, unclean, exotic, criminal, interesting* **Discarded:** *religious, isolationist, quiet, freezing, gangster, discernible, vodka(3), allies, borscht* **Demographic:** *Jewish, white, Russian*
Bengali	**Included:** *dirty(2), stinky, poor(2), dangerous, smelly(2), cold, primitive, wild, tough, rich, smart, indecent, adventorous* **Discarded:** *macho, dancing, conservative, curry, rice, tiger, eastern, india, tigers(2)* **Demographic:** []
Hispanic	**Included:** *poor(3), lazy(2), dirty(2), nasty, uneducated, shady, illegal(2), aggressive, hardworking(2), beautiful, greasy, undocumented* **Discarded:** *loud, smelly, immigrants, farmworker, spicy, spanish, gardener, season, equator* **Demographic:** *Catholic*
Target groups from the 'profession' category	
psychologist	**Included:** intelligent(3), smart(3), empathetic(3), compassionate, witty, analytical(3), educated(2), opinionated, careful, observant, curious, interesting, calm, rigid **Discarded:** calming, nosy, wackjobs, helps **Demographic:** []
manager	**Included:** bossy(8), overbearing, rude, selfcentered, boss, lazy, authoritative, leader **Discarded:** strict(2), controlling, boring, knowledgeable, corrupt, organized(3), mean, seriousness **Demographic:** white, man(2), he(2)
guitarist	**Included:** energetic, passionate, talented(2), musical(3), young, sensitive, emotional, skilled, cool, stimulating, attitude, lazy, hippie **Discarded:** skinny, drunk, rock, stoned, dedication, druggies **Demographic:** male, Caucasian
civil servant	**Included:** lazy, tired, overworked(2), poor, bad, careless, selfish, bored, stressed, corrupt, smart, dependable **Discarded:** bureaucratic, kind, competent, honorable, bureaucrat(2), disrespected, democrat, state, job, volunteering **Demographic:** []
butcher	**Included:** brutal, bloody(2), violent, bloodthirsty, filthy, strong(2), sloppy, dumb, burly, macho, cuts **Discarded:** meat, fat(3), knives, foreign, uneducated(2), experienced, overweight **Demographic:** he(2), white, man, male
commander	**Included:** strong, ruthless, harsh, wise, friendly, honest, intelligent, brave, hard, trustworthy, loud(2), important, sharp, protective, strict(2), dominant **Discarded:** leading, experienced, leader(2), organized, bossy(2), controlling **Demographic:** male(2), him
chess player	**Included:** smart(4), strategic(4), clever, brilliant, analytical, methodical, careful, nerdy, strategy, quiet, dorky, calm, geek **Discarded:** reserved, frail, calculated, glasses **Demographic:** Asian(3), man, Russian
performing artist	**Included:** talented(4), charismatic, creative, skilled, egotistical(2), opinionated, attractive, expressive, outgoing, dancer **Discarded:** selfish, ballerina, musical, big, emotional, ostentatious, broke **Demographic:** gay
plumber	**Included:** uneducated(3), lazy, illiterate, dirty(2), uncouth, fat(3), unkempt, crooked, redneck, lackadaisical, hardworking **Discarded:** greasy, expensive(2), lower, fix, installs **Demographic:** male, Italian
tailor	**Included:** precise(2), meticulous, detailed(2), careful, professional(2), neat, skilled(2), creative, quiet(2), short **Discarded:** snarky, crafty, foreigner, james **Demographic:** Italian(2), gay, Asian, man, Jew, female
chemist	**Included:** smart(5), intelligent(3), nerdy(3), geeky(2), studious, crafty, analytical, introverted, careful, brillant **Discarded:** shy, science, laboratory **Demographic:** white, man, he
musician	**Included:** creative(4), artistic(2), talented(4), professional, good, emotional, cute, intense, informal **Discarded:** emotion, poor, employed, flakey, famous, handsome, shiftless, addict, unreliable, broke **Demographic:** []
software developer	**Included:** nerdy(7), geeky(2), smart(6), nerd, intelligent(2), genius, math, unhealthy **Discarded:** boys, rich(2), coder, glasses, skilled **Demographic:** male
mover	**Included:** strong(18), burly, hardworking, dumb, tall **Discarded:** creepy, careless **Demographic:** men(2), Mexican
football player	**Included:** dumb(4), stupid(3), strong(5), tough(2), aggressive(2), fast, talented **Discarded:** bully, muscular, hulking, athletic(3) injured, buff, athlete, doping **Demographic:** black (2), male

producer	**Included:** shrewd, smart, egotistical, selfish, flashy, industrious, aloof, rich, visionary, knowledgable, businesslike, creepy, busy, critical, cheap **Discarded:** predatory, tired, leader, ideas, elitist, demanding(3), film **Demographic:** white, male(2)
politician	**Included:** dishonest(3), corrupt(8), deceptive, crooked, liars(3), shifty, lying(2), disingenuine, cheating, fake, slippery, argumentative **Discarded:** [] **Demographic:** female, male
prosecutor	**Included:** aggressive(5), ruthless, agressive, vindictive, harsh(2), tough, unfair, unrelenting, aggressively, sneaky, biased, contentious, persuasive, racist, corrupt, confident **Discarded:** educated, guilty(2), lawyer, justice **Demographic:** male(2)
bartender	**Included:** friendly(3), sociable(2), personable, gregarious, outgoing(2), beautiful, sexy, lovely, quick, fast, stupid **Discarded:** uneducated, criminal, alcoholic(2), drunk(3), listens, drinker, provocatively, listener, listeners **Demographic:** []
physicist	**Included:** smart(7), intelligent(4), nerdy(5), brilliant, eccentric, genius(2), educated, dynamic, focused **Discarded:** pety **Demographic:** man, his
assistant	**Included:** helpful(4) efficient, diligent, organized, helper **Discarded:** harried, independent, meek, young, obediant, slutty, meak, unnoticed, clerical, coffee, grunt, inferior, coward **Demographic:** female
historian	**Included:** bookish(4), nerdy, studious, bookworm, smart(3), academic, professorial, library **Discarded:** boring(5), dull, educated, ancient, wise, correct **Demographic:** []
researcher	**Included:** studious, nerdy(3), inquisitive, smart, analytical(2), careful, meticulous, thorough, introverted, competent, dedicated(2), awkward, shy, inquiring **Discarded:** boring, coldhearted, educator, overworked, intern, library, computer **Demographic:** man, Asian
delivery man	**Included:** young(3), hardworking(2), poor, careless(2), stupid, teenager, strong, late(2), athletic **Discarded:** kind, fit, screwing, efficient, rushed, timely, distribute **Demographic:** white, black, male
entrepreneur	**Included:** savvy(3), ambitious(2), smart, greedy(3), ruthless, young(2), creative(2), industrious, wise, rich(2), successful, inventive, talkative, sleazy, educated, extroverted **Discarded:** hardworker, driven, free, boss **Demographic:** male
policeman	**Included:** corrupt(4), racist(2), megalomaniacal, bad, aggressive **Discarded:** fat, murderer, strong, safe, bully, macho, agressive, brave, hero, assaults, stern, mean, helpful **Demographic:** white(2)
engineer	**Included:** smart(6), intelligent, nerdy(3), geeky, knowledgeable, egotistical, wealthy, awkward, rich, boring, intellegant, methodical **Discarded:** technical, antisocial, update **Demographic:** man, Chinese, he, male
mathematician	**Included:** smart(6), intelligent(3), nerdy(4), geek, analytical, analytical, nerds, good, geniuses, intelligence, meek **Discarded:** logic, antisocial, introvert, numbers, algebra **Demographic:** he, man(2)
nurse	**Included:** caring(8), compassionate, hardworking, supportive, patient, kind, dedicated **Discarded:** overworked, profession, nice, busy, tired, hot, underqualified **Demographic:** woman(2), her, female(2), she, male
prisoner	**Included:** violent(5), dangerous(2), brutal, cruel, evil(2), criminal, bad, dishonest, untrustworthy, hopeless, thug, lazy, perpetrator **Discarded:** smelly, guilty(2), mean(2) **Demographic:** black(6)

Table B.1: Target groups and associated stereotype words in StereoSet. Words which occur more than once for a given group have their frequency indicated in parentheses. Words that are included in the analysis are ranked by closeness to the cluster mean; thus the first words in the list are most representative of the stereotype for that group.