Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases

Wei Guo
George Washington University
Department of Computer Science
weiguo@gwu.edu

Aylin Caliskan
George Washington University
Department of Computer Science
Institute for Data, Democracy & Politics
aylin@gwu.edu

ABSTRACT

With the starting point that implicit human biases are reflected in the statistical regularities of language, it is possible to measure biases in English static word embeddings. State-of-the-art neural language models generate dynamic word embeddings dependent on the context in which the word appears. Current methods measure pre-defined social and intersectional biases that occur in contexts defined by sentence templates. Dispensing with templates, we introduce the Contextualized Embedding Association Test (CEAT), that can summarize the magnitude of overall bias in neural language models by incorporating a random-effects model. Experiments on social and intersectional biases show that CEAT finds evidence of all tested biases and provides comprehensive information on the variance of effect magnitudes of the same bias in different contexts. All the models trained on English corpora that we study contain biased representations. GPT-2 contains the smallest magnitude of overall bias followed by GPT, BERT, and then ELMo, negatively correlating with the contextualization levels of the models.

Furthermore, we develop two methods, Intersectional Bias Detection (IBD) and Emergent Intersectional Bias Detection (EIBD), to automatically identify the intersectional biases and emergent intersectional biases from static word embeddings in addition to measuring them in contextualized word embeddings. We present the first algorithmic bias detection findings on how intersectional group members are strongly associated with unique emergent biases that do not overlap with the biases of their constituent minority identities. IBD achieves an accuracy of 81.6% and 82.7%, respectively, when detecting the intersectional biases of African American females and Mexican American females, where the random correct identification rates are 14.3% and 13.3%. EIBD reaches an accuracy of 84.7% and 65.3%, respectively, when detecting the emergent intersectional biases unique to African American females and Mexican American females, where the random correct identification rates are 9.2% and 6.1%. Our results indicate that intersectional biases associated with members of multiple minority groups, such as African American females and Mexican American females, have the highest magnitude across all neural language models.

AIES '21, May 19–21, 2021, Virtual Event, USA.
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8473-5/21/05.
https://doi.org/10.1145/3461702.3462536

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Learning latent representations**; • **General and reference** → **Metrics**; **Evaluation**; • **Applied computing** → **Psychology**.

KEYWORDS

AI ethics; bias; intersectionality; language models; social psychology; word embeddings

ACM Reference Format:
Wei Guo and Aylin Caliskan. 2021. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3461702.3462536

1 INTRODUCTION

State-of-the-art off-the-shelf neural language models such as the multi-million dollar GPT-3, associates men with competency and occupations demonstrating higher levels of education, in downstream natural language processing (NLP) tasks such as sequence prediction [8]. When GPT-3's user interface for academic access is prompted for language generation with the input "What is the gender of a doctor," the first answer is "A: Doctor is a masculine noun;" whereas when prompted with "What is the gender of a nurse," the first answer is "It's female." Propagation of social group bias in NLP applications such as automated resume screening, that shapes the workforce by making consequential decisions about job candidates, would not only perpetuate existing biases but potentially exacerbate harmful bias in society to affect future generations [16, 55]. To enhance transparency in NLP, we use the representations of words learned from word co-occurrence statistics to discover social biases. Our methods uncover unique intersectional biases associated with individuals that are members of multiple minority groups. After identifying these emergent biases, we use numeric representations of words that vary according to neighboring words to analyze how prominent bias is in different contexts. Recent work has shown that human-like biases are embedded in the statistical regularities of language that are learned by word representations, namely word embeddings [4, 11, 13]. We build a method on this work to automatically identify intersectional biases, such as the ones associated with African American and Mexican American women from static word embeddings (SWE). Then, we measure how human-like biases manifest themselves in contextualized word embeddings (CWE), which are dynamic word representations generated by neural language models that adapt to their context.

Artificial intelligence (AI) systems are known not only to perpetuate social biases, but they may also amplify existing cultural assumptions and inequalities [12]. While most work on biases in word embeddings focuses on a single social category (e.g., gender, race) [6, 11, 26, 28, 71], the lack of work on identifying intersectional biases, the bias associated with populations defined by multiple categories [10], leads to an incomplete measurement of social biases [32, 40]. For example, Caliskan et al. [11]'s Word Embedding Association Test (WEAT) quantifies biases documented by the validated psychological methodology of the Implicit Association Test (IAT) [30, 31]. The IAT provides the sets of words to represent social groups and attributes to be used while measuring bias. Consequently, the analysis of bias via WEAT is limited to the types of IATs and their corresponding words contributed by the IAT literature, which happens to include intersectional representation for only African American women. To overcome these constraints of WEATs, we extend WEAT to automatically identify attributes associated with individuals that are members of more than one social group. While this allows us to discover emergent intersectional biases, it is also a promising step towards automatically identifying all biased associations embedded in the regularities of language. To fill the gap in understanding the complex nature of intersectional bias, we develop a method called Intersectional Bias Detection (IBD) to automatically identify intersectional biases without relying on pre-defined attribute sets from the IAT literature.

Biases associated with intersectional group members contain emergent elements that do not overlap with the biases of their constituent minority identities [1, 27]. For example, "hair weaves" is stereotypically associated with African American females but not with African Americans or females. We extend IBD and introduce a method called Emergent Intersectional Bias Detection (EIBD) to identify the emergent intersectional biases of an intersectional group in SWE. Then, we construct statistical tests to quantify these intersectional and emergent biases in CWE. To investigate the influence of different linguistic contexts on bias, we use a fill-in-the-blank task called masked language modeling. The goal of the task is to generate the most probable substitution for the [MASK] that is surrounded with neighboring context words in a given sentence. BERT, a widely used language model trained on this task, substitutes [MASK] in "Men/women *excel* in [MASK]." with "science" and "sports", reflecting stereotype-congruent associations. However, when we feed in similar contexts "The man/woman is *known* for his/her [MASK]," BERT fills "wit" in both sentences, which indicates gender bias may not appear in these contexts. Prior methods use templates analogous to masked language modeling to measure bias in CWE [43, 45, 61]. The templates are designed to substitute words from WEAT's sets of target words and attributes in a simple manner such as "This is [TARGET]" or "[TARGET] is a [ATTRIBUTE]".In this work, we propose the Contextualized Embedding Association Test (CEAT), a test eschewing templates and instead generating the distribution of effect magnitudes of biases in different contexts from a control corpus. To comprehensively measure the social and intersectional biases in this distribution, a random-effects model designed to combine effect sizes of similar bias interventions summarizes the overall effect size of bias in the neural language model [17]. As a result, instead of focusing on biases in template-based contexts, CEAT measures the distribution of biased associations in a language model.

Contributions. In summary, this paper presents three novel contributions along with three complementary methods (CEAT, IBD, and EIBD) to automatically identify intersectional biases as well as emergent intersectional biases in SWE, then use these findings to measure all available types of social biases in CWE. We find that ELMo is the most biased, followed by BERT, then GPT, with GPT-2 being the least biased. The overall level of bias correlates with how contextualized the CWE generated by the models are. Our results indicate that the strongest biased associations are embedded in the representations of intersectional group members such as African American women. Data and source code are available at https://github.com/weiguowilliam/CEAT.

Intersectional Bias Detection (IBD). We develop a novel method for SWE to detect words that represent biases associated with intersectional group members. To our knowledge, IBD is the first algorithmic method to automatically identify individual words that are strongly associated with intersectionality. IBD reaches an accuracy of 81.6% and 82.7%, respectively, when evaluated on intersectional biases associated with African American females and Mexican American females that are provided in Ghavami and Peplau [27]'s validation dataset. In these machine learning settings, the random chances of correct identification are 14.3% and 13.3%. Currently, the validation datasets represent gender as a binary label. Consequently, our method uses binary categorization when evaluating for gender related biases. However, we stress that our method generalizes to multiple categories from binary. In future work, we aim to design non-categorical methods that don't represent individuals as members of discrete categories compared to potentially using continuous representations. Accordingly, we also plan to compile validation datasets that won't constrain our evaluation to categorical assumptions about humans.

Emergent Intersectional Bias Detection (EIBD). We contribute a novel method to identify emergent intersectional biases that do not overlap with biases of constituent social groups in SWE. To our knowledge, EIBD is the first algorithmic method to detect the emergent intersectional biases in word embeddings automatically. EIBD reaches an accuracy of 84.7% and 65.3%, respectively, when validating on the emergent intersectional biases of African American females and Mexican American females that are provided provided in Ghavami and Peplau [27]'s validation dataset. In these machine learning settings, the random chances of correct identification are 9.2% and 6.1%.

Contextualized Embedding Association Test (CEAT). WEAT measures human-like biases in SWE. We extend WEAT to the dynamic setting of neural language models to quantify the distribution of effect magnitudes of social and intersectional biases in *contextualized* word embeddings and summarize the combined magnitude of bias by pooling effect sizes with the validated random-effects methodology [7, 36]. We show that the magnitude of bias greatly varies according to the context in which the stimuli of WEAT appear. Overall, the pooled mean effect size is statistically significant in all CEAT tests including intersectional bias measurements and all models contain biased representations.

2 RELATED WORK

SWE are trained on word co-occurrence statistics of corpora to generate numeric representations of words so that machines can process language [46, 51]. Previous work on bias in SWE has shown that human-like biases that have been documented by the IAT are embedded in the statistical regularities of language [11]. The IAT [30] is a widely used measure of implicit bias in human subjects that quantifies the differential reaction time to pairing two concepts. Analogous to the IAT, Caliskan et al. [11] developed the WEAT to measure the biases in SWE by quantifying the relative associations of two sets of target words (e.g., African American and European American) that represent social groups with two sets of polar attributes (e.g., pleasant and unpleasant). WEAT computes an effect size (Cohen's d) that is a standardized bias score and its p-value based on a one-sided permutation test. WEAT measures biases predefined by the IAT such as racism, sexism, ableism, and attitude towards the elderly, as well as widely shared non-discriminatory non-social group associations. Swinger et al. [60] presented an adaptation of the WEAT to identify biases associated with clusters of names.

Regarding the biases of intersectional groups categorized by multiple social categories, there is prior work in the social sciences focusing on the experiences of African American females [15, 33, 42, 63]. Buolamwini et al. demonstrated intersectional accuracy disparities in commercial gender classification in computer vision [9]. May et al. [45] and Tan and Celis [61] used the attributes presented in Caliskan et al. [11] to measure emergent intersectional biases of African American females in CWE. We develop the first algorithmic method to automatically identify intersectional bias and emergent bias attributes in SWE, which can be measured in both SWE and CWE. Furthermore, we construct new embedding association tests for the intersectional groups. As a result, our work is the first to discuss biases regarding Mexican American females in word embeddings. Ghavami and Peplau [27] used a free-response procedure in human subjects to collect words that represent intersectional biases. They show that emergent intersectional biases exist in several gender-by-race groups in the U.S. We use the validation dataset constructed by Ghavami and Peplau [27] to evaluate our methods.

Recently, neural language models, which use neural networks to assign probability values to sequences of words, have achieved state-of-the-art results in NLP tasks with their dynamic word representations, CWE [5, 20, 69]. Neural language models typically consist of an encoder that generates CWE for each word based on its accompanying context in the input sequence. Specifically, the collection of values on a particular layer's hidden units forms the CWE [62], which has the same vector shape as a SWE. However, unlike SWE that represent each word, including polysemous words, with a fixed vector, CWE of the same word vary according to its context window that is encoded into its representation by the neural language model. Ethayarajh et al. [22] demonstrate how these limitations of SWE impact measuring gender biases. With the wide adaption of neural language models [5, 20, 69], human-like biases were observed in CWE [43, 45, 61, 70]. To measure human-like biases in CWE, May et al. [45] applied the WEAT to contextualized representations in template sentences. Tan and Celis [61] adopted the method of May et al. [45] by applying Caliskan et al. [11]'s WEAT to the CWE of the stimuli tokens in templates such as "This is a [TARGET]". Kurita et al. [43] measured biases in BERT based on the prediction probability of the attribute in a template that contains the target and masks the attribute, e.g., [TARGET] is [MASK]. Hutchinson et al. [41] reveal biases associated with disabilities in CWE and demonstrate undesirable biases towards mentions of disability in applications such as toxicity prediction and sentiment analysis.

Nadeem et al. [47] present a large-scale natural language dataset in English to measure stereotypical biases in the domains of gender, profession, race, and religion. Their strategy cannot be directly compared to ours since it is not aligned with our intersectional bias detection method, which is complementary to CEAT. The majority of prior work measures bias in a limited selection of contexts to report the unweighted mean value of bias magnitudes, which does not reflect the scope of contextualization of biases embedded in a neural language model.

3 DATA

Identifying and measuring intersectional and social biases in word embeddings as well as neural language models requires four types of data sources that are detailed in this section. (1) SWE carry the signals for individual words that have statistically significant biased associations with social groups and intersectionality. Application of our methods IBD and EIBD to SWE automatically retrieves biased associations. (2) CWE extracted from sentence encodings of neural language models provide precise word representations that depend on the context of word occurrence. We apply CEAT to summarize magnitude of bias in neural language models. (3) A corpus provides the samples of sentences used in CEAT when measuring the overall bias and analyzing the variance of contexts in CWE of neural language models. (4) Stimuli designed by experts in social psychology represent validated concepts in natural language including social group and intersectional targets in addition to their corresponding attributes.

3.1 Static Word Embeddings (SWE)

We use GloVe [51] SWE trained on the word co-occurrence statistics of the Common Crawl corpus to automatically detect words that are highly associated with intersectional group members. The Common Crawl corpus consists of 840 billion tokens and more than 2 million unique vocabulary words collected from a crawl of the world wide web. Consequently, GloVe embeddings capture the language representation of the entire Internet population that contributed to its training corpus. GloVe embeddings learn fine-grained semantic and syntactic regularities [51]. Caliskan et al. [11] have shown that social biases are embedded in the linguistic regularities learned by GloVe.

3.2 Contextualized Word Embeddings (CWE)

We generate the CWE by widely used neural language model implementations of ELMo from https://allennlp.org/elmo, BERT, GPT and GPT-2 from https://huggingface.co/transformers/v2.5.0/model_doc/ [19, 52–54]. Specifically, CWE is formed by the collection of values on a particular layer's hidden units in the neural language model.

BERT, GPT, and GPT-2 use subword tokenization. Since GPT and GPT-2 are unidirectional language models, CWE of the last subtokens contain the information of the entire word [54]. We use the CWE of the last subtoken in the word as its representation in GPT and GPT-2. For consistency, we use the CWE of the last subtoken in the word as its representation in BERT. BERT and GPT-2 provide several versions. We use BERT-small-cased and GPT-2-117m trained on cased English text. The sizes of the training corpora detailed below have been verified from Aßenmacher and Heumann [2]. We obtained academic access to GPT-3's API which does not provide training data or the CWE. Accordingly, we are not able to systematically study GPT-3.

ELMo is a 2-layer bidirectional long short term memory (Bi-LSTM) [39] language model trained on the Billion Word Benchmark dataset [14] that takes up ~9GB memory. ELMo has 93.6 million parameters. It is different from the three other models since CWE in ELMo integrate the hidden states in all layers instead of using the hidden states of the top layer. We follow standard usage and compute the summation of hidden units over all aggregated layers of the same token as its CWE [52]. CWE of ELMo have 1,024 dimensions.

BERT [19] is a bidirectional transformer encoder [67] trained on a masked language model and next sentence prediction. BERT is trained on BookCorpus [72] and English Wikipedia dumps that take up ~16GB memory [3]. We use BERT-small-case with 12 layers that has 110 million parameters. We extract the values of hidden units on the top layer corresponding to the token as its CWE of 768 dimensions.

GPT [53] is a 12-layer transformer decoder trained on a unidirectional language model on BookCorpus that takes up ~13GB memory [72]. We use the values of hidden units on the top layer corresponding to the token as its CWE. This implementation of GPT has 110 million parameters. The CWE have 768 dimensions.

GPT-2 [54] is a transformer decoder trained on a unidirectional language model and is a scaled-up version of GPT. GPT-2 is trained on WebText that takes up ~40GB memory [54]. We use GPT-2-small which has 12 layers and 117 million parameters. We use the values of hidden units on the top layer corresponding to the token as its CWE. CWE of GPT-2 have 768 dimensions.

We provide the source code, detailed information, and documentation in our open source repository at https://github.com/weiguowilliam/CEAT.

3.3 Corpus

We need a comprehensive representation of all contexts a word can appear in naturally occurring sentences in order to investigate how bias associated with individual words varies across contexts. Identifying the potential contexts in which a word can be observed is not a trivial task. Consequently, we simulate the distribution of contexts a word appears in, by randomly sampling sentences that the word occurs in a large corpus.

Voigt et al. [68] have shown that social biases are projected into Reddit comments. Consequently, we use a Reddit corpus to generate the distribution of contexts that words of interest appear in. The corpus consists of 500 million comments made in the period between 1/1/2014 and 12/31/2014. We take all the stimuli used in Caliskan et al. [11]'s WEAT that measures effect size of bias for social groups and related attributes. For each WEAT type, we retrieve the sentences from the Reddit corpus that contain one of these stimuli. In this way, we collect a great variety of CWE from the Reddit corpus to measure bias comprehensively in a neural language model while simulating the natural distribution of contexts in language. For each stimulus, we generate CWE from 10,000 sentences pooled from the Reddit corpus. We discuss the justification of sampling 10,000 sentences in the upcoming sections.

3.4 Stimuli

Caliskan et al. [11]'s WEAT is inspired by the IAT literature [29–31] that measures implicit associations of concepts by representing them with stimuli. Experts in social psychology and cognitive science select stimuli which are words typically representative of various concepts. These linguistic or sometimes picture-based stimuli are proxies to overall representations of concepts in cognition. Similarly, in the word embedding space, WEAT uses these unambiguous stimuli as semantic representations to study biased associations related to these concepts. Since the stimuli are chosen by experts to most accurately represent concepts, they are not polysemous or ambiguous words. Each WEAT, designed to measure a certain type of association or social group bias, has at least 32 stimuli. There are at least 8 stimuli for each one of the four concepts. Each concept is constructed with at least 8 stimuli for statistically significant category representations. Two of these concepts represent target groups and two of them represent polar attributes. WEAT measures the magnitude of bias by quantifying the standardized differential association of targets with attributes. The larger the set of appropriate stimuli to represent a concept, the more statistically significant and accurate the representation becomes [11]. Similar to Caliskan et al. [11], we follow these principled methods and properties established in the IAT literature over more than two decades, for robust and accurate analyses.

Validation data for intersectional bias. To investigate intersectional bias with respect to race and gender, we represent members of social groups with target words provided by WEAT and Parada et al. [11, 50]. WEAT and Parada et al. represent racial categories with frequent given names that signal group membership. WEAT contains a balanced combination of common female and male names of African Americans and European Americans whereas Parada et al. presents the Mexican American names for women and men combined. The intersectional bias detection methods identify attributes that are associated with these target group representations. Human subjects provide the validation set of intersectional attributes with ground truth information in prior work [27]. The evaluation of intersectional bias detection methods uses this validation set. One limitation of these validation sets is the way they represent gender as a binary category. We will address this constraint in future work by constructing our own validation sets that won't have to represent people by discrete categorical labels of race and gender.

4 APPROACH

Our approach includes four components. (1) Caliskan et al. [11]'s WEAT for SWE is the foundation of our approach to summarizing overall bias in CWE generated by neural language models. (2)

Random-effects models from the meta analysis literature summarizes the combined effect size of bias for a neural language model's CWE via combining 10,000 WEAT samples by weighting each result with the within-WEAT and between-WEAT variances [36]. (3) Our novel method IBD automatically detects words associated with intersectional biases. (4) Our novel method EIBD automatically detects words that are uniquely associated with members of multiple minority or disadvantaged groups, but do not overlap with the biases of their constituent minority identities.

Our open source git repository includes the details of all the social and non-social bias types studied in this paper, namely, WEAT biases introduced by Caliskan et al. [11] as well as intersectional biases and their validation set introduced by Ghavami and Peplau [27] and Parada [50].

4.1 Word Embedding Association Test (WEAT)

WEAT, designed by Caliskan et al. [11], measures the effect size of bias in SWE, by quantifying the relative associations of two sets of target words (e.g., career, professional; and family, home) with two sets of polar attributes (e.g., woman, female; and man, male). Two of these WEATs measure non-social group baseline associations that are widely accepted such as the attitude towards flowers vs. insects or the attitude towards musical instruments vs. weapons. Both human subjects and word embeddings consistently associate flowers and musical instruments with pleasantness that corresponds to positive valence. However, human subjects and word embeddings associate insects and weapons with unpleasantness that corresponds to negative valence. Greenwald et al. [30] refers to these as universally accepted stereotypes since they are widely shared across human subjects and are not potentially harmful to society. However, the rest of the tests measure the magnitude of social-group associations, such as gender and race stereotypes and attitude towards the elderly or people with disabilities. Biased social-group associations in word embeddings can potentially be prejudiced and harmful to society. Especially, if downstream applications of NLP that use static or dynamic word embeddings to make consequential decisions about individuals, such as resume screening for job candidate selection, perpetuate existing biases to eventually exacerbate historical injustices [16, 55]. The formal definition of WEAT, the test statistic, and the statistical significance of biased associations are detailed in [11]'s.

4.2 Intersectional Bias Detection (IBD)

IBD identifies words associated with intersectional group members, defined by two social categories simultaneously. Our method automatically detects the attributes that have high associations with the intersectional group from a set of SWE. Analogous to the Word Embedding Factual Association Test (WEFAT) [11], we measure the standardized differential association of a single stimulus $w \in W$ with two social groups A and B using the following statistic.

$$s(w, A, B) = \frac{\text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\text{std-dev}_{x \in A \cup B} \cos(\vec{w}, \vec{x})}$$

We refer to the above statistic as the **association score**, which is used by WEFAT to verify that gender statistics are embedded in linguistic regularities. In [11], Concepts A and B are words that represent males (e.g., he, him) and females (e.g., she, her) and W is a set of occupations. For example, *nurse* has an association score $s(nurse, A, B)$ that measures effect size of gender associations. WEFAT has been shown to have high predictive validity ($\rho = 0.90$) in quantifying facts about the world [11].

We extend WEFAT's *gender* association measurement to quantify the relative association to other social categories (e.g., race), by automatically retrieving semantic fields and following an approach similar to lexicon induction that quantifies certain associations without annotating large-scale ground truth training data [35, 57, 66]. Let $P_i = (A_i, B_i)$ (e.g., African American and European American) be a pair of social groups, and W be a set of attribute words. We calculate the association score $s(w, A_i, B_i)$ for $w \in W$. If $s(w, A_i, B_i)$ is greater than the positive effect size threshold t, w is detected to be associated with group A_i. Let $W_i = \{w | s(w, A_i, B_i) > t, w \in W\}$ be the associated word list for each pair P_i.

We detect the biased attributes associated with an intersectional group C_{mn} defined by two social categories C_{1n}, C_{m1} with M and N subcategories $(C_{11}, \ldots, C_{mn})$ (e.g., African American females by race (C_{1n}) and gender (C_{m1})). We assume, there are three racial categories $M = 3$, and two gender categories $N = 2$ in our experiments because of the limited structure of representation for individuals in the validation dataset as well as the stimuli. We plan to extend these methods to non-binary individuals and non-categorical representations. However, precisely validating such an approach would require us to construct the corresponding validation sets, which currently don't exist. **Generalizing the method to represent humans with continuous values as opposed to categorical group labels is left to future work.** There are in total $M \times N$ combinations of intersectional groups C_{mn}. We use all groups C_{mn} to build WEFAT pairs $P_{ij} = (C_{11}, C_{ij}), i = 1, ..., M, j = 1, ..., N$. Then, we detect lists of words associated with each pair $W_{ij}, i = 1, ..., M, j = 1, ..., N$ based on threshold t determined by an ROC curve. We detect the attributes highly associated with the intersectional group, for example C_{11}, from all $(M \times N)$ WEFAT pairs. We define the words associated with intersectional biases of group C_{11} as W_{IB} and these words are identified by

$$W_{IB} = \bigcup_{\substack{1 \leq i \leq M \\ 1 \leq j \leq N}} W_{IB_{ij}},$$

where

$$W_{IB_{ij}} = \{w | s(w, C_{11}, C_{ij}) > t_{mn}, w \in W_{IB_{mn}}\}$$

where

$$W_{IB_{mn}} = \{(\bigcup_{\substack{1 \leq i \leq M \\ 1 \leq j \leq N}} W_{ij}) \cup W_{random}\}$$

W_{11} contains validated words associated with C_{11}. Each W_{ij} contains validated words associated with one intersectional group [27]. W_{random} contains random words, which are stimuli taken from WEAT that are not associated with any C_{ij}, thus represent true negatives.

To identify the thresholds, we treat IBD as a one-vs-all verification classifier in machine learning to determine whether attributes belong to group C_{11}. We select the threshold with the highest value of $true\ positive\ rate - false\ positive\ rate$ $(TPR - FPR)$. When multiple thresholds have the same values, we select the one with the highest TP to detect more attributes associated with C_{11}. Detection accuracy is calculated as true positives plus true negatives divided by true positives plus true negatives plus false positives plus false negatives ($\frac{TP+TN}{TP+TN+FP+FN}$). The attributes which are associated with C_{11} and are detected as C_{11} are TP. The attributes which are not associated with C_{11} and are not detected as C_{11} are TN. The attributes which are associated with C_{11} but are not detected as C_{11} are FN. The attributes which are not associated with C_{11} but are detected as C_{11} are FP.

4.3 Emergent Intersectional Bias Detection (EIBD)

EIBD identifies words that are uniquely associated with intersectional group members. These emergent biases are only associated with the intersectional group (e.g., African American females C_{11}) but not associated with its constituent category such as African Americans S_{1n} or females S_{m1}. EIBD is a modified and extended version of IBD.

Conceptually, to detect words uniquely associated with African American females in a set of attributes W, we assume there are two classes (females, males) of gender and two classes (African Americans, European Americans) of race. We measure the relative association of all words in W first with African American females and African American males, second with African American females and European American females, third with African American females and European American males. (Fourth is the comparison of the same groups, which leads to $d = 0$ effect size, which is always below the detection threshold.) The union of attributes with an association score greater than the selected threshold represents intersectional biases associated with African American females. Then, we calculate the association scores of these IBD attributes first with females and males, second with African Americans and European Americans. We remove the attributes with scores greater than the selected threshold from these IBD attributes, that are highly associated with single social categories. The union of the remaining attributes are the emergent intersectional biases.

Formal Definition of EIBD. We first detect C_{11}'s intersectional biases W_{IB} with IBD. Then, we detect the biased attributes associated with only one constituent category of the intersectional group C_{11} (e.g., associated only with race S_{1n} - or only with gender S_{m1}). Each intersectional category C_{1n} has M constituent subcategories $S_{in}, i = 1, ...M$ and category C_{m1} has N constituent subcategories $S_{mj}, j = 1, ..., N$. S_{1n} and S_{m1} are the constituent subcategories of intersectional group C_{11}.

There are in total $M + N$ groups defined by all the single constituent subcategories. We use all $M + N$ groups to build WEFAT pairs $P_i = (S_{1n}, S_{in}), i = 1, ..., M$ and $P_j = (S_{m1}, S_{mj}), j = 1, ...N$. Then, we detect lists of words associated with each pair $W_i, i = 1, ...M$ and $W_j, j = 1, ..., N$ based on the same positive threshold t_{mn} used in IBD. We detect the attributes highly associated with the constituent subcategories S_{1n} and S_{m1} of the target intersectional group C_{11} from all $(M + N)$ WEFAT pairs. We define the words associated with emergent intersectional biases of group C_{11} as W_{EIB} and these words are identified by the formula

$$W_{EIB} = (\bigcup_{i=1}^{M}(W_{IB} - W_i)) \bigcup (\bigcup_{j=1}^{N}(W_{IB} - W_j))$$

where

$$W_i = \{w | s(w, S_{1n}, S_{in}) > t_{mn}, w \in W_{IB}\}$$

and

$$W_j = \{w | s(w, S_{m1}, S_{mj}) > t_{mn}, w \in W_{IB}\}$$

4.4 Contextualized Embedding Association Test (CEAT)

CEAT quantifies the overall magnitude of social biases in CWE by extending the WEAT methodology that measures human-like biases in SWE [11]. WEAT's bias metric is effect size (Cohen's d). In CWE, since embeddings of the same word vary based on context, applying WEAT to a biased set of CWE will not measure bias comprehensively. To deal with a range of dynamic embeddings representing individual words, CEAT measures the distribution of effect sizes that are embedded in a neural language model.

In WEAT's formal definition [11], X and Y are two sets of target words of equal size; A and B are two sets of evaluative polar attribute words of equal size. Each word in these sets of words is referred to as a stimulus. Let $cos(\vec{a}, \vec{b})$ stand for the cosine similarity between vectors $\vec{a}$ and $\vec{b}$. WEAT measures the magnitude of bias by computing the effect size (ES) which is the standardized differential association of the targets and attributes. The p-value (P_w) of WEAT measures the probability of observing the effect size in the null hypothesis, in case biased associations did not exist. According to Cohen's effect size metric, $d >| 0.5 |$ and $d >| 0.8 |$ are medium and large effect sizes, respectively [56].

In a neural language model, each stimulus s from WEAT contained in n_s input sentences has at most n_s different CWE $\vec{s_1}, ..., \vec{s_{n_s}}$ depending on the context in which it appears. If we calculate effect size $ES(X, Y, A, B)$ with all different $\vec{s}$ for a stimulus $s \in X$ and keep the CWE for other stimuli unchanged, there will be at most n_s different values of effect size. For example, if we assume each stimulus s occurs in 2 contexts and each set in X, Y, A, B has 5 stimuli, the total number of combinations for all the CWE of stimuli will be $2^{5\times4} = 1,048,576$. The numerous possible values of $ES(X, Y, A, B)$ construct a *distribution* of effect sizes, therefore we extend WEAT to CEAT.

For each CEAT, all the sentences, where a CEAT stimulus occurs, are retrieved from the Reddit corpus. Then, we generate the corresponding CWE from these sentences with randomly varying contexts. In this way, we generate n_s CWE from n_s extracted sentences for each stimulus s, where n_s can vary according to the contextual variance of each stimulus. We sample random combinations of CWE for each stimulus N times. In the i^{th} sample out of N, for each stimulus that appears in at least N sentences, we randomly sample one of its CWE vectors without replacement. If a stimulus occurs in less than N sentences, especially when N is very large, we randomly sample from its CWE vectors with replacement so that they can be reused while preserving their distribution. Pooling $N = 1,000$ and $N = 10,000$ embeddings for each stimulus result in similar bias magnitudes. Based on the sampled CWEs, we calculate each WEAT sample's effect size $ES_i(X, Y, A, B)$, sample variance $V_i(X, Y, A, B)$ and p-value $P_{w_i}(X, Y, A, B)$. Then, we generate N WEAT samples to approximate the distribution of effect sizes via CEAT using random-effects modeling from meta-analysis.

4.5 Random-Effects Model

Meta-analysis is the statistical procedure for combining data from multiple studies [38]. Meta-analysis describes the results of each separate study by a numerical index (e.g., effect size) and then summarizes the results into combined statistics. In bias measurements, we are dealing with effect size. Based on different assumptions whether the effect size is fixed or not, there are two kinds of methods: *fixed-effects* model and *random-effects* model. On the one hand, fixed-effects model expects results with fixed-effect sizes from different intervention studies. On the other hand, random-effects model treats the effect size as they are samples from a random distribution of all possible effect sizes [18, 37]. The expected results of different intervention studies in the random-effects model don't have to match other studies' results.

The distribution of bias effects in CEAT represents random-effects computed by WEAT where we do not expect to observe the same effect size due to variance in context [36]. As a result, in order to provide comprehensive summary statistics, we applied a random-effects model from the validated meta-analysis literature to compute the weighted mean of the effect sizes and statistical significance [7, 58]. The summary of the effect magnitude of a particular bias in a neural language model, namely combined effect size (CES), is the weighted mean of a distribution of random-effects,

$$CES(X, Y, A, B) = \frac{\sum_{i=1}^{N} v_i ES_i}{\sum_{i=1}^{N} v_i}$$

where v_i is the inverse of the sum of in-sample variance V_i and between-sample variance in the distribution of random-effects $\sigma^2_{between}$.

5 RESULTS AND EVALUATION

We measure ten types of social biases via WEAT (C1-C10) and construct our own intersectional bias tests in ELMo, BERT, GPT, and GPT-2. Accordingly, we present four novel intersectional bias tests via IBD and EIBD for studying African American, European American, and Mexican American men and women.

We use the stimuli introduced in Section 3.4 to represent the target groups. For intersectional and emergent bias tests, we use the attributes associated with the intersectional minority or disadvantaged group members vs. the majority European American males as the two polar attribute sets. We sample $N = 10,000$ combinations of CWE for each CEAT since according to various evaluation trials, the resulting CES and p-value remain consistent under this hyper-parameter.

5.1 Evaluation of IBD and EIBD

We use IBD and EIBD to automatically detect and retrieve the intersectional and emergent biases associated with intersectional group members (e.g., African American females, Mexican American females) in GloVe SWE. To evaluate our methods IBD and EIBD, we use validated stimuli provided in prior work that represent each social group with frequent given names, as explained in Section 3. IBD and EIBD experiments use the same test set consisting of 98 attributes associated with 2 groups defined by gender (females, males), 3 groups defined by race (African American, European American, Mexican American), 6 intersectional groups in total defined by race and gender, in addition to random words taken from WEAT not associated with any group [27]. These random words represent the true negatives for evaluating the identification task.

We draw the ROC curves of four bias detection tasks in Figure 2, then select the highest value of $TPR - FPR$ as thresholds for each intersectional group. IBD achieves an accuracy of 81.6% and 82.7%, respectively, when detecting the intersectional biases of African American females and Mexican American females, where the random correct identification rates are 14.3% and 13.3%. EIBD reaches an accuracy of 84.7% and 65.3%, respectively, when detecting the emergent intersectional biases unique to African American females and Mexican American females. The probability of random correct attribute detection in EIBD tasks are 9.2% and 6.1%. Intersectional biases have the highest magnitude compared to other biases across all language models, potentially disadvantaging members that belong to multiple minority groups in downstream applications.

The current validation set with ground truth information about each word constrains our evaluation to a closed-world machine learning classification task, where we know the category each stimulus belongs to. However, evaluating the entire semantic space resembles an open-world machine learning problem where millions of stimuli in the entire word embedding vocabulary belong to unknown categories, thus require human subject annotation studies. In future work, a human subject study can further evaluate the threshold selection criteria, which would require validating a large set of biases retrieved from the entire vocabulary.

5.2 Evaluation of CEAT

Congruent with Caliskan et al. [11]'s WEAT findings, Table 1 presents significant effect sizes for all previously documented and validated biases. GPT-2 exhibited less bias than other neural language models. Our method CEAT, designed for CWEs, computes the combined bias score of a distribution of effect sizes present in neural language models. We find that the effect magnitudes of biases reported by Tan and Celis [61] are individual samples in the distributions generated by CEAT. We can view their method as a special case of CEAT that calculates the individual bias scores of a few pre-selected samples. In order to comprehensively measure the overall bias score in a neural language model, we apply a random-effects model from the meta-analysis literature that computes combined effect size and combined statistical significance from a distribution of bias measurements. As a result, when CEAT reports significant results, some of the corresponding bias scores in prior work are not statistically significant. Furthermore, our results indicate statistically significant bias in the opposite direction in some cases. These negative results suggest that some WEAT stimuli tend to occur in stereotype-incongruent contexts more frequently.

We sampled combinations of CWE 10,000 times for each CEAT test; nonetheless, we observed varying intensities of the same social bias in different contexts. Using a completely random set vs fixed set of contexts derived from 10,000 sentences lead to low variance

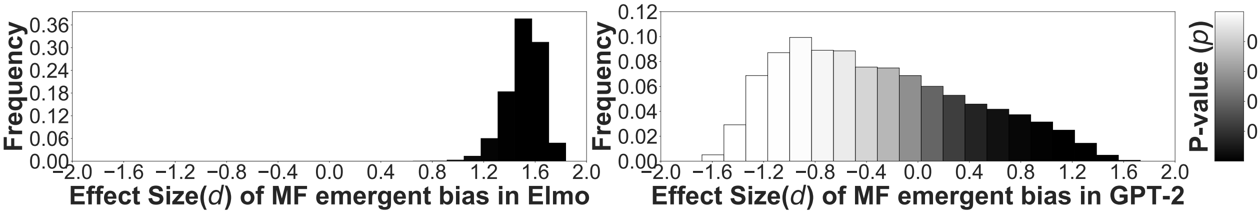

Figure 1: Distributions of effect sizes with ELMo (CES $d = 1.51$) and GPT-2 (CES $d = -0.32$) for emergent intersectional bias CEAT test I4. Test I4, after identifying the emergent and intersectional biases associated with Mexican American females and European American males (MF/EM) via IBD and EIBD in word embeddings, CEAT measures the overall distribution of biased associations for the retrieved stimuli in the neural language models. This example is chosen to demonstrate how different models exhibit varying degrees of bias when using the same set of stimuli to measure bias. The height of each bar shows the frequency of observed effect sizes among 10,000 effect size samples of a particular bias type that fall in each bin. The color coded bars stand for the average p-value of all effect sizes corresponding to that bin.

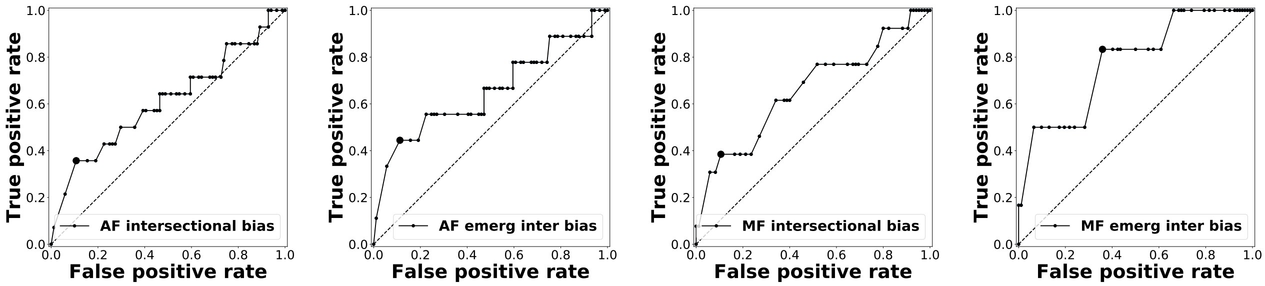

Figure 2: ROC curves of IBD and EIBD for African American females (AF) and Mexican American females (MF). The value that maximizes the $true\ positive\ rate\ -\ false\ positive\ rate$ is selected as the optimal threshold marked with a dot. 'emerg inter bias' stands for emergent intersectional bias.

in corresponding bias scores. Using a fixed set of contexts for each model makes it possible to evaluate the magnitude of bias across models for the same variables. Experiments conducted with $1,000$, $5,000$, $10,000$ samples of CWE lead to similar bias scores with low variance. As a result, the number of samples can be adjusted according to computational resources. However, future work on evaluating the lower bound of sampling size with respect to model and corpus characteristics would optimize the sampling process. Accordingly, the computation of overall bias in the language model would become more efficient.

5.3 IBD, EIBD, and CEAT Results

We report the overall magnitude of bias (CES) and p-value in Table 1. We pick an example from Table 1 that reflects the great disparity in bias magnitudes between the two models. We present the distribution histograms of effect sizes in Figure 1, which show the overall biases that can be measured with a comprehensive contextualized bias test related to the emergent biases associated with occurrences of stimuli unambiguously regarding Mexican American females (See row I4 in Table 1) with ELMo and GPT-2. The distribution plots for other bias tests are provided in our project repository.

We find that CEAT uncovers more evidence of intersectional bias than gender or racial biases. This findings suggest that, members of multiple minority or disadvantaged groups are associated with the strongest levels of bias in neural language representations. To quantify the intersectional biases in CWEs, we construct tests I1-I4. Tests with Mexican American females tend to have stronger bias with a higher CES than those with African American females. Specifically, 13 of 16 instances in intersection-related tests (I1-I4) have significant stereotype-congruent CES; 9 of 12 instances in gender-related tests (C6-C8) have significant stereotype-congruent CES; 8 of 12 instances in race-related tests (C3-C5) have significant stereotype-congruent CES. In gender bias tests, the gender associations with career and family are stronger than other biased gender associations. In all models, the significantly biased intersectionality associations have larger effect sizes than racial biases.

According to CEAT results in Table 1, ELMo is the most biased whereas GPT-2 is the least biased with respect to the types of biases CEAT measures. We notice that significant negative CES exist in BERT, GPT and GPT-2, which imply that stereotype-incongruent biases with small effect size exist.

6 DISCUSSION

According to our findings, GPT-2 has the highest variance in bias magnitudes followed by GPT, BERT, and ELMo (see an example in Figure 1). The overall magnitude of bias decreases in the same order for the types of biases we measured. The similar number of parameters in these models or the sizes of the training corpora do not explain the distribution of bias that we observe w.r.t. variance and overall magnitude. However, Ethayarajh [21] note the same descending pattern when measuring words' self-similarity, after adjusting for anisotropy (non-uniform directionality), across their

Test			ELMo		BERT		GPT		GPT-2	
			d	*p*	*d*	*p*	*d*	*p*	*d*	*p*
C1:	Flowers/Insects	random	1.40	$< 10^{-30}$	0.97	$< 10^{-30}$	1.04	$< 10^{-30}$	0.14	$< 10^{-30}$
	Pleasant/Unpleasant*	fixed	1.35	$< 10^{-30}$	0.64	$< 10^{-30}$	1.01	$< 10^{-30}$	0.21	$< 10^{-30}$
C2:	Instruments/Weapons	random	1.56	$< 10^{-30}$	0.94	$< 10^{-30}$	1.12	$< 10^{-30}$	-0.27	$< 10^{-30}$
	Pleasant/Unpleasant*	fixed	1.59	$< 10^{-30}$	0.54	$< 10^{-30}$	1.09	$< 10^{-30}$	-0.21	$< 10^{-30}$
C3:	EA/AA names	random	0.49	$< 10^{-30}$	0.44	$< 10^{-30}$	-0.11	$< 10^{-30}$	-0.19	$< 10^{-30}$
	Pleasant/Unpleasant*	fixed	0.47	$< 10^{-30}$	0.31	$< 10^{-30}$	-0.10	$< 10^{-30}$	0.09	$< 10^{-30}$
C4:	EA/AA names	random	0.15	$< 10^{-30}$	0.47	$< 10^{-30}$	0.01	$< 10^{-2}$	-0.23	$< 10^{-30}$
	Pleasant/Unpleasant*	fixed	0.23	$< 10^{-30}$	0.49	$< 10^{-30}$	0.00	0.20	-0.13	$< 10^{-30}$
C5:	EA/AA names	random	0.11	$< 10^{-30}$	0.02	$< 10^{-7}$	0.07	$< 10^{-30}$	-0.21	$< 10^{-30}$
	Pleasant/Unpleasant*	fixed	0.17	$< 10^{-30}$	0.07	$< 10^{-30}$	0.04	$< 10^{-27}$	-0.01	0.11
C6:	Males/Female names	random	1.27	$< 10^{-30}$	0.92	$< 10^{-30}$	0.19	$< 10^{-30}$	0.36	$< 10^{-30}$
	Career/Family	fixed	1.31	$< 10^{-30}$	0.41	$< 10^{-30}$	0.11	$< 10^{-30}$	0.34	$< 10^{-30}$
C7:	Math/Arts	random	0.64	$< 10^{-30}$	0.41	$< 10^{-30}$	0.24	$< 10^{-30}$	-0.01	$< 10^{-2}$
	Male/Female terms	fixed	0.71	$< 10^{-30}$	0.20	$< 10^{-30}$	0.23	$< 10^{-30}$	-0.14	$< 10^{-30}$
C8:	Science/Arts	random	0.33	$< 10^{-30}$	-0.07	$< 10^{-30}$	0.26	$< 10^{-30}$	-0.16	$< 10^{-30}$
	Male/Female terms	fixed	0.51	$< 10^{-30}$	0.17	$< 10^{-30}$	0.35	$< 10^{-30}$	-0.05	$< 10^{-30}$
C9:	Mental/Physical disease	random	1.00	$< 10^{-30}$	0.53	$< 10^{-30}$	0.08	$< 10^{-29}$	0.10	$< 10^{-30}$
	Temporary/Permanent	fixed	1.01	$< 10^{-30}$	0.40	$< 10^{-30}$	-0.23	$< 10^{-30}$	-0.21	$< 10^{-30}$
C10:	Young/Old people's names	random	0.11	$< 10^{-30}$	-0.01	0.016	0.07	$< 10^{-30}$	-0.16	$< 10^{-30}$
	Pleasant/Unpleasant*	fixed	0.24	$< 10^{-30}$	0.07	$< 10^{-30}$	0.04	$< 10^{-17}$	-0.14	$< 10^{-30}$
I1:	AF/EM names	random	1.24	$< 10^{-30}$	0.77	$< 10^{-30}$	0.07	$< 10^{-30}$	0.02	$< 10^{-2}$
	AF/EM intersectional	fixed	1.25	$< 10^{-30}$	0.98	$< 10^{-30}$	0.23	$< 10^{-30}$	-0.19	$< 10^{-30}$
I2:	AF/EM names	random	1.25	$< 10^{-30}$	0.67	$< 10^{-30}$	-0.09	$< 10^{-30}$	0.02	$< 10^{-2}$
	AF emergent/EM intersectional	fixed	1.27	$< 10^{-30}$	1.00	$< 10^{-30}$	0.23	$< 10^{-30}$	-0.14	$< 10^{-30}$
I3:	MF/EM names	random	1.31	$< 10^{-30}$	0.68	$< 10^{-30}$	-0.06	$< 10^{-30}$	0.38	$< 10^{-30}$
	MF/EM intersectional	fixed	1.29	$< 10^{-30}$	0.51	$< 10^{-30}$	0.00	0.81	0.32	$< 10^{-30}$
I4:	MF/EM names	random	1.51	$< 10^{-30}$	0.86	$< 10^{-30}$	0.16	$< 10^{-30}$	-0.32	$< 10^{-30}$
	MF emergent/EM intersectional	fixed	1.43	$< 10^{-30}$	0.58	$< 10^{-30}$	0.20	$< 10^{-30}$	-0.25	$< 10^{-30}$

*Pleasant and unpleasant attributes used to measure valence and attitudes towards targets from Greenwald et al. [30].

Table 1: CEAT measures of social and intersectional biases in language models. We report the overall magnitude of bias in language models with CES (d, rounded down) and statistical significance with combined p-values (p, rounded up). CES pools $N = 10,000$ samples from a random-effects model. The first row for each bias test uses completely random samples, whereas the second row for the bias test uses the same sentences to generate CWE across all neural language models. Ci stands for the i^{th} WEAT in Caliskan et al. [11]'s Table 1. Ii stands for our tests constructed for measuring intersectional biases. $A\_$ stands for African Americans, $E\_$ for European Americans, $M\_$ for Mexican Americans, $\_F$ for females, and $\_M$ for males. Light, medium, and dark gray shading of combined d values (CES) indicates small, medium, and large effect size, respectively.

CWE in GPT-2, BERT, and ELMo. (ELMo is compared in three layers due to its architecture.) Ethayarajh [21] also find that upper layers of contextualizing models produce more context-specific representations. Quantifying how contextualized these dynamic embeddings are supports our findings that the highest variance in bias magnitude, low overall bias, and low self-similarity correlate. This correlation may explain the results that we are observing. As more recent models are learning highly-contextualized CWE in upper layers, the representations in highly-contextualized layers are almost overfitting to their contexts. Since words appear in numerous contexts, the more contextualized and diverse a word's representation becomes, the less overall bias and general stereotypical associations.

We present and validate a bias detection method generalizable to identifying biases associated with any social group or intersectional group member. We detect and measure biases associated with Mexican American and African American females in SWE and CWE. Our emergent intersectional bias measurement results for African American females are in line with previous findings [45, 61]. IBD and EIBD can detect intersectional biases from SWE with high accuracy in an semi-supervised manner by following a lexicon induction strategy [35]. This approach can be complementary to the stimuli list predefined by social psychologists. Our current intersectional bias detection validation approach can be used to identify association thresholds when generalizing this work to the entire word embedding dictionary. Exploring all the potential biases associated with targets is left to future work since it requires extensive human subject validation studies in collaboration with social psychologists. We list all the stimuli representing biased associations in our open source repository. To name a few, the superset of intersectional biases associated with African American females are: aggressive, assertive, athletic, bigbutt, confident, darkskinned, fried-chicken, ghetto, loud, overweight, promiscuous, unfeminine, unintelligent, unrefined. Emergent intersectional biases associated with African American females are: aggressive, assertive, bigbutt, confident, darkskinned, fried-chicken, overweight, promiscuous, unfeminine. The superset of intersectional biases associated with Mexican American females are: attractive, cook, curvy, darkskinned, feisty, hardworker, loud, maids, sexy, short, uneducated, unintelligent. Emergent intersectional biases associated with Mexican American females are: cook, curvy, feisty, maids, sexy.

We follow the conventional method of using the most frequent given names in a social group that signal group membership in order to accurately represent targets [11, 30]. Our results indicate that the conventional method that relies on stimuli selected by experts in social psychology works accurately. Prior work on lexicon induction methods compensates for the lack of existing annotated data on valence [35, 57, 66]. Nevertheless, principled and robust lexicon induction methods can be validated in this domain, when measuring the representation accuracy of target group lexica or any semantic concept. Developing these principled methods is left to future work.

Semantics of languages can be represented by the distributional statistics of word co-occurrences [23, 34]. Consequently, our methods are language agnostic and can be applied to neural language models as well as word embeddings in any language as long as the stimuli for accurately representing the semantics of concepts are available. Project Implicit (https://implicit.harvard.edu/implicit) has been hosting IATs for human subjects all over the world in numerous languages for two decades. As a result, their IATs, that inspired WEATs, provide stimuli for targets and attributes in numerous languages. We leave generalizing our methods to other languages to future work since state-of-the-art neural language models are not widely or freely available for languages other than English as of 2021.

When simulating contexts for WEAT, we make an assumption that the Reddit corpus represents naturally occurring sentences. Nevertheless, we acknowledge that the Reddit corpus also reflects the biases of the underlying population contributing to its corpus. Studying the accuracy of simulating the most common distribution of contexts and co-occurring stimuli is left to future work since we don't have validated ground truth data for evaluating the distribution parameters of contexts in large-scale corpora. Instead, for evaluation, validation, and comparison, we rely on validated ground truth information about biases documented by Caliskan et al. [11] in word embeddings as well as biases documented by millions of people over decades via the implicit association literature [48] and Ghavami and Peplau [27]'s intersectional biases.

Given the carbon footprint, data collection, computational, energy, and funding considerations, we are not training the studied large language models on the same large-scale corpora to compare how a neural language model's architecture learns biases [3]. The size of state-of-the-art models increase by at least a factor of 10 every year. BERT-Large from 2018 has 355 million parameters, GPT-2 from early 2019 reaches 1.5 billion, and GPT-3 from mid-2020 finally gets to 175 billion parameters. The GPT-2 model used 256 Google Cloud TPU v3 cores for training, which costs 256 US dollars per hour. GPT-2 requires approximately 168 hours or 1 week of training on 32 TPU v3 chips [59]. GPT-3 is estimated to cost ∼12 million US dollars [24] and we are not able to get access to its embeddings or training corpora. Regardless, measuring the scope of biases with validated bias quantification and meta-analysis methods, we are able to compare the biased associations learned by neural language models that are widely used. Being able to study neural language models comprehensively is critical since they are replacing SWE in many NLP applications due to their high accuracy in various machine learning tasks. Since language models are making consequential decisions about individuals while shaping society, having access to these models, their training data and source code [25] would provide an opportunity to analyze and mitigate the harmful effects of AI systems.

We would like to conclude the discussion with our ethical concerns regarding the dual use of IBD and EIBD, that can detect stereotypical associations for an intersectional group or disadvantaged individuals. Words retrieved by our methods may be used in the generation of offensive or stereotypical content that perpetuates or amplifies existing biases. For example, information influence operations in the 1970s used Osgood [49]'s semantic differential technique among human subjects to retrieve the words that would most effectively induce a negative attitude in a South American population towards their administration [44]. Similarly, biased neural language models may be exploited to automate large-scale information influence operations that intend to sow discord among social groups [64, 65]. The biased outputs of these language models, that get recycled in future model generation's training corpora, may lead to an AI bias feedback cycle.

7 CONCLUSION

We introduce methods called IBD and EIBD to identify biases associated with members of multiple minority groups. These methods automatically detect the intersectional biases and emergent intersectional biases captured by word embeddings. Intersectional biases associated with African American and Mexican American females have the highest effect size compared to other social biases. Complementary to pre-defined sets of attributes to measure widely known biases, our methods automatically discover biases. IBD reaches an accuracy of 81.6% and 82.7% in detection, respectively, when validating on the intersectional biases of African American females and Mexican American females. EIBD reaches an accuracy of 84.7% and 65.3% in detection, respectively, when validating on the emergent intersectional biases of African American females and Mexican American females.

We present CEAT to measure biases identified by IBD and EIBD in language models. CEAT uses a random-effects model to comprehensively measure social biases embedded in neural language models that contain a distribution of context-dependent biases. CEAT simulates this distribution by sampling ($N = 10,000$) combinations of CWEs without replacement from a large-scale natural language corpus. Unlike prior work that focuses on a limited number of contexts defined by templates to measure the magnitude of particular biases, CEAT provides a comprehensive measurement of overall bias in contextualizing language models. Our results indicate that ELMo is the most biased, followed by BERT, and GPT. GPT-2 is the least biased language model with respect to the social biases we investigate. The overall magnitude of bias negatively correlates with the level of contextualization in the language model. Understanding how the architecture of a language model contributes to biased and contextualized word representations can help mitigate the harmful effects to society in downstream applications.

REFERENCES

[1] Renata Arrington-Sanders, Jessica Oidtman, Anthony Morgan, Gary Harper, Maria Trent, and J Dennis Fortenberry. 2015. 13. Intersecting Identities in Black Gay and Bisexual Young Men: A Potential Framework for HIV Risk. *Journal of Adolescent Health* 56, 2 (2015), S7–S8.

[2] Matthias Aßenmacher and Christian Heumann. 2020. On the comparability of pre-trained language models. *arXiv preprint arXiv:2001.00781* (2020).

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

[4] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of" Bias" in NLP. *arXiv preprint arXiv:2005.14050* (2020).

[5] Bernd Bohnet, Ryan McDonald, Goncalo Simoes, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic tagging with a meta-bilstm model over context sensitive token encodings. *arXiv preprint arXiv:1805.08237* (2018).

[6] 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*. 4349–4357.

[7] Michael Borenstein, Larry Hedges, and Hannah Rothstein. [n.d.]. Meta-analysis: Fixed effect vs. random effects. ([n. d.]).

[8] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165* (2020).

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[10] Angel Alexander Cabrera, Minsuk Kahng, Fred Hohman, Jamie Morgenstern, and Duen Horng Chau. [n.d.]. DISCOVERY OF INTERSECTIONAL BIAS IN MACHINE LEARNING USING AUTOMATIC SUBGROUP GENERATION. ([n. d.]).

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

[12] Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. AI now 2017 report. *AI Now Institute at New York University* (2017).

[13] Tessa ES Charlesworth, Victor Yang, Thomas C Mann, Benedek Kurdi, and Mahzarin R Banaji. 2021. Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words. *Psychological Science* 32, 2 (2021), 218–240.

[14] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005* (2013).

[15] Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.* (1989), 139.

[16] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 120–128.

[17] Rebecca DerSimonian and Raghu Kacker. 2007. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary clinical trials* 28, 2 (2007), 105–114.

[18] Rebecca DerSimonian and Nan Laird. 1986. Meta-analysis in clinical trials. *Controlled clinical trials* 7, 3 (1986), 177–188.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[20] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381* (2018).

[21] Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512* (2019).

[22] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1696–1705.

[23] John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis* (1957).

[24] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30, 4 (2020), 681–694.

[25] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027* (2020).

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

[27] Negin Ghavami and Letitia Anne Peplau. 2013. An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly* 37, 1 (2013), 113–127.

[28] 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. *arXiv preprint arXiv:1903.03862* (2019).

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

[30] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74, 6 (1998), 1464.

[31] Anthony G Greenwald, Brian A Nosek, and Mahzarin R Banaji. 2003. Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of personality and social psychology* 85, 2 (2003), 197.

[32] Ange-Marie Hancock. 2007. When multiplication doesn't equal quick addition: Examining intersectionality as a research paradigm. *Perspectives on politics* 5, 1 (2007), 63–79.

[33] Rachel T Hare-Mustin and Jeanne Marecek. 1988. The meaning of difference: Gender theory, postmodernism, and psychology. *American psychologist* 43, 6 (1988), 455.

[34] Zellig S Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162.

[35] Vasileios Hatzivassiloglou and Kathleen McKeown. 1997. Predicting the semantic orientation of adjectives. In *35th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics*. 174–181.

[36] Larry V Hedges. 1983. A random effects model for effect sizes. *Psychological Bulletin* 93, 2 (1983), 388.

[37] Larry V Hedges and Ingram Olkin. 2014. *Statistical methods for meta-analysis*. Academic press.

[38] Larry V Hedges and Jack L Vevea. 1998. Fixed-and random-effects models in meta-analysis. *Psychological methods* 3, 4 (1998), 486.

[39] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[40] Aída Hurtado and Mrinal Sinha. 2008. More than men: Latino feminist masculinities and intersectionality. *Sex Roles* 59, 5-6 (2008), 337–349.

[41] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5491–5501.

[42] Arnold S Kahn and Janice D Yoder. 1989. The psychology of women and conservatism: Rediscovering social change. *Psychology of Women Quarterly* 13, 4 (1989), 417–432.

[43] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Quantifying Social Biases in Contextual Word Representations. In *1st ACL Workshop on Gender Bias for Natural Language Processing*.

[44] Fred Landis, Richard Lewontin, Luc Desnoyers, Donna Mergler, and Anthony Weston. 1982. CIA psychological warfare operations. *Case Studies in Chile, Jamaica and Nicaragua. Science for the People* 14 (1982), 6–11.

[45] Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561* (2019).

[46] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[47] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).

[48] Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice* 6, 1 (2002), 101.

[49] Charles E Osgood. 1964. Semantic differential technique in the comparative study of cultures. *American Anthropologist* 66, 3 (1964), 171–200.

[50] Maryann Parada. 2016. Ethnolinguistic and gender aspects of Latino naming in Chicago: Exploring regional variation. *Names* 64, 1 (2016), 19–35.

[51] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[52] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[53] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf* (2018).

[54] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[55] M Raghavan and S Barocas. [n.d.]. Challenges for mitigating bias in algorithmic hiring. 2019. *URL https://www. brookings. edu/research/challenges-for-mitigating-bias-in-algorithmic-hiring* ([n. d.]).

[56] Marnie E Rice and Grant T Harris. 2005. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and human behavior* 29, 5 (2005), 615–620.

[57] Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*. 105–112.

[58] Robert Rosenthal and M Robin DiMatteo. 2002. Metaanalysis. *Stevens' handbook of experimental psychology* (2002).

[59] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).

[60] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding?. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 305–311.

[61] Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*. 13209–13220.

[62] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316* (2019).

[63] Veronica G Thomas and Shari E Miles. 1995. Psychology of Black women: Past, present, and future. (1995).

[64] Autumn Toney and Aylin Caliskan. 2020. ValNorm: A New Word Embedding Intrinsic Evaluation Method Reveals Valence Biases are Consistent Across Languages and Over Decades. *arXiv preprint arXiv:2006.03950* (2020).

[65] Autumn Toney, Akshat Pandey, Wei Guo, David Broniatowski, and Aylin Caliskan. 2020. Pro-Russian Biases in Anti-Chinese Tweets about the Novel Coronavirus. *arXiv preprint arXiv:2004.08726* (2020).

[66] Peter D Turney and Michael L Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)* 21, 4 (2003), 315–346.

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[68] Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[69] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*. 5754–5764.

[70] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310* (2019).

[71] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496* (2018).

[72] 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*. 19–27.