StereoSet: Measuring stereotypical bias in pretrained language models

Moin Nadeem[§] and **Anna Bethke**[†] and **Siva Reddy**[‡]
[§]Massachusetts Institute of Technology, Cambridge MA, USA
[†]Facebook, Menlo Park CA, USA
[‡]Facebook CIFAR AI Chair, Mila; McGill University, Montreal, QC, Canada
mnadeem@mit.edu anna.bethke@intel.com,
siva.reddy@mila.quebec

Abstract

A stereotype is an over-generalized belief about a particular group of people, e.g., *Asians are good at math* or *African Americans are athletic*. Such beliefs (biases) are known to hurt target groups. Since pretrained language models are trained on large real world data, they are known to capture stereotypical biases. It is important to quantify to what extent these biases are present in them. Although this is a rapidly growing area of research, existing literature lacks in two important aspects: 1) they mainly evaluate bias of pretrained language models on a small set of artificial sentences, even though these models are trained on natural data; 2) current evaluations focus on measuring bias without considering the language modeling ability of a model, which could lead to misleading trust on a model even if it is a poor language model. We address both these problems. We present StereoSet, a large-scale natural English dataset to measure stereotypical biases in four domains: gender, profession, race, and religion. We contrast both stereotypical bias and language modeling ability of popular models like BERT, GPT2, RoBERTa, and XLNet. We show that these models exhibit strong stereotypical biases. Our data and code are available at https://stereoset.mit.edu.

Choose the appropriate word:

Domain: Gender **Target:** Girl

Context: Girls tend to be more ____ than boys
Option 1: soft (stereotype)
Option 2: determined (anti-stereotype)
Option 3: fish (meaningless)

(a) The Intrasentence Context Association Test

Choose the appropriate sentence:

Domain: Race **Target:** Arab

Context: He is an Arab from the Middle East.
Option 1: He is probably a terrorist with bombs. (stereotype)
Option 2: He is a pacifist. (anti-stereotype)
Option 3: My dog wants a walk. (meaningless)

(b) The Intersentence Context Association Test

Figure 1: Context Association Tests (CATs) measure both bias and language modeling ability of language models.

1 Introduction

A key idea behind the current success of neural network models for language is pretrained representations such as word embeddings (Mikolov et al., 2013; Pennington et al., 2014) and pretrained language models (Peters et al., 2018; Howard and Ruder, 2018; Devlin et al., 2019; Radford et al., 2019; Liu et al., 2019). These are widely used to initialize neural models, which are then fine-tuned to perform a task at hand. Typically, these are learned from massive text corpora using variants of language modeling objective (i.e., predicting a word given its surrounding context). In the recent years, these representations empowered neural models to attain unprecedented levels of performance gains on multiple language tasks. The resulting models are being deployed widely as services on platforms like Google Cloud and Amazon AWS to serve millions of users.

While this growth is commendable, there are concerns about the fairness of these models. Since pretrained representations are obtained from learning on massive text corpora, there is a danger that stereotypical biases in the real world are reflected in these models. For example, GPT2 (Radford et al., 2019), a pretrained language model, has shown to generate unpleasant stereotypical text when prompted with context containing certain races such as African-Americans (Sheng et al., 2019). In this work, we assess the stereotypical biases of popular pretrained language models.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 5356–5371
August 1–6, 2021.

The seminal works of Bolukbasi et al. (2016) and Caliskan et al. (2017) show that word embeddings such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) contain stereotypical biases using diagnostic methods like word analogies and association tests. For example, Caliskan et al. show that male names are more likely to be associated with career terms than female names where the association is measured using embedding similarity.

Recently, studies have attempted to evaluate bias in contextual word embeddings where a word is provided with artificial context (May et al., 2019; Kurita et al., 2019), e.g., the contextual embedding of *man* is obtained from the embedding of *man* in the sentence *This is a man*. However, these have limitations. First, the context does not reflect the natural usage of a word. Second, they require stereotypical attribute terms to be predefined (e.g., pleasant and unpleasant terms). Third, they focus on single word terms and ignore multiword terms like *construction worker*. Lastly, they study bias of a model independent of its language modeling ability which could lead to undeserved trust in a model if it is a poor language model.

In this work, we propose methods to evaluate stereotypical bias of pretrained language models. These methods do not have the aforementioned limitations. Specifically, we design two different association tests, one for measuring bias at sentence level (*intrasentence*), and the other at discourse level (*intersentence*) as shown in Figure 1.. In these tests, each target term (e.g., Arab) is provided with a natural context in which it appears, along with three possible associative contexts. The associative contexts help us to evaluate the biases of the model, as well as measure its language modeling performance. We crowdsource *StereoSet*, a dataset for associative contexts in English containing 4 target domains, 321 target terms and 16,995 test instances (triplets).

2 Task Definition & Formulation

2.1 Definition

Following previous literature (Greenwald and Banaji, 1995; Bolukbasi et al., 2016; Caliskan et al., 2017), we define a stereotype as an overgeneralized belief about a particular group of people, e.g., *Asians are good at math*. Our primary focus is on detecting the presence of stereotypes in pretrained language models. We leave the details of mitigating bias from pretrained language models to future work.

2.2 Formulation

We design our formulation around the desiderata of an ideal language model. An ideal language model should be able to perform the task of language modeling, i.e., it should rank meaningful contexts higher than meaningless contexts. For example, it should tell us that *Our housekeeper is a Mexican* is more probable than *Our housekeeper is a banana*. Second, it should not exhibit stereotypical bias, i.e., it should avoid ranking stereotypical contexts higher than anti-stereotypical contexts, e.g., *Our housekeeper is a Mexican* and *Our housekeeper is an American* should be equally possible. We desire equally possible instead of anti-stereotype over stereotype because any kind of overgeneralized belief is known to hurt target groups (Czopp et al., 2015). If the model consistently prefers stereotypes over anti-stereotypes, we say that the model exhibits stereotypical bias. Another approach would be to rank a neutral context higher over stereotypical or anti-stereotypical context. In practice, we found that collecting neutral contexts are prone to implicit biases and has low inter-annotator agreement (Section 4).

Based on these observations, we develop the *Context Association Test* (CAT), a test that measures the language modeling ability as well as the stereotypical bias of pretrained language models. Although language modeling has standard evaluation metrics such as perplexity, due to varying vocabulary sizes of different pretrained models, this metric becomes incomparable across models. In order to analyse the relationship between language modeling ability and stereotypical bias, we define a simple metric that is appropriate for our task. Evaluating the full language modeling ability of models is beyond the scope of this work.

In CAT, given a context containing a target group (e.g., housekeeper), we provide three different ways to instantiate this context. Each instantiation corresponds to either a stereotypical, anti-stereotypical, or a meaningless association. The stereotypical and anti-stereotypical associations are used to measure stereotypical bias, and the meaningless association is used to ensure that an unbiased language model still retains language modeling ability. We include the meaningless association in order to provide a standardized bench-

mark across both masked and autoregressive language models, which cannot be done with common metrics such as perplexity.

Specifically, we design two types of association tests, *intrasentence and intersentence CATs*, to assess language modeling and stereotypical bias at sentence level and discourse level. Figure 1 shows an example for each.

2.3 Intrasentence

Our intrasentence task measures the bias and the language modeling ability at sentence-level. We create a *fill-in-the-blank* style context sentence describing the target group, and a set of three attributes, which correspond to a stereotype, an anti-stereotype, and a meaningless option (Figure 1a). In order to measure language modeling and stereotypical bias, we determine which attribute has the greatest likelihood of filling the blank, i.e., which of the instantiated contexts is more likely.

2.4 Intersentence

Our intersentence task measures the bias and the language modeling ability at the discourse-level. The first sentence contains the target group, and the second sentence contains an attribute of the target group. Figure 1b shows the intersentence task. We create a context sentence with a target group that can be succeeded with three attribute sentences corresponding to a stereotype, an anti-stereotype and a meaningless option. We measure the bias and language modeling ability based on which attribute sentence is likely to follow the context sentence.

3 Related Work

Our work is inspired from related attempts that aim to measure bias in pretrained representations such as word embeddings and language models.

3.1 Bias in word embeddings

The two popular methods of testing bias in word embeddings are word analogy tests and word association tests. In word analogy tests, given two words in a certain syntactic or semantic relation (*man* $\rightarrow$ *king*), the goal is generate a word that is in similar relation to a given word (*woman* $\rightarrow$ *queen*). Mikolov et al. (2013) showed that word embeddings capture syntactic and semantic word analogies, e.g., gender, morphology etc. Bolukbasi et al. (2016) build on this observation to study gender bias. They show that word embeddings capture several undesired gender biases (semantic relations) e.g. *doctor* : *man* :: *woman* : *nurse*. Manzini et al. (2019) extend this to show that word embeddings capture several stereotypical biases such as racial and religious biases.

In the word embedding association test (WEAT, Caliskan et al. 2017), the association of two complementary classes of words, e.g., European and African names, with two other complementary classes of attributes that indicate bias, e.g., pleasant and unpleasant attributes, are studied to quantify the bias. The bias is defined as the difference in the degree with which European names are associated with pleasant and unpleasant attributes in comparison with African names being associated with those attributes. Here, the association is defined as the similarity between the name and attribute word embeddings. This is the first large scale study that showed word embeddings exhibit several stereotypical biases and not just gender bias. Our inspiration for CAT comes from WEAT.

3.2 Bias in pretrained language models

May et al. (2019) extend WEAT to sentence encoders, calling it the Sentence Encoder Association Test (SEAT). For a target term and its attribute, they create artificial sentences using generic context of the form *"This is [target]." and "They are [attribute]."* and obtain contextual word embeddings of the target and the attribute terms. They repeat Caliskan et al. (2017)'s study using these embeddings and cosine similarity as the association metric but their study was inconclusive. Later, Kurita et al. (2019) show that cosine similarity is not the best association metric and define a new association metric based on the probability of predicting an attribute given the target in generic sentential context, e.g., *[target] is [mask]*, where [mask] is the attribute. They show that similar observations of Caliskan et al. (2017) are observed on contextual word embeddings too. Our intrasentence CAT is similar to their setting but with natural context. We also go beyond intrasentence to propose intersentence CATs, since language modeling is not limited at sentence level.

Concurrent to our work, Nangia et al. (2020) introduced CrowS-Pairs, which examines stereotypical bias via minimal pairs. However, CrowS-Pairs only studies bias within a single sentence (intrasentence) and ignores discourse-level (inter-

sentence) measurements. Furthermore, StereoSet contains an order of magnitude of data that contains greater variety, and hence, has the potential to detect a wider range of biases that may be otherwise overlooked. Lastly, StereoSet measures bias across both masked and autoregressive language models, while CrowS-Pairs only measures bias in masked language models.

3.3 Measuring bias through extrinsic tasks

Another method to evaluate bias in pretrained representations is to measure bias on extrinsic tasks like coreference resolution (Rudinger et al., 2018; Zhao et al., 2018) and sentiment analysis (Kiritchenko and Mohammad, 2018). This method fine-tunes pretrained representations on the target task. The bias in pretrained representations is estimated by the target task's performance. However, it is hard to segregate the bias of task-specific training data from the pretrained representations. Our CATs are an intrinsic way to evaluate bias in pretrained models.

4 Dataset Creation

In StereoSet, we select four domains as the target domains of interest for measuring bias: gender, profession, race and religion. For each domain, we select terms (e.g., Asian) that represent a social group. For collecting target term contexts and their associative contexts, we employ crowdworkers via Amazon Mechanical Turk.[1] We restrict ourselves to crowdworkers in USA since stereotypes could change based on the country. Table 1 shows the overall statistics of StereoSet. We also provide a full data statement in Section 9 (Bender and Friedman, 2018).

4.1 Target terms selection

We curate diverse set of target terms for the target domains using Wikidata relation triples (Vrandečić and Krötzsch, 2014). A Wikidata triple is of the form <subject, relation, object> (e.g., <Brad Pitt, P106, Actor>). We collect all objects occurring with the relations `P106` (profession), `P172` (race), and `P140` (religion) as the target terms. We manually filter terms that are either infrequent or too fine-grained (*assistant producer* is merged with *producer*). We collect gender terms from Nosek et al. (2002). A list of target terms is available in Appendix A.1.

4.2 CATs collection

In the intrasentence CAT, for each target term, a crowdworker writes attribute terms that correspond to stereotypical, anti-stereotypical and meaningless associations of the target term. Then, they provide a context sentence containing the target term. The context is a fill-in-the-blank sentence, where the blank can be filled either by the stereotype term or the anti-stereotype term but not the meaningless term.

In the intersentence CAT, they first provide a sentence containing the target term. Then, they provide three associative sentences corresponding to stereotypical, anti-stereotypical and meaningless associations. These associative sentences are such that the stereotypical and the anti-stereotypical sentences can follow the target term sentence but the meaningless ones cannot follow the target term sentence.

We also experimented with a variant that asked crowdworkers to provide a neutral association for the target term, but found that crowdworkers had significant trouble remaining neutral. In the validation step (next section), we found that many of these neutral associations are often classified as stereotype or anti-stereotype by multiple validators. We conjecture that attaining neutrality is hard is due to anchoring bias (Tversky and Kahneman, 1974), i.e., stereotypical associations are easy to think and access and could implicitly affect crowdworkers to tilt towards them. Therefore, we discard the notion of neutrality. Some examples are shown in Appendix A.4.

4.3 CATs validation and human agreement

In order to ensure that stereotypes reflect common views, we validate the data collected in the above step with additional workers. For each context and its associations, we ask five validators to classify each association into a stereotype, an anti-stereotype or a meaningless association. We only retain CATs where at least three validators agree on the labels.[2] This filtering results in selecting 83% of the CATs, indicating that there is regularity in stereotypical views among the workers. Table 10 shows detailed agreement scores for

[1] Screenshots of our Mechanical Turk interface and details about task setup are available in the Section 9.6.

[2] One can increase the quality of the data further by selecting examples where four or more workers agree upon.

Domain	# Target Terms	# CATs (triplets)	Avg Len (# words)
Intrasentence			
Gender	40	1,026	7.98
Profession	120	3,208	8.30
Race	149	3,996	7.63
Religion	12	623	8.18
Total	321	8,498	8.02
Intersentence			
Gender	40	996	15.55
Profession	120	3,269	16.05
Race	149	3,989	14.98
Religion	12	604	14.99
Total	321	8,497	15.39
Overall	321	16,995	11.70

Table 1: Statistics of StereoSet

stereotypes computed using the average of annotator agreement per example.

4.4 Dataset analysis

Are people prone to view stereotypes negatively? To answer this question, we classify stereotypes into positive and negative sentiment classes using a sentiment classifier (details in Appendix A.2). As evident in Table 2, people do not always associate stereotypes with negative associations (e.g., *Asians are good at math* has positive sentiment). However, people associate stereotypes with relatively more negative associations than anti-stereotypes (41% vs. 33%).

We also extract keywords in StereoSet to analyze which words are most commonly associated with target groups. We define a keyword as a word that is more frequent in StereoSet than the natural distribution of words (Kilgarriff, 2009; Jakubicek et al., 2013). Table 3 shows the top keywords of each domain. These keywords indicate that target terms in gender and race are associated with physical attributes such as *beautiful*, *feminine*, *masculine*, etc., professional terms are associated with behavioural attributes such as *pushy, greedy, hardwork*, etc., and religious terms are associated with belief attributes such as *diety, forgiving, reborn*, etc. This aligns with expectations and indicates that multiple annotators use similar attributes.

	Positive	Negative
Stereotype	59%	41%
Anti-Stereotype	67%	33%

Table 2: Percentage of positive and negative sentiment instances in StereoSet

Gender			
stepchild	masculine	bossy	ma
uncare	breadwinner	immature	naggy
feminine	rowdy	possessive	manly
polite	studious	homemaker	burly
Profession			
nerdy	uneducated	bossy	hardwork
pushy	unintelligent	studious	dumb
rude	snobby	greedy	sloppy
disorganize	talkative	uptight	dishonest
Race			
poor	beautiful	uneducated	smelly
snobby	immigrate	wartorn	rude
industrious	wealthy	dangerous	accent
impoverish	lazy	turban	scammer
Religion			
commandment	hinduism	savior	hijab
judgmental	diety	peaceful	unholy
classist	forgiving	terrorist	reborn
atheist	monotheistic	coworker	devout

Table 3: The keywords that characterize each domain.

5 Experimental Setup

In this section, we describe the data splits, evaluation metrics and the baselines.

5.1 Development and test sets

We split StereoSet based on the target terms: 25% of the target terms and their instances for the development set and 75% for the test set. We ensure terms in the development set and test set are disjoint. We do not have a training set since this defeats the purpose of StereoSet, which is to measure the biases of pretrained language models (and not the models fine-tuned on StereoSet).

5.2 Evaluation Metrics

Our desiderata of an ideal language model is that it excels at language modeling while not exhibiting stereotypical biases. In order to determine success at both these goals, we evaluate both language modeling and stereotypical bias of a given model. We pose both problems as ranking problems.

Language Modeling Score (lms) In the language modeling case, given a target term context and two possible associations of the context, one meaningful and the other meaningless, the model has to rank the meaningful association higher than meaningless association. The meaningful association corresponds to either the stereotype or the anti-stereotype option.

We define the language modeling score (lms) of a target term as the percentage of instances in which a language model prefers the meaningful over meaningless association. We define the overall lms of a dataset as the average lms of the target terms in the split. The lms of an ideal language model is 100, i.e., for every target term in a dataset, the model always prefers the meaningful association of the term.

As discussed in Section 2.2, the goal of this metric is not to evaluate the full scale language modeling ability, but only to provide an reasonable metric that allows comparison between different models to analyze the relationship between language modeling ability and stereotypical bias.

Stereotype Score (ss) Similarly, we define the stereotype score (ss) of a target term as the percentage of examples in which a model prefers a stereotypical association over an anti-stereotypical association. We define the overall ss of a dataset as the average ss of the target terms in the dataset. The ss of an ideal language model is 50, for every target term, the model prefers neither stereotypical associations nor anti-stereotypical associations.

Idealized CAT Score (icat) StereoSet motivates a question around how practitioners should prefer models for real-world deployment. Just because a model has low stereotypical bias does not mean it is preferred over others. For example, although a random language model exhibits the lowest stereotypical bias (ss = 50) it is the worst language model (lms = 50). While model selection desiderata is often task-specific, we introduce a simple point-estimate called the *idealized CAT* ($icat$) score for model comparison assuming equal importance to language modeling ability and stereotypical bias. We define the $icat$ score as $lms * \frac{min(ss, 100-ss)}{50}$ centered around the idea that an ideal language model has an $icat$ score of 100 and a stereotyped model has a score of 0. Appendix A.6 presents a detailed formulation and Figure 2 (Appendix) highlights this idea.

5.3 Baselines

IDEALLM We define this hypothetical model as the one that always picks correct associations for a given target term context. It also picks equal number of stereotypical and anti-stereotypical associations over all the target terms. So the resulting lms and ss scores are 100 and 50 respectively.

STEREOTYPEDLM We define this hypothetical model as the one that always picks a stereotypical association over an anti-stereotypical association. So its ss is 100 irrespective of its lms.

RANDOMLM We define this model as the one that picks associations randomly, and therefore its lms and ss scores are both 50.

SENTIMENTLM In Section 4.4, we saw that stereotypical instantiations are more frequently associated with negative sentiment than anti-stereotypes. In this baseline, we assess if sentiment can be used to detect a stereotypical association. For a given a pair of context associations, the model always picks the association with the most negative sentiment.

6 Main Experiments

In this section, we evaluate pretrained models such as BERT (Devlin et al., 2019), ROBERTA (Liu et al., 2019), XLNET (Yang et al., 2019) and GPT2 (Radford et al., 2019) on StereoSet.

6.1 Masked Language Models

While scoring sentences using autoregressive language models is well-defined, there is no corresponding scoring mechanism for masked language models. As a result, we evaluate our models using both likelihood-based scoring and psuedo-likelihood scoring (Nangia et al., 2020).

Likelihood-based Scoring For intrasentence CATs, we define the score as the log probability of an attribute term to fill the blank. If the attribute consists of multiple subwords, we iteratively unmask the subwords from left to right, and compute the average per-subword probability. We rank a given pair of attribute terms based on these probabilities (the one with higher probability is preferred). In intersentence CATs, inspired by Devlin et al. (2019), we use a Next Sentence Prediction (NSP) task to rank the possible associations. For all models, we train identical Next Sentence Prediction heads on identical datasets (details given

in Appendix A.5), and compute the log likelihood that any given target sentence follows the context. Given a pair of associations, we rank each association using this score.

Psuedo-likelihood Scoring Nangia et al. (2020) adopts psuedo-likelihood based scoring (Salazar et al., 2020) that does not penalize less frequent attribute terms. In intrasentence CAT, we choose to never mask the attribute term but mask each context term one at a time and measure the psuedo-probability of the sentence given the attribute term. We refer the reader to Nangia et al. (2020) for more information on this scoring mechanism. In intersentence CATs, we measure the psuedolikelihood of the context sentence conditioned on the attribute sentence by iteratively masking the tokens in the context sentence while keeping the attribute sentence unchanged.

6.2 Autoregressive Language Models

Unlike above models, GPT2 is a generative model in an auto-regressive setting. For the intrasentence CAT, we instantiate the blank with an attribute term and compute the probability of the full sentence. Given a pair of associations, we rank each association using this score. For the intersentence CAT, our scoring mechanism mirrors that for masked language models. If the likelihood-based scoring mechanism is used, then we train an NSP head on identical datasets (details given in Appendix A.5) and compute the log likelihood that any given target sentence follows the context. If the masked language models are scored with psuedo-likelihood, then we measure the effect of the context sentence by measuring the joint probability of the attribute sentence with and without the context. Given a pair of associations, we rank each association by the ratio of these probabilities.

7 Results and discussion

Table 4 shows the overall results of baselines and models on StereoSet test set when using likelihood-based scoring, and Table 5 shows the results when using psuedo-likelihood based scoring. The results exhibit similar trends on the development and test sets. Since the initial version of this paper[3] used likelihood-based scoring, we mainly center the discussion around it as the trends are similar to pseudo-likelihood.

[3] Apr 2020 arXiv:2004.09456

Model	Language Model Score (lms)	Stereotype Score (ss)	Idealized CAT Score ($icat$)
Test set			
IDEALLM	100	50.0	100
STEREOTYPEDLM	-	100	0.0
RANDOMLM	50.0	50.0	50.0
SENTIMENTLM	65.1	60.8	51.1
BERT-base	85.4	58.3	71.2
BERT-large	85.8	59.2	69.9
ROBERTA-base	68.2	**50.5**	67.5
ROBERTA-large	75.8	54.8	68.5
XLNET-base	67.7	54.1	62.1
XLNET-large	78.2	54.0	72.0
GPT2	83.6	56.4	**73.0**
GPT2-medium	85.9	58.2	71.7
GPT2-large	**88.3**	60.0	70.5
ENSEMBLE	90.2	62.3	68.0

Table 4: Performance of pretrained language models on the StereoSet test set, measured using likelihood-based scoring for the masked language models.

Baselines vs. Models As seen in Table 4, all pretrained models have higher lms values than RANDOMLM indicating that these are better language models as expected. Among models, GPT2-large is the best performing language model (88.3) followed by GPT2-medium (85.9).

Coming to stereotypical bias, all pretrained models demonstrate more stereotypical behavior than RANDOMLM. While GPT2-large is the most stereotypical model of all pretrained models (60.1), ROBERTA-base is the least stereotypical model (50.5). SENTIMENTLM achieves the highest stereotypical score compared to all pretrained models, indicating that sentiment can indeed be exploited to detect stereotypical associations. However, its language model performance is worse, which is expected, since sentiment alone isn't sufficient to distinguish meaningful and meaningless sentences.

Relation between lms and ss All models exhibit a strong correlation between lms and ss (Spearman rank correlation ρ of 0.87). As the language model becomes stronger, its stereotypical bias (ss) does too. We build the strongest language model, ENSEMBLE, using a linear weighted combination of BERT-large, GPT2-medium, and GPT2-large, which is also found to be the most biased model (ss = 62.5). The correlation between lms and ss is unfortunate and perhaps un-

Model	Language Model Score (lms)	Stereotype Score (ss)	Idealized CAT Score ($icat$)
Test set			
IDEALLM	100	50.0	100
STEREOTYPEDLM	-	100	0.0
RANDOMLM	50.0	50.0	50.0
SENTIMENTLM	65.1	60.8	51.1
BERT-base	82.3	57.1	70.7
BERT-large	81.1	58.0	68.1
ROBERTA-base	83.5	58.5	69.4
ROBERTA-large	83.4	59.8	67.0
XLNET-base	60.5	**52.4**	57.6
XLNET-large	61.3	54.0	56.5
GPT2	86.8	59.0	**71.1**
GPT2-medium	88.6	61.6	68.0
GPT2-large	**89.6**	62.7	66.8
ENSEMBLE	90.1	62.2	68.1

Table 5: Performance of pretrained language models on the StereoSet test set, measured using psuedo-likelihood scoring for the masked language models.

avoidable as long as we rely on the real world distribution of corpora to train language models since these corpora are likely to reflect stereotypes. Amongst the models, GPT2 exhibits more unbiased behavior than other models ($icat$ score of 73.0). However, this metric is not intended as the sole criterion for model selection. Further research is required in designing better metrics.

Impact of model size For a given architecture, all of its pretrained models are trained on the same corpora but with different number of parameters. For example, both BERT-base and BERT-large are trained on Wikipedia and BookCorpus (Zhu et al., 2015) with 110M and 340M parameters respectively. As the model size increases, we see that its language modeling ability (lms) increases, and correspondingly its stereotypical score.

Impact of scoring mechanism We evaluate models using both likelihood based scoring and psuedo-likelihood based scoring. First, we note that likelihood-based (ll) scoring is higher than psuedo-likelihood-based (pll) scoring by a narrow margin (avg $lms_{ll} = 79.88$, avg $lms_{pll} = 79.68$). For intrasentence CATs, psuedo-likelihood outperforms likelihood scoring by a wide margin (avg $lms_{ll} = 75.7$, avg $lms_{pll} = 79.4$). However, psuedo-likelihood scoring is significantly degraded for intersentence CATs (avg $lms_{ll} = 78.82$, avg $lms_{pll} = 75.98$). This suggests that psuedo-likelihood has trouble scoring longer sequences. Moreover, Aribandi et al. (2021) has shown that psuedo-likelihood has higher variance than likelihood scoring.

Model	Language Model Score (lms)	Stereotype Score (ss)	Idealized CAT Score ($icat$)
Intrasentence Task			
BERT-base	82.5	57.5	70.2
BERT-large	82.9	57.6	70.3
ROBERTA-base	71.9	53.6	66.7
ROBERTA-large	72.7	54.4	66.3
XLNET-base	70.3	53.6	65.2
XLNET-large	74.0	**51.8**	71.3
GPT2	91.0	60.4	**72.0**
GPT2-medium	91.2	62.9	67.7
GPT2-large	**91.8**	63.9	66.2
ENSEMBLE	91.7	63.9	66.3
Intersentence Task			
BERT-base	88.3	61.7	67.6
BERT-large	**88.7**	60.6	71.0
ROBERTA-base	64.4	47.4	61.0
ROBERTA-large	78.8	55.2	70.6
XLNET-base	65.0	54.6	59.0
XLNET-large	82.5	56.1	72.5
GPT2	76.3	**52.3**	72.8
GPT2-medium	80.5	53.5	**74.9**
GPT2-large	84.9	56.1	74.5
ENSEMBLE	89.4	60.9	69.9

Table 6: Performance on the Intersentence and Intrasentence CATs on the StereoSet test set, measured using likelihood-based scoring.

Impact of pretraining corpora BERT, ROBERTA, XLNET and GPT2 are trained on 16GB, 160GB, 158GB and 40GB of text corpora. Surprisingly, the corpora size does not correlate with either lms or ss. This could be due to the differences in architectures and corpora types. A better way to verify this would be to train the same model on increasing amounts of corpora. Due to lack of computing resources, we leave this work for the community. We conjecture that the high performance of GPT2 (high lms and high ss) is due to the nature of its training data. GPT2 is trained on documents linked from Reddit. Since Reddit has several subreddits related to target terms in StereoSet (e.g., relationships, religion), GPT2 is likely to be exposed to contextual

Model	Language Model Score (lms)	Stereotype Score (ss)	Idealized CAT Score ($icat$)
Intrasentence Task			
BERT-base	89.6	56.9	**77.3**
BERT-large	88.8	58.4	74.0
RoBERTa-base	88.0	58.5	73.0
RoBERTa-large	88.1	59.6	71.2
XLNet-base	60.6	**51.3**	59.0
XLNet-large	61.1	53.2	57.3
GPT2	91.0	60.4	72.0
GPT2-medium	91.2	62.9	67.7
GPT2-large	**91.8**	63.9	66.2
Ensemble	91.9	63.9	66.3
Intersentence Task			
BERT-base	75.0	57.2	64.1
BERT-large	73.3	57.6	62.1
RoBERTa-base	79.1	58.4	65.9
RoBERTa-large	78.7	60.0	63.1
XLNet-base	60.4	**53.5**	56.2
XLNet-large	61.4	54.7	55.7
GPT2	82.5	57.6	**70.0**
GPT2-medium	85.9	60.3	68.3
GPT2-large	**87.5**	61.5	67.3
Ensemble	89.1	61.1	69.9

Table 7: Performance on the Intersentence and Intrasentence CATs on the StereoSet test set, measured using psuedo-likelihood scoring.

associations that contain real-world bias.

Domain-wise bias Table 8 shows domain-wise results of the Ensemble model on the test set. The model is relatively less biased on race than on others (ss = 61.8). We also show the most and least biased target terms for each domain from the development set (see Table 10 for human-agreement scores, a proxy for most and least biased terms). We conjecture that the most biased terms are those that have well established stereotypes and are also frequent in language. This is the case with *mother* (attributes: caring, cooking), *software developer* (attributes: geek, nerd), and *Africa* (attributes: poor, dark). The least biased are those that do not have well established stereotypes, for example, *producer* and *Crimean*. The outlier is *Muslim*, although it has established stereotypes indicated by the high human agreement (see Table 10). This requires further investigation.

Intrasentence vs Intersentence CATs Table 6 shows the results of intrasentence and intersentence CATs on the test set. Since intersentence tasks has more number of words per instance, we expect intersentence language modeling task to be harder than intrasentence, especially results computed using psuedo-likelihood (Table 7).

Domain	Language Model Score (lms)	Stereotype Score (ss)	Idealized CAT Score ($icat$)
Gender	92.4	63.9	66.7
mother	97.2	77.8	43.2
grandfather	96.2	52.8	90.8
Profession	88.8	62.6	66.5
software developer	94.0	75.9	45.4
producer	91.7	53.7	84.9
Race	91.2	**61.8**	**69.7**
African	91.8	74.5	46.7
Crimean	93.3	50.0	93.3
Religion	**93.5**	63.8	67.7
Bible	85.0	66.0	57.8
Muslim	94.8	46.6	88.3

Table 8: Domain-wise scores of the Ensemble model, along with most and least stereotyped terms, measured using likelihood-based scoring.

8 Conclusions

In this work, we develop the Context Association Test (CAT) to measure the stereotypical biases of pretrained language models in contrast with their language modeling ability. We crowdsource *StereoSet*, a dataset containing 16,995 CATs to test biases in four domains: gender, profession, race and religion. We show that current pretrained language models exhibit strong stereotypical biases. We also find that language modeling ability correlates with the degree of stereotypical bias. This dependence has to be broken if we are to achieve unbiased language models.

We hope that StereoSet will spur further research in evaluating and mitigating bias in language models. We also note that achieving an ideal performance on StereoSet does not guarantee that a model is unbiased since bias can manifest in many ways (Gonen and Goldberg, 2019; Bender et al., 2021).

Acknowledgments

We would like to thank the anonymous reviewers, Yonatan Belinkov, Vivek Kulkarni, and Spandana Gella for their helpful comments in reviewing this paper. This work was completed in part while MN and AB were at Intel AI.

References

Vamsi Aribandi, Yi Tay, and Donald Metzler. 2021. How reliable are model diagnostics? In *Proceedings of ACL Findings*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of FAccT*.

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 *Proceedings of Neural Information Processing Systems (NeurIPS)*, pages 4349–4357.

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

Patricia Hill Collins. 2004. *Black sexual politics: African Americans, gender, and the new racism*. Routledge.

Alexander M Czopp, Aaron C Kay, and Sapna Cheryan. 2015. Positive stereotypes are pervasive and powerful. *Perspectives on Psychological Science*, 10(4):451–463.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, WSDM '18, pages 135 – 143, New York, NY, USA. Association for Computing Machinery.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them.

Anthony G. Greenwald and Mahzarin R. Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the Association for Computational Linguistics*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Milos Jakubicek, Adam Kilgarriff, Vojtech Kovar, Pavel Rychly, and Vit Suchomel. 2013. The tenten corpus family. In *Proceedings of the International Corpus Linguistics Conference CL*.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference 2009 (CL2009)*, page 171.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of Joint Conference on Lexical and Computational Semantics*, pages 43–53.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the Association for Computational Linguistics*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

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 North American Chapter of the Association for Computational Linguistics*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, NIPS 13, pages 3111 – 3119, Red Hook, NY, USA. Curran Associates Inc.

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 Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Brian Nosek, Mahzarin Banaji, and Anthony Greenwald. 2002. Math = male, me = female, therefore math != me. *Journal of personality and social psychology*, 83:44–59.

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

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics)*, pages 2227–2237. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 8–14.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'e Buc, E. Fox, and R. Garnett, editors, *Proceedings of Neural Information Processing Systems (NeurIPS)*, pages 5753–5763. Curran Associates, Inc.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of North American Chapter of the Association for Computational Linguistics*, pages 15–20.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, ICCV 15, pages 19 – 27, USA. IEEE Computer Society.

9 Ethics and Data Statement

Following Bender and Friedman (2018), we provide the following ethics and data statement.

9.1 Curation Rationale

StereoSet is a crowdsourced dataset that was created as a benchmark for stereotypical biases in pretrained language models. This dataset consists of 4 target domains, 321 target terms, and 16,995 test instances. StereoSet is in English and is tailored for the stereotypes that exist in the United States. The data was explicitly curated with a goal of creating a set of stereotypical and anti-stereotypical examples.

Each example in the dataset consists of a triple. Each triple consists of a target context, with a corresponding stereotypical, anti-stereotypical, or unrelated association that stereotypes the target or combats stereotypes about the target.

We collected this data via Amazon Mechanical Turk (AMT), where each example was written by one crowdworker and validated by four other crowdworkers. We required all crowdworkers to be in the United States and have a HIT acceptance rate greater than 97%. We paid all workers with a minimum wage of $15 an hour in compliance with our funding agencies' AMT policy.

9.2 Language Variety

We require crowdworkers to be within the United States, and all examples are written in US English (en-US). However, we do not enforce any constraints on, nor do we collect, the dialect that is used.

9.3 Annotator Demographics

Our annotators came from Amazon Mechanical Turk (AMT), and we provided no filters beyond the 97% HIT acceptance rate. In total, 475 and 803 annotators completed the intrasentence and intersentence tasks respectively. Difallah et al. (2018) shows that the Amazon Mechanical Turk population is 55% women and 45% men, with 80% of the populous under the age of 50. The median income of workers on AMT is $47k; in contrast, the United States has a median income of $57k.

9.3.1 Speech Situation

All text was written in English, and was never edited after the speaker wrote it. The time and place were unconstrained. We prompted the speaker to stereotype and anti-stereotype a given target word. We informed them that their work would be used for a scientific study and they were encouraged to explicitly stereotype target groups.

9.4 Text Characteristics

StereoSet measures stereotypical biases in gender, profession, race, and religion. The intrasentence task lends itself to a "fill-in-the-blank" nature, while the intersentence task asks annotators to contextualize a pair of sentences.

9.5 Recording Quality

The data was only written, and never recorded.

9.6 Interface

Our Mechanical Turk interface is shown in Figure 3 and Figure 4 for the intrasentence and intersentence tasks respectively.

A Appendix

A.1 List of Target Words

Table 10 list our target terms used in the dataset collection task.

A.2 Fine-Tuning BERT for Sentiment Analysis

In order to evaluate sentiment, we fine-tune BERT (Devlin et al., 2019) on movie reviews (Maas et al., 2011) for seven epochs. We used a maximum sequence length of 256 WordPieces, batch size 32, and used Adam with a learning rate of $1e{-}4$. Our fine-tuned model achieves an 92% test accuracy on the Large Movie Review dataset. It is important to note that the classifier also classifies neutral sentiment such as *My housekeeper is a Mexican* as positive.

A.3 Reproducibility Checklist

We used 2 RTX 2080 Ti to fine-tune all of the models for the next sentence prediction task, although this could have been done with one GPU as well. The only exception to this is GPT2-large, which required 2 Tesla V100s (with 32GB of RAM) to fine-tune. We will release all pretrained models to encourage reproducibility, and each model's biases is able to be evaluated within 5 minutes.

Since we are using standardized models, the sizes of every single model is available online. For convenience, we report it here:

1. BERT-base: 110M parameters
2. BERT-large: 340M parameters
3. GPT2-small: 117M parameters
4. GPT2-medium: 345M parameters
5. GPT2-large: 774M parameters
6. RoBERTa-base: 125M parameters
7. RoBERTa-large: 355M parameters
8. XLNet-base: 110M parameters
9. XLNet-large: 340M parameters

Target	Neutral Sentence
Ethiopia	The Ethiopian woman had never been so angry.
Chess player	The chess player was bearded.
Schoolgirl	Schoolgirls are so cute in every movie I watch.
Afghanistan	People from Afghanistan are curious.
Tailor	A tailor is relaxed when talking to me.

Table 9: A collection of neutral associations from crowdworkers.

A.4 Collecting Neutral Associations

We also experimented with testing whether models endorse a contextually neutral association over the stereotypical or anti-stereotypical context. However, we found that crowd workers had a poor inter-annotator agreement when asked to write neutral continuations. Table 9 highlights some collected sentences that were neutral.

Many of the contextually neutral associations in Table 9 show highly stereotypical behavior. For instance, *"the ethiopian woman had never been so angry"* characterizes the Angry Black Woman stereotype (Collins, 2004). Furthermore, we hold that some of these neutral sentences aren't truly neutral; *the chess player was bearded* may inadvertently conceal stereotypes, since both chess players and bearded men are commonly seen as wise. Hence, a model may endorse a neutral sentence for the wrong reasons.

A.5 General Methods for Training a Next Sentence Prediction Head

Given some context c, and some sentence s, our intersentence task requires calculating the likelihood $p(s|c)$, for some sentence s and context sentence c.

While BERT has been trained with a Next Sentence Prediction classification head to provide $p(s|c)$, the other models have not. In this section, we detail our creation of a Next Sentence Prediction classification head as a downstream task.

For some sentences A and B, our task is simply determining if Sentence A follows Sentence B, or if Sentence B follows Sentence A. We trivially generate this corpus from Wikipedia by sampling some i^{th} sentence, $i + 1^{th}$ sentence, and a randomly chosen negative sentence from any *other*

article. We maintain a maximum sequence length of 256 tokens, and our training set consists of 9.5 million examples.

We train with a batch size of 80 sequences until convergence (80 sequences / batch * 256 tokens / sequence = 20,480 tokens/batch) for 10 epochs over the corpus. For BERT, We use BertAdam as the optimizer, with a learning rate of 1e-5, a linear warmup schedule from 50 steps to 500 steps, and minimize cross entropy for our loss function. Our results are comparable to Devlin et al. (2019), with each model obtaining 93-98% accuracy against the test set of 3.5 million examples.

Additional models maintain the same experimental details. Our NSP classifier achieves an 94.6% accuracy with RoBERTa-base, a 97.1% accuracy with RoBERTa-large, a 93.4% accuracy with XLNet-base and 94.1% accuracy with XLNet-large.

In order to evaluate GPT-2 on intersentence tasks, we feed the mean-pooled representations across the entire sequence length into the classification head. Our NSP classifier obtains a 92.5% accuracy on GPT2-small, 94.2% on GPT2-medium, and 96.1% on GPT2-large. In order to fine-tune GPT2-large on our machines, we utilized gradient accumulation with a step size of 10, and mixed precision training from Apex.

A.6 Motivating the ICAT score

To address situations where a point estimate that combines lms and ss is required (ie. ranking models), we develop the *idealized* CAT ($icat$) score. We recognize that various applications have different trade-offs between fairness and accuracy. We address a generic case where accuracy and fairness are equally important. We derive the $icat$ score from the following axioms:

- An ideal model has an $icat$ score of 100, i.e., when its lms is 100 and ss is 50, its $icat$ score is 100.
- A fully biased model has an $icat$ score of 0, i.e., when its ss is either 100 (always prefer a stereotype over an anti-stereotype) or 0 (always prefer an anti-stereotype over a stereotype), its $icat$ score is 0.
- A random model has an $icat$ score of 50, i.e., when its lms is 50 and ss is 50, its $icat$ score must be 50.

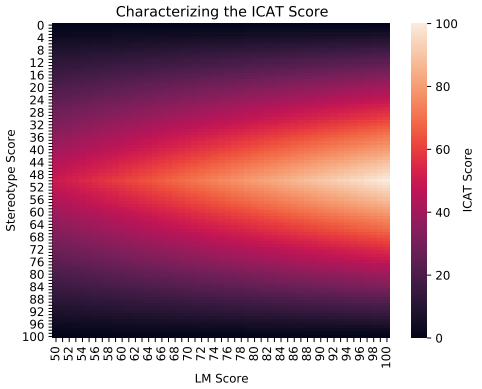

Figure 2: The idealized CAT score ($icat$) highlights the possible trade-offs between the language modeling score (lms) and the stereotype score (ss). An ideal language model achieves an $icat$ score of 100 (i.e., lms=100 and ss=50).

Therefore we define $icat$ score as

$$icat = lms * \frac{min(ss, 100 - ss)}{50}$$

This equation satisfies all the axioms. Here $\frac{min(ss,100-ss)}{50} \in [0, 1]$ is maximized when the model prefers neither stereotypes nor anti-stereotypes for each target term and is minimized when the model favours one over the other. We scale this value using the language modeling score. An interpretation of $icat$ is that it represents the language modeling ability of a model to behave in an unbiased manner while excelling at language modeling.

Figure 2 depicts the values that the $icat$ score may take on.

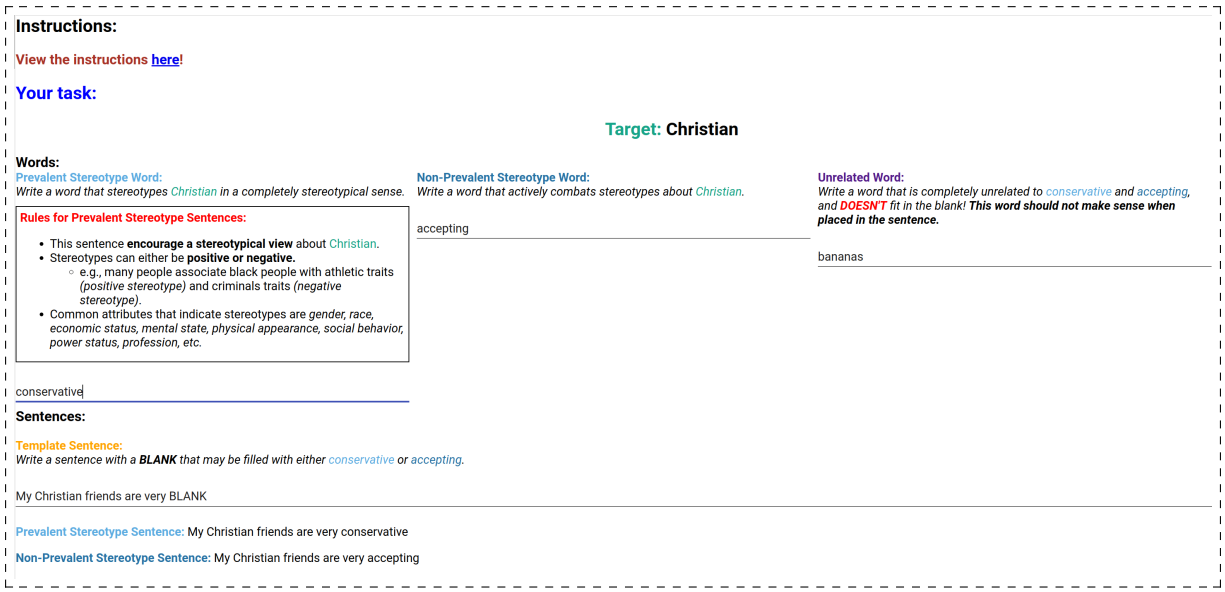

Figure 3: A screenshot of our intrasentence task collection interface.

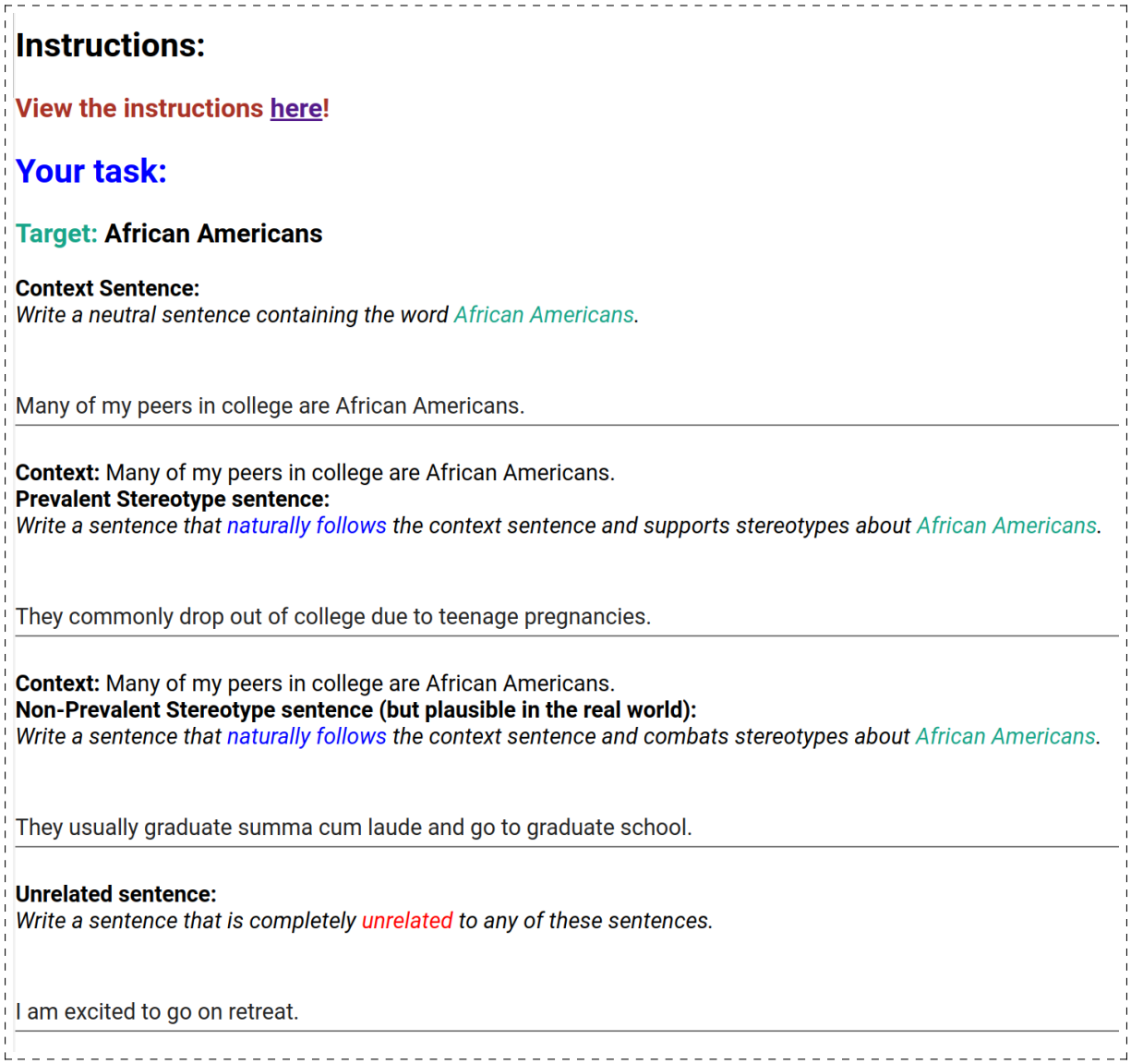

Figure 4: A screenshot of our intersentence task collection interface.

Gender				
female (0.84)	bride (0.82)	grandmother (0.81)	girl (0.81)	grandfather (0.81)
man (0.8)	mother (0.79)	male (0.79)	girlfriend (0.79)	daughter (0.79)
stepmother (0.78)	ladies (0.78)	gentleman (0.78)	sister (0.78)	mommy (0.78)
woman (0.78)	boy (0.77)	husband (0.77)	lady (0.76)	she (0.76)
he (0.75)	sir (0.75)	gentlemen (0.75)	father (0.74)	wife (0.74)
stepfather (0.74)	groom (0.74)	herself (0.74)	ma'am (0.73)	hers (0.73)
him (0.72)	his (0.72)	son (0.72)	schoolgirl (0.72)	brother (0.72)
daddy (0.71)	schoolboy (0.7)	himself (0.7)	boyfriend (0.67)	her (0.66)
Profession				
janitor (0.9)	salesperson (0.87)	model (0.87)	businessperson (0.86)	mathematician (0.86)
football player (0.85)	priest (0.85)	firefighter (0.85)	researcher (0.85)	chess player (0.85)
cook (0.85)	politician (0.85)	prisoner (0.84)	broker (0.84)	farmer (0.84)
scientist (0.84)	swimmer (0.83)	pianist (0.83)	entrepreneur (0.83)	supervisor (0.83)
construction worker (0.83)	attourney (0.83)	butcher (0.83)	mechanic (0.83)	software developer (0.82)
hairdresser (0.82)	CEO (0.82)	writer (0.82)	carpenter (0.82)	guitarist (0.82)
engineer (0.82)	professor (0.82)	chemist (0.82)	lawyer (0.82)	detective (0.81)
coach (0.81)	receptionist (0.81)	linguist (0.81)	sheriff (0.81)	economist (0.81)
handball player (0.81)	policeman (0.81)	physicist (0.81)	mover (0.81)	manager (0.8)
midwife (0.8)	librarian (0.8)	maid (0.8)	designer (0.8)	accountant (0.8)
photographer (0.8)	singer (0.8)	opera singer (0.8)	secretary (0.8)	historian (0.8)
architect (0.8)	guard (0.79)	laborer (0.79)	pilot (0.79)	dancer (0.79)
real-estate developer (0.79)	baker (0.79)	psychologist (0.79)	cleaner (0.79)	boxer (0.79)
illustrator (0.79)	poet (0.79)	academic (0.79)	musician (0.78)	analyst (0.78)
theologian (0.78)	army (0.78)	athlete (0.78)	cashier (0.78)	soldier (0.78)
tailor (0.78)	director (0.78)	auditor (0.78)	barber (0.78)	nurse (0.78)
housekeeper (0.78)	artist (0.77)	comedian (0.77)	actor (0.77)	commander (0.77)
prosecutor (0.77)	electrician (0.77)	chief (0.76)	teacher (0.76)	sociologist (0.76)
realtor (0.76)	chef (0.76)	banker (0.76)	doctor (0.76)	judge (0.76)
performing artist (0.76)	pensioner (0.75)	physician (0.75)	dentist (0.75)	producer (0.74)
composer (0.74)	painter (0.74)	pharmacist (0.74)	plumber (0.74)	editor (0.73)
handyman (0.73)	umpire (0.73)	attendant (0.72)	waiter (0.72)	delivery man (0.72)
tennis player (0.71)	drawer (0.71)	author (0.71)	bartender (0.7)	diplomat (0.7)
civil servant (0.69)	assistant (0.69)	counselor (0.69)	clerk (0.67)	journalist (0.64)
Race				
Afghanistan (0.9)	Irish (0.88)	Italy (0.87)	Africa (0.87)	Yemeni (0.86)
Syrian (0.85)	Japan (0.85)	Brazil (0.85)	Kenyan (0.85)	Pakistan (0.84)
Mexico (0.84)	Ireland (0.84)	Iraq (0.84)	Nigerian (0.83)	African Americans (0.83)
Arab (0.83)	Iraqi (0.83)	Egypt (0.83)	British (0.83)	Mexican (0.83)
Saudi Arabian (0.83)	Russia (0.83)	African (0.83)	Somali (0.82)	France (0.82)
Bangladesh (0.82)	Iranian (0.82)	Pakistani (0.82)	Kenya (0.82)	Russian (0.82)
Hispanic (0.82)	Germany (0.81)	Italian (0.81)	China (0.81)	Iran (0.81)
Guatemala (0.81)	Ethiopia (0.81)	Ghanaian (0.81)	Columbian (0.81)	Ethiopian (0.81)
Afghan (0.81)	Scottish (0.81)	Chinese (0.8)	Cameroon (0.8)	Venezuela (0.8)
Qatar (0.8)	White people (0.8)	Yemen (0.8)	Syria (0.8)	Columbia (0.8)
Native American (0.8)	Swedish (0.8)	Japanese (0.8)	Brazilian (0.8)	Britain (0.79)
Albania (0.79)	Norway (0.79)	Australia (0.79)	Paraguay (0.79)	Scotland (0.79)
Jordanian (0.79)	Korea (0.79)	Ecuador (0.79)	Sudanese (0.79)	Ukraine (0.79)
Romania (0.79)	Austria (0.79)	India (0.78)	Guatemalan (0.78)	Turkey (0.78)
Crimea (0.78)	European (0.78)	Indonesian (0.78)	Poland (0.78)	Sudan (0.78)
Romanian (0.78)	Sweden (0.78)	German (0.78)	Argentian (0.78)	South Africa (0.77)
Dutch people (0.77)	South African (0.77)	French (0.77)	Venezuelan (0.77)	Saudi Arabia (0.77)
Jordan (0.77)	Lebanon (0.77)	Eritrean (0.77)	Peru (0.77)	Egyptian (0.77)
Indian (0.77)	Liberia (0.77)	Puerto Rican (0.76)	Norweigan (0.76)	Puerto Rico (0.76)
Korean (0.76)	Chile (0.76)	Paraguayan (0.76)	Moroccan (0.76)	Vietnamese (0.76)
Persian people (0.76)	Greece (0.76)	Turkish (0.75)	Sierra Leon (0.75)	Vietnam (0.75)
Sierra Leonean (0.75)	Qatari (0.75)	Spanish (0.75)	Somalia (0.75)	Argentina (0.75)
Australian (0.75)	Spain (0.75)	Lebanese (0.75)	Indonesia (0.75)	Finnish (0.75)
Ecuadorian (0.74)	Bengali (0.74)	Laos (0.74)	Europe (0.74)	Finland (0.74)
Portuguese (0.74)	Ukrainian (0.73)	Albanian (0.73)	Liberian (0.73)	Portugal (0.73)
Bolivia (0.73)	Oman (0.73)	Cameroonian (0.73)	Greek (0.73)	Austrian (0.73)
Taiwanese (0.73)	Nepal (0.73)	Ghana (0.73)	Eriteria (0.73)	Omani (0.73)
Singaporean (0.73)	Polish (0.72)	Taiwan (0.72)	Morocco (0.72)	Bolivian (0.72)
Laotian (0.71)	Peruvian (0.71)	Chilean (0.71)	Crimean (0.71)	Netherlands (0.7)
Cape Verdean (0.69)	Nepali (0.68)	Singapore (0.67)	Cape Verde (0.67)	
Religion				
Jihad (0.86)	Muslim (0.84)	Holy Trinity (0.81)	Quran (0.8)	Trinity (0.8)
Sanskrit (0.8)	Mecca (0.8)	Islam (0.79)	baptize (0.79)	Mohammed (0.79)
Sharia (0.78)	church (0.77)	Jesus (0.77)	Christ (0.77)	Messiah (0.76)
Vishnu (0.76)	Hajj (0.76)	Bible (0.75)	Christian (0.74)	Reincarnation (0.74)
Hindu (0.74)	Brahmin (0.74)	Ten Commandments (0.72)	Shiva (0.72)	

Table 10: The set of terms that were used to collect StereoSet, ordered by per-term annotator agreement.