Mapping the Multilingual Margins: Intersectional Biases of Sentiment Analysis Systems in English, Spanish, and Arabic

António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, Richard Zemel
Department of Computer Science, Columbia University
{ac4443,nat2142,ta2553}@columbia.edu
{eallaway,zemel}@cs.columbia.edu

Abstract

As natural language processing systems become more widespread, it is necessary to address fairness issues in their implementation and deployment to ensure that their negative impacts on society are understood and minimized. However, there is limited work that studies fairness using a multilingual and intersectional framework or on downstream tasks. In this paper, we introduce four multilingual Equity Evaluation Corpora, supplementary test sets designed to measure social biases, and a novel statistical framework for studying unisectional and intersectional social biases in natural language processing. We use these tools to measure gender, racial, ethnic, and intersectional social biases across five models trained on emotion regression tasks in English, Spanish, and Arabic. We find that many systems demonstrate statistically significant unisectional and intersectional social biases.[1]

1 Introduction

Large-scale transformer-based language models, such as BERT (Devlin et al., 2018), are now the state-of-the-art for a myriad of tasks in natural language processing. However, these models are well-documented to perpetuate harmful social biases, specifically by regurgitating the social biases present in their training data which are scraped from the Internet without careful consideration (Bender et al., 2021). While steps have been taken to "debias", or remove, gender and other social biases from word embeddings (Bolukbasi et al., 2016; Manzini et al., 2019), these methods have been demonstrated to be cosmetic (Gonen and Goldberg, 2019). Furthermore, these studies neglect to recognize both the impact of social biases on downstream task results as well as the complex and interconnected nature of social biases. In this paper, we detect and discuss unisectional[2] and intersectional social biases in multilingual language models applied to downstream tasks using a novel statistical framework and novel multilingual datasets.

Intersectionality is a framework introduced by Crenshaw (1990) to study how the composite identity of an individual across different social cleavages (e.g., race and gender) informs that individual's social advantages and disadvantages. For example, individuals who identify with multiple disadvantaged social cleavages (e.g., Black women) face a greater and altered risk for discrimination and oppression than individuals with a subset of those identities (e.g., white women). This framework for understanding overlapping systems of discrimination has been explored in some studies of fairness in machine learning, including by Buolamwini and Gebru (2018) who show that face detection systems perform markedly worse for female users of color, compared to female users or users of color.

Although work has begun to study intersectional social biases in natural language processing, to the best of our knowledge no work has explored fairness in an intersectional framework on downstream tasks (e.g. sentiment analysis). Social biases in downstream tasks expose users with multiple disadvantaged sensitive attributes to unknown but potentially harmful outcomes, especially when models trained on downstream tasks are used in real-world decision making, such as for screening résumes or predicting recidivism in criminal proceedings (Bolukbasi et al., 2016; Angwin et al., 1999). In this work, we choose emotion regression as a downstream task because social biases are often realized through emotion recognition (Elfenbein and Ambady, 2002) and machine learning models have been shown to reflect gender bias in emotion recognition tasks (Domnich and Anbarjafari, 2021). For

[1] We make our code and datasets available for download at https://github.com/ascamara/ml-intersectionality.

[2] In this paper, we refer to biases against a single social cleavage, such as racial bias or gender bias, as unisectional.

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 90 - 106

example, sentiment analysis and emotion regression may be used by companies to measure product engagement for different social groups.

In addition, while some work has studied gender biases across different languages (Zhou et al., 2019; Zhao et al., 2020), no work to our knowledge has studied racial, ethnic, and intersectional social biases across different languages. This lack of a multilingual analysis neglects non-English speaking users and their complex social environments.

In this paper, we demonstrate the presence of gender, racial, ethnic, and intersectional social biases on five language models trained on an emotion regression task in English, Spanish, and Arabic. We do so by introducing novel supplementary test sets designed to measure social biases and a novel statistical framework for detecting the presence of unisectional and intersectional social biases in models trained on sentiment analysis tasks.

Our contributions are summarized as:

- Following Kiritchenko and Mohammad (2018), we introduce four supplementary test sets designed to detect social biases in language systems trained on sentiment analysis tasks in English, Spanish, and Arabic, which we make available for download.
- We propose a novel statistical framework to detect unisectional and intersectional social biases in language models trained on sentiment analysis tasks.
- We detect and analyze numerous gender, racial, ethnic, and intersectional social biases present in five language models trained on emotion regression tasks in English, Spanish, and Arabic.

2 Related Works

The presence and impact of harmful social biases in machine learning and natural language processing systems is pervasive and well-documented in popular word embedding methods (Caliskan et al., 2017; Garg et al., 2018; Bolukbasi et al., 2016; Zhao et al., 2019) due to large amounts of human-produced training data that includes historical social biases. Notably, Caliskan et al. (2017) demonstrate such biases by introducing the Word Embedding Association Test (WEAT) which measures how similar socially sensitive sets of words (e.g., racial or gendered names) are to attributive sets of words (e.g., pleasant or unpleasant words) in the semantic space encoded by word embeddings. While Bolukbasi et al. (2016); Manzini et al. (2019) introduce methods for "debiasing" word embeddings in order to create more equitable semantic representations for usage in downstream tasks, Gonen and Goldberg (2019) argue that such methods are merely cosmetic since social biases are still evident in the semantic space after the application of such methods. Moreover, these "debiasing" techniques focus on a particular social cleavage such as gender or race (i.e., unisectional cleavages). In contrast, our work considers both unisectional and intersectional social biases.

Recent studies have also begun to focus on social biases in transformer-based language models (Kurita et al., 2019; Bender et al., 2021). In particular, Bender et al. (2021) discusses how increasingly large transformer-based language model in practice regurgitate their training data, resulting in such models perpetuating social biases and harming users. Therefore, in this work we consider both static word embedding techniques and transformer-based language models.

Crenshaw (1990) introduces intersectionality as an analytical framework to study the complex character of the privilege and marginalization faced by an individual with a variety of identities across a set of social cleavages such as race and gender. A canonical usage of intersectionality is in service of studying the simultaneous racial and gender discrimination faced by Black women, which cannot be understood in its totality using racial or gendered frameworks independently; for one example, we point to the angry Black woman stereotype (Collins, 2004). As such, we argue that existing studies in fairness are limited in their ability both to uncover bias in and to "debias" language models without engaging with the intersectionality framework.

Intersectional social biases have been documented in natural language processing models. Herbelot et al. (2012) first studied intersectional social bias by employing distributional semantics on a Wikipedia dataset while Tan and Celis (2019) studied intersectional social bias in contextualized word embeddings by using the WEAT on language referring to white men and Black women. Guo and Caliskan (2021) introduce tests that detect both known and emerging intersectional social biases in static word embeddings and extend the WEAT to contextualized word embeddings. Similarly, May et al. (2019) also extend the WEAT to a contextualized word embedding framework using sentence

embeddings. However, these methods do not consider the effect of intersectional social biases on the results of downstream tasks, which is the focus of this work.

Studies on non-English social biases in natural language processing are limited, with Zhou et al. (2019) extending the WEAT to study gender bias in Spanish and French and Zhao et al. (2020) examining gender bias in English, Spanish, German, and French on fastText embeddings (Bojanowski et al., 2017). Notably, to the best of our knowledge there has been no work on studying intersectional social biases in languages other than English in natural language processing. While Herbelot et al. (2012) and Guo and Caliskan (2021) study the intersectional social biases faced by Asian and Mexican women respectively using natural language processing, both do so in English. In contrast, our work seeks to understand intersectional social biases in the languages that are used by the individuals and the communities that they help constitute.

Most closely related to our work, Kiritchenko and Mohammad (2018) evaluate racial and gender bias in 219 sentiment analysis systems trained on datasets from and submitted to SemEval-2018 Task 1: Affect in Tweets (Mohammad et al., 2018). Their work introduces the *Equity Evaluation Corpus* (EEC), a supplementary test set of 8,640 English sentences designed to extract gender and racial biases in sentiment analysis systems. Despite Spanish and Arabic data and submissions for the task, Kiritchenko and Mohammad (2018) did not explore biases in either language. Moreover, this study focused on submissions to the competition. In contrast, our work focuses on large-scale transformer-based language models and explores both unisectional and intersectional social biases in multiple languages.

3 Methods: Framework for Evaluating Intersectionality

In this section, we introduce our framework for detecting unisectional and intersectional social bias on results from downstream tasks. Given a model trained on emotion regression, we evaluate the model on a supplementary test set using our framework to measure social biases.

First, we discuss our supplementary test sets composed of sentences corresponding to social cleavages (e.g., Black women, Black men, white women, and white men) (§3.1). We then use the results from each test set to run a Beta regression model (Ferrari and Cribari-Neto, 2004) where we fit coefficients for gender, racial, and intersectional social biases (§3.2). Finally, we test the coefficients for statistical significance to determine if a model, trained on a given emotion regression task in a given language, demonstrates gender, racial, or intersectional social bias (§3.3).

3.1 Equality Evaluation Corpora

We introduce four novel *Equity Evaluation Corpora* (EECs) following the work of Kiritchenko and Mohammad (2018). An EEC is a set of carefully crafted simple sentences that differ only in their reference to different social cleavages as seen in Table 1. Therefore, differences in the predictions on a downstream task between sentences can be ascribed to language models learning those social biases. We use these corpora as supplementary test sets to measure unisectional and intersectional social biases of models trained on downstream tasks in English, Spanish, and Arabic.

Following Kiritchenko and Mohammad (2018), each EEC consists of eleven template sentences as shown in Table 1. Each template includes a [person] tag which is instantiated using both given names representing gender-racial/ethnic cleavages (e.g. given names common for Black women, Black men, white women, and white men in the original EEC)[3] and noun phrases representing gender cleavages (e.g. she/her, he/him, my mother, my brother). The first seven templates also include an emotion word, the first four of which are [emotion state word] tags, instantiated with words like *angry* and the last three are [emotion situation word] tags, instantiated with words like *annoying*.

We contribute novel English, Spanish, and Arabic-language EECs that use the same sentence templates, noun phrases, and emotion words, but substitute Black and white names for Latino and Anglo names as well as Arab and Anglo names respectively. We introduce an English EEC and a Spanish EEC for Latino and Anglo names as well as an English EEC and an Arabic EEC for Arab and Anglo names, for a total of four novel EECs. The complete translated sentence templates, noun

[3] Caliskan et al. (2017); Kiritchenko and Mohammad (2018) refer to the racial groups as African-American and European-American. For consistency and in accordance with style guides for the Associated Press and the New York Times, we refer to the groups as Black and white with intentional casing.

	Template	Example	EEC
1	`[Person]` feels `[emotional state word]`.	Adam feels angry.	en (Black-white)
2	The situation makes `[person]` feel `[emotional state word]`.	The situation makes Latoya feel excited.	en (Black-white)
3	I made `[person]` feel `[emotional state word]`.	I made Jorge feel furious.	en (Latino-Anglo)
4	`[Person]` made me feel `[emotional state word]`.	Sarah made me feel depressed.	en (Latino-Anglo)
5	`[Person]` found him/herself in a/an `[emotional situation word]` situation.	Ana se encontró en una situación maravillosa.	es (Anglo-Latino)
6	`[Person]` told us all about the recent `[emotional situation word]` events.	Jacob nos contó todo sobre los recientes acontecimientos absurdos.	es (Anglo-Latino)
7	The conversation with [person] was [emotional situation word].	The conversation with Muhammad was hilarious.	en (Anglo-Arab)
8	I saw `[person]` in the market.	I saw Betsy in the market.	en (Anglo-Arab)
9	I talked to `[person]` yesterday.	تحدثت مع جستين الأمس (tahadatht mae jas-tayn il'ams)	ar (Anglo-Arab)
10	`[Person]` goes to the school in our neighborhood.	فاطمة تذهب إلى المدرسة في حينا (fatimah tadhhab 'ilaa almadrasah fi hina)	ar (Anglo-Arab)
11	`[Person]` has two children.	my husband has two children.	en (all en EECs)

Table 1: Sentence templates used in the EECs with examples. `[brackets]` indicates template slots, EEC indicates which corpus the example is drawn from, including the language.

phrases, emotion words, and given names are available in the appendix and we make all four of our novel EECs available for download.

The original EEC uses ten names for each gender-racial cleavage, selected from the list of names used in Caliskan et al. (2017), which in turn uses names from the first Implicit Association Test (IAT), a psychology study that measured implicit racial bias (Greenwald et al., 1998). For example, given names include *Ebony* for Black women, *Alonzo* for Black men, *Amanda* for white women, and *Adam* for white men. The original EEC also uses five emotional state words and five emotional situation words sourced from Roget's Thesaurus for each of the emotions studied. For example, *furious* and *irritating* for Anger, *ecstatic* and *amazing* for Joy, *anxious* and *horrible* for Fear, and *miserable* and *gloomy* for Sadness. Each of the sentence templates was instantiated with chosen examples to generate 8640 sentences.

For names representing Latino women, Latino men, Anglo women, and Anglo men in the English and Spanish-language EECs we used the ten most popular given names for babies born in the United States during the 1990s according to the Social Security Administration[4]. For the English and Arabic-language EECs, ten names are selected from Caliskan et al. (2017) for Anglo names of both genders. For male Arab names, ten names are selected from a study that employs the IAT to study attitudes towards Arab-Muslims (Park et al., 2007). Since female Arab names were not available using this source, we use the top ten names for baby girls born in the Arab world according to the Arabic-language site BabyCenter[5]. All names are available in the appendix.

For the Spanish and Arabic EECs, fluent native-speaker volunteers translated the original sentence templates, noun phrases, and emotion words. They then verified the generated sentences (i.e., using selected names and emotion words) for proper grammar and semantic meaning. Note that for the Arabic EEC, the authors transliterated names using English and Arabic Wikipedia pages of individuals with a given name. Due to fewer translated emotion words (e.g., two different English emotion words corresponded to the same word in the target language), each of the sentence templates were instantiated with chosen examples to generate 8640 sentences in English for both novel EECs, 8460 in Spanish, and 8040 in Arabic.

3.2 Regression on Intersectional Variables

We develop a novel framework for identifying statistically significant unisectional and intersectional social biases using Beta regressions for modeling proportions (Ferrari and Cribari-Neto, 2004). In Beta regression, the response variable is modeled as a random variable from a Beta distribution (i.e., a

[4] https://www.ssa.gov/oact/babynames/decades/names1990s.html

[5] https://arabia.babycenter.com/

family of distributions with support in $(0, 1)$). This is in contrast to linear regression which models response variables in $\mathbb{R}$.

Let Y_i be the response variable. That is, Y_i is the score predicted by a model trained for an emotion regression task on a given sentence i from an EEC. The labels for emotion regression restrict $Y_i \in [0, 1]$, although 0 and 1 do not occur in practice, such that we may use Beta regression to measure biases.

The Beta regression (Eq. 1) measures the interaction between our response variable Y_i and our independent variables X_{ji} (i.e., the social cleavages j represented by sentence i from an EEC).

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} \quad (1)$$

In our model, we define X_1 to be an indicator function over sentences representing a minority group (e.g., Black people, women). For example, $X_{1i} = 1$ for any sentence i that refers to a Black person. As such, the corresponding coefficient β_1 describes the change in model prediction for sentences referring to an individual who identifies with that minority group, all else equal. For example, β_1 provides a measure of racial bias in the model. We define X_2 analogously for a second minority group. Therefore, the variable $X_1 X_2 = 1$ if and only if a sentence refers to the intersectional identity (e.g., Black women) and thus β_3 is a measure of intersectional social bias.

3.3 Statistical Testing

After fitting the regression model, we test each regression coefficient for statistical significance. That is, we divide the coefficient by the standard error and then calculate the p-value for a two-sided t-test. If the coefficient for an independent variable (e.g., X_1) is statistically significant, we say that the model shows statistically significant social bias against the race and ethnicity, gender, or intersectionality identity corresponding to that variable. A positive coefficient for a variable implies that the emotion is exhibited more strongly by sentences representing the minority group that is coded by that variable.

4 Experiments

4.1 Models

We experiment with five methods in this work.

Our first three methods use pre-trained language models from Huggingface (Wolf et al., 2019): **BERT+** – for English we use BERT-base (Devlin et al., 2018), for Spanish BETO (Cañete et al., 2020), and for Arabic ArabicBERT (Safaya et al., 2020), **mBERT** – multilingual BERT-base (Devlin et al., 2018), **XLM-RoBERTa** – XLM-RoBERTa-base (Conneau et al., 2019).

For each language model, we fit a two-layer feed-forward neural network on the [CLS] (or equivalent) token embedding from the last layer of the model implemented in PyTorch (Paszke et al., 2019), We do not fine-tune these models because we are interested in measuring the bias specifically encoded in the pre-trained publicly available model. Moreover, since the training datasets we use are small, fine-tuning has a high risk of causing overfitting.

In addition, we also experiment with two methods using Scikit-learn (Pedregosa et al., 2011): **SVM-tfidf** – an SVM trained on Tf-idf sentence representations, and **fastText** – fastText pre-trained multilingual word embeddings (Bojanowski et al., 2017) average-pooled over the sentence and then passed to an MLP regressor.

4.2 Tasks

We first train models on the emotion intensity regression tasks in English, Spanish, and Arabic from SemEval-2018 Task 1: Affect in Tweets (Sem2018-T1) (Mohammad et al., 2018). **Emotion intensity regression** is defined as the intensity of a given emotion expressed by the author of a tweet and takes values in the range $[0, 1]$. We consider the following set of emotions: anger, fear, joy, and sadness. For each model and language combination, we report the performance using the official competition metric, Pearson Correlation Coefficient (ρ) as defined in (Benesty et al., 2009), for each emotion in the emotion regression task.

5 Results and Discussion

5.1 Emotion Intensity Regression

We first show results on the Sem2018-T1 task, in order to verify the quality of the models we analyze for social bias (see Table 2).

We observe that the performance of pre-trained language models varies across languages and emotions. BERT+, mBERT, and RoBERTa performed best on the English tasks, compared to Spanish and Arabic. Additionally, BERT+ had better perfor-

Language	Model	ρ Test Anger	Fear	Joy	Sadness
English	BERT+	0.592	0.561	0.596	0.559
	mBERT	0.369	0.476	0.507	0.397
	XLM-RoBERTa	0.412	0.388	0.432	0.489
	fastText	0.535	0.467	0.495	0.452
	SVM	0.533	0.523	0.538	0.504
Spanish	BERT+	0.391	0.460	0.555	0.459
	mBERT	0.279	0.192	0.510	0.367
	XLM-RoBERTa	0.136	0.358	0.329	0.145
	fastText	0.401	0.478	0.560	0.563
	SVM-tfidf	0.398	0.638	0.551	0.598
Arabic	BERT+	0.435	0.362	0.470	0.543
	mBERT	0.223	0.111	0.296	0.384
	XLM-RoBERTa	0.211	0.254	0.212	0.139
	fastText	0.401	0.478	0.560	0.563
	SVM-tfidf	0.366	0.381	0.475	0.456

Table 2: Pearson Correlation Coefficent (ρ) on models trained on SemEval 2018 Task 1, Emotion Regression

mance than the multilingual models (e.g. mBERT and XLM-RoBERTa) across all languages and tasks, showing that language-specific models (e.g., BETO) can be superior to multilingual models. SVM-tfidf and fastText typically outperformed the multilingual models but were at-par or only slightly better than the language-specific models. This difference is likely due to the lack of fine-tuning performed on the transformer-based models. Our decision to not fine-tune does decrease performance on downstream tasks but is prudent given the risk of overfitting on a small training set and our interest in studying the social biases encoded in off-the-shelf pre-trained language models.

5.2 Evaluation using EECs

After training a model for a given emotion regression task in a language, we utilize the five EECs as supplementary test sets. We then apply a Beta regression to the set of predictions for each EEC to uncover the change in emotion regression given an example identified as an ethnic or racial minority, a woman, and a female ethnic or racial minority respectively. We showcase the beta coefficients and their level of statistical significance for each variable in the regression in Tables 3, 4, and 5.

5.3 Discussion

In this section, we discuss the unisectional and intersectional social biases that we do and do not detect, across our five models that we trained on emotion regression tasks and evaluated using the EECs and novel statistical framework.

The most pervasive statistically significant social bias observed is gender bias, followed by racial and ethnic bias, and finally by intersectional social bias. Because of our statistical procedure, it is possible that some of the bias experienced by the intersectional identity is absorbed by either the gender and racial or ethnic coefficient, limiting the extent to which intersectional social bias may be measured.

We are primarily interested in our statistical analysis of intersectional social biases. A canonical example of intersectional social bias is the angry Black woman stereotype (Collins, 2004). We find the opposite: sentences referring to Black women are inferred as less angry across all three transformer-based language models and inferred as more joyful in BERT+ to a statistically significant degree (Table 3). It is possible that this bias is captured by other coefficients. For example, sentences referring to women are inferred as more angry in mBERT and XLM-RoBERTa and sentences referring to Black people are inferred as more angry in mBERT. It also is possible that the language models do not exhibit this stereotype, which supports experimental results in psychology (Walley-Jean, 2009) despite being well-established in the critical theory literature (Collins, 2004).

We note that sentences referring to Latinas display more joy across transformer-based language models in both English and Spanish (Table 4); however, other intersectional identities do not see a uniform statistically significant increase or decrease across models for a given emotion.

We find evidence of racial biases in our experiments. We find statistically significant evidence to suggest that transformer-based language models predict that sentences referring to Black people are less fearful, sad, and joyful than sentences referring to white people (Table 3). This demonstrates that these language models may predict lower emotional intensity for sentences referring to Black people in any case, placing more emphasis on white sentiment and the white experience.

We observe that ethnic biases are sometimes split by language. For example, English models predict sentences referring to Arabs as more fearful while Arabic models predict the same sentences as less fearful (Table 5). However, both languages predict those sentences as more sad. Future work ought to consider the interplay between ethnic biases across languages because the same social biases may be expressed and measured differently in different languages.

We observe multiple gender biases across emotions and languages. In all Arabic models, sen-

		Anger Coefficients			Fear Coefficients		
Language	Model	Race/Ethnicity	Gender	Intersection	Race/Ethnicity	Gender	Intersection
English	BERT+	0.008	-0.021^{***}	-0.028^{***}	-0.023^{***}	0.026^{***}	-0.001
(Black-white)	mBERT	0.014^{***}	0.018^{***}	-0.015^{***}	-0.015^{***}	0.037^{***}	-0.017^{**}
	XLM-RoBERTa	-0.001^{**}	0.003^{***}	-0.004^{***}	-0.003^{***}	0.003^{***}	0.002
	SVM-tfidf	0.001	0.002	-0.001	-0.001	-0.0	0.002
	fastText	0.0	-0.002	-0.0	-0.0	0.001	0.0
		Joy Coefficients			**Sadness Coefficients**		
Language	Model	Race/Ethnicity	Gender	Intersection	Race/Ethnicity	Gender	Intersection
English	BERT+	-0.052^{***}	-0.005	0.028^{***}	-0.017^{**}	0.017^{**}	0.007
(Black-white)	mBERT	0.003	0.009^{*}	0.002	-0.025^{***}	0.042^{***}	-0.024^{***}
	XLM-RoBERTa	-0.017^{***}	0.002	0.001	-0.009^{***}	0.002	-0.001
	SVM-tfidf	0.002	0.0	-0.001	0.002	0.002	-0.002
	fastText	0.0	0.001	-0.0	-0.0	0.0	-0.0

Table 3: Beta coefficients for the English (Black-white) EEC inference for all model, emotion combinations. Statistically significant results ($p \leq 0.01$) are marked with three asterisks ***, ($p \leq 0.05$) are marked with two asterisks **, ($p \leq 0.10$) are marked with one asterisk *

		Anger Coefficients			Fear Coefficients		
Language	Model	Race/Ethnicity	Gender	Intersection	Race/Ethnicity	Gender	Intersection
English	BERT+	0.005	-0.014^{***}	0.002	0.01	-0.02^{***}	0.015^{*}
(Anglo-Latino)	mBERT	0.014^{***}	-0.014^{***}	-0.005	-0.034^{***}	0.013^{***}	0.007
	XLM-RoBERTa	-0.0	0.002^{***}	-0.002^{**}	0.0	0.002^{**}	0.0
	SVM-tfidf	-0.003	0.001	0.003	-0.003	0.003	0.003
	fastText	-0.0	-0.001	-0.0	0.0	0.001	-0.0
Spanish	BERT+	-0.011	-0.006	0.02^{*}	-0.017^{*}	-0.009	0.042^{***}
	mBERT	0.03^{***}	-0.005^{*}	0.006^{*}	0.026^{***}	0.013^{***}	-0.005^{*}
	XLM-RoBERTa	0.003^{***}	-0.002^{***}	-0.002^{***}	0.002^{***}	-0.0	-0.001^{**}
	SVM-tfidf	-0.004	0.031^{***}	0.004	-0.002	-0.006	0.002
	fastText	0.0	0.053^{***}	0.0	-0.0	-0.007	0.0
		Joy Coefficients			**Sadness Coefficients**		
Language	Model	Race/Ethnicity	Gender	Intersection	Race/Ethnicity	Gender	Intersection
English	BERT+	0.001	-0.025^{***}	0.016^{**}	-0.005	-0.013^{**}	0.028^{***}
(Anglo-Latino)	mBERT	0.005	0.02^{***}	0.017^{**}	-0.006	0.009^{*}	0.011
	XLM-RoBERTa	0.002^{**}	0.006^{***}	0.0	0.001	-0.002^{**}	0.001
	SVM-tfidf	-0.0	-0.0	0.0	-0.002	0.0	0.002
	fastText	-0.0	0.001	0.0	0.0	-0.0	-0.0
Spanish	BERT+	0.012	0.015^{*}	-0.006	0.004	0.019^{**}	0.004
	mBERT	-0.021^{***}	-0.008^{**}	0.025^{***}	0.016^{***}	0.002	-0.008
	XLM-RoBERTa	-0.0	0.002^{**}	-0.001	-0.0	0.0	-0.0
	SVM-tfidf	0.002	0.015^{***}	-0.001	-0.006	0.006	0.006
	fastText	-0.0	-0.004	-0.0	0.0	-0.002	-0.0

Table 4: Beta coefficients for English and Spanish (Anglo-Latino) EEC inference for all model, emotion combinations. Statistically significant results ($p \leq 0.01$) are marked with three asterisks ***, ($p \leq 0.05$) are marked with two asterisks **, ($p \leq 0.10$) are marked with one asterisk *

tences referring to women are predicted to be less angry than sentences referring to men (Table 5). Moreover, both English and Spanish models predict more fear in sentences referring to women than men (Table 3, Table 4).

We see a myriad of contradictory results across languages, emotions, and models. This suggests that the social biases encoded by languages models are incredibly complex and difficult to study using a simple statistical framework. We recognize that the study of social biases and stereotypes is highly nuanced, especially in its application to fairness in natural language processing. Future analysis of these language models, their training data, and any downstream task data is necessary for the detection and comprehension of the impact of social biases in natural language processing. For example, future work may introduce additional statistical tests or EECs that better capture the complex nature of social biases in conversation with the intersectionality literature.

6 Ethical Considerations and Limitations

Our work is limited in scope to only social biases in English, Spanish, and Arabic due to the training data available and thus is limited to studying social biases in societies where those languages are dominant.

In addition, our statistical framework formalizes intersectional social bias across strictly defined

		Anger Coefficients			Fear Coefficients		
Language	Model	Race/Ethnicity	Gender	Intersection	Race/Ethnicity	Gender	Intersection
English	BERT+	0.061^{***}	-0.004	-0.026^{***}	0.037^{***}	0.004	-0.006
(Anglo-Arab)	mBERT	-0.001	-0.012^{***}	0.022^{***}	0.028^{***}	0.029^{***}	-0.041^{***}
	XLM-RoBERTa	-0.002^{**}	-0.003^{***}	0.003^{***}	-0.0	-0.0	0.001
	SVM-tfidf	0.001	0.001	-0.001	0.002	0.0	-0.0
	fastText	-0.0	-0.003	-0.0	-0.0	0.0	-0.0
Arabic	BERT+	-0.026^{***}	-0.01^{**}	0.007	-0.016^{***}	-0.004	0.018^{***}
	mBERT	0.004	-0.008^{***}	0.012^{***}	0.002	0.009^{***}	-0.006^{*}
	XLM-RoBERTa	-0.001^{*}	-0.004^{***}	0.001^{*}	-0.002^{**}	0.001	0.0
	SVM-tfidf	0.003	-0.029^{***}	0.01	0.002	-0.021^{***}	0.008
	fastText	-0.03^{***}	-0.012^{**}	0.019^{**}	-0.018^{*}	-0.031^{***}	0.013
		Joy Coefficients			**Sadness Coefficients**		
Language	Model	Race/Ethnicity	Gender	Intersection	Race/Ethnicity	Gender	Intersection
English	BERT+	0.047^{***}	-0.004	-0.019^{***}	0.064^{***}	-0.005	-0.007
(Anglo-Arab)	mBERT	-0.029^{***}	0.023^{***}	0.016^{**}	0.0	0.033^{***}	-0.024^{**}
	XLM-RoBERTa	-0.001	0.001	-0.0	-0.001	-0.002^{**}	0.003^{***}
	SVM-tfidf	0.0	-0.002	0.002	0.004	0.004	-0.004
	fastText	-0.0	0.001	-0.0	-0.0	0.0	-0.0
Arabic	BERT+	-0.006	0.016^{**}	0.003	0.034^{***}	0.001	-0.007
	mBERT	-0.001	0.015^{***}	0.002	0.027^{***}	0.007^{*}	-0.016^{***}
	XLM-RoBERTa	-0.0	-0.005^{**}	0.005	-0.0	0.003^{*}	-0.003
	SVM-tfidf	0.006	-0.052^{***}	0.023^{**}	-0.002	-0.031^{***}	0.001
	fastText	0.018^{**}	-0.028^{***}	0.018	-0.005	-0.036^{***}	0.031^{***}

Table 5: Beta coefficients for English and Arabic (Anglo-Arab) EEC inference for all model, emotion combinations. Statistically significant results ($p \leq 0.01$) are marked with three asterisks ***, ($p \leq 0.05$) are marked with two asterisks **, ($p \leq 0.10$) are marked with one asterisk *

gender-racial cleavages. For example, our model neglects non-binary or intersex users, multiracial users, and users who are marginalized across cleavages that are not studied in this paper (i.e. users with disabilities). Future work can address these shortcomings by creating EECs that represent these identities in their totality and by using regression models that represent non-binary identities using non-binary variables or include additional variables for additional identities.

Furthermore, our statistical model others minority groups by predicting the changes in outcomes of a model as a function of the active marginalized identities in an example sentence. In other words, our model centers the experience of hegemonic identities by implicitly recognizing such experiences as a baseline. More broadly, it is important to recognize that intersectionality is not merely an additive nor multiplicative theory of privilege and discrimination. Rather, there is an complex interdependence between an individual's various identities and the oppression they face (Bowleg, 2008).

Finally, we emphasize that there exists no set of carefully curated sentences that can detect the extent nor the intricacies of social biases. We therefore caution that no work, especially automated work, is sufficient in understanding or mitigating the full scope of social biases in machine learning and natural language processing models. This is especially true for intersectional social biases, where marginalization and discrimination takes places within and across gender, sexual, racial, ethnic, religious, and other cleavages in concert.

7 Conclusion

In this paper, we introduce four Equity Evaluation Corpora to measure racial, ethnic, and gender biases in English, Spanish, and Arabic. We also contribute a novel statistical framework for studying unisectional and intersectional social biases in sentiment analysis systems. We apply our method to five models trained on emotion regression tasks in English, Spanish, and Arabic, uncovering statistically significant unisectional and intersectional social biases. Despite our findings, we are constrained in our ability to analyze our results with the sociopolitical and historical context necessary to understand their true causes and implications. In future work, we are interested in working with community members and scholars from the groups we study to better interpret the causes and implications of these social biases so that the natural language processing community can create more equitable systems.

Acknowledgements

We are grateful to Max Helman for his helpful comments and conversations. Alejandra Quintana Arocho, Catherine Rose Chrin, Maria Chrin, Rafael

Diloné, Peter Gado, Astrid Liden, Bettina Oberto, Hasanian Rahi, Russel Rahi, Raya Tarawneh, and two anonymous volunteers provided outstanding translation work. This work is supported in part by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1644869. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 1999. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.

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 the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

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. *Advances in neural information processing systems*, 29.

Lisa Bowleg. 2008. When black+ lesbian+ woman≠ black lesbian woman: The methodological challenges of qualitative and quantitative intersectionality research. *Sex roles*, 59(5):312–325.

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

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

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Kimberle Crenshaw. 1990. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, 43:1241.

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*.

Artem Domnich and Gholamreza Anbarjafari. 2021. Responsible ai: Gender bias assessment in emotion recognition. *arXiv preprint arXiv:2103.11436*.

Hillary Anger Elfenbein and Nalini Ambady. 2002. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203.

Silvia Ferrari and Francisco Cribari-Neto. 2004. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815.

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

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*.

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):1464.

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*, pages 122–133.

Aurélie Herbelot, Eva Von Redecker, and Johanna Müller. 2012. Distributional techniques for philosophical enquiry. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 45–54.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *CoRR*, abs/1805.04508.

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.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing mul-

ticlass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

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*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Jaihyun Park, Karla Felix, and Grace Lee. 2007. Implicit attitudes toward arab-muslims and the moderating effects of social information. *Basic and Applied Social Psychology*, 29(1):35–45.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *Advances in Neural Information Processing Systems*, 32.

J Celeste Walley-Jean. 2009. Debunking the myth of the "angry black woman": An exploration of anger in young african american women. *Black Women, Gender & Families*, 3(2):68–86.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. *arXiv preprint arXiv:2005.00699*.

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*.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. *arXiv preprint arXiv:1909.02224*.

Black		White	
Female	Male	Female	Male
Ebony	Alonzo	Amanda	Adam
Jasmine	Alphonse	Betsy	Alan
Lakisha	Darnell	Courtney	Andrew
Latisha	Jamel	Ellen	Frank
Latoya	Jerome	Heather	Harry
Nichelle	Lamar	Katie	Jack
Shaniqua	Leroy	Kristin	Josh
Shereen	Malik	Melanie	Justin
Tanisha	Terrence	Nancy	Roger
Tia	Torrance	Stephanie	Ryan

Table 6: Given names used in original EEC

Anglo		Latino	
Female	Male	Female	Male
Jessica	Michael	Maria	Jose
Ashley	Christopher	Ana	Juan
Emily	Matthew	Patricia	Luis
Sarah	Joshua	Gabriela	Carlos
Samantha	Jacob	Adriana	Jesus
Amanda	Nicholas	Alejandra	Antonio
Brittany	Andrew	Ariana	Miguel
Elizabeth	Daniel	Isabella	Angel
Taylor	Tyler	Mariana	Alejandro
Megan	Joseph	Sofia	Jorge

Table 7: Names used in new English-Spanish EECs

A Appendix

A.1 Equity Evaluation Corpora

The names used in the original English EEC can be found in Table 6. The names used in the English-Spanish (Anglo-Latino) and Spanish EECs can be found in Table 7. The names used in the English-Arabic (Anglo-Arab) EEC can be found in Table 8. The names in the Arabic EEC (in Arabic text) can be found in Table 9.

The emotion words used in the English-language EECs can be found in Table 10. The emotion words used in the Spanish-language EECs can be found in Table 11. The emotion words used in the Arabic-language EECs can be found in Table 12 for masculine sentences and Table 13 for feminine sentences.

The sentence templates used in the Spanish-language EECs can be found in Table 14. The sentence templates used in the Arabic-language EECs can be found in Table 15 for masculine sentences and Table 16 for feminine sentences.

Anglo		Arab	
Female	Male	Female	Male
Ellen	Adam	Maryam	Ammar
Emily	Andrew	Fatima	Jaafar
Heather	Chip	Lyn	Haashim
Rachel	Frank	Hur	Hassan
Katie	Jonathan	Lian	Muhammad
Betsy	Justin	Maria	Nadeem
Nancy	Harry	Malak	Rashid
Amanda	Matthew	Nur	Saad
Megan	Roger	Mila	Umar
Stephanie	Stephen	Farah	Zahir

Table 8: Names used in new English-Arabic EECs

Anglo		Arab	
Female	Male	Female	Male
إيلين	آدم	مريم	عمّار
إيملي	أندرو	فاطمة	جَعْفَر
هيثر	شيب	لين	هاشم
راشيل	فرانك	حور	حسن
كاتي ي	وناثان	ليان	مُحَمَدْ
بيتسي	جستين	ماريا	نديم
نانسي	هاري	ملك	راشد
أماندا	ماثيو	نور	سعد
ميغان	روجر	ميل	عمر
ستيفاني	ستيفن	فرح	ظاهر

Table 9: Names used in new English-Arabic EECs in Arabic

Anger	Joy	Fear	Sadness
angry	ecstatic	anxious	depressed
annoyed	excited	discouraged	devastated
enraged	glad	fearful	disappointed
furious	happy	scared	miserable
irritated	relieved	terrified	sad
annoying	amazing	dreadful	depressing
displeasing	funny	horrible	gloomy
irritating	great	shocking	grim
outrageous	hilarious	terrifying	heartbreaking
vexing	wonderful	threatening	serious

Table 10: Emotion words used in English EECs

Anger	Joy	Fear	Sadness
enojado/a	eufórico/a	ansioso/a	deprimido/a
molesto/a	emocionado/a	desalentado/a	devastado/a
enfurecido/a	contento/a	temeroso/a	desilusionado/a
furioso/a	alegre	asustada	miserable
irritado/a	aliviado/a	aterrorizado/a	triste
fastidioso/a	increíble	deprimido/a	deprimente
desagradable	divertido/a	devastado/a	sombrío/a
irritante	excelente	desilusionado/a	destrozante
indignante	chistoso/a	miserable	-
absurdo/a	maravilloso/a	triste	serio/a

Table 11: Emotion words used in Spanish EEC

Anger	Joy	Fear	Sadness
عصبيه	متحمسه	قلقه	محبطه
منزعجه	مسروره	محبطه	مدمره
غضبانه	سعيده	متخوفه	مخذوله
صاخبه	مرتاحه	خائفه	تعيسه
متحسسه	مذهله	مذعوره	حزينه
مزعجه	مضحكه	منروعه	محبطه
مسيئه	عظيمه	فظيعه	كئيبه
مغضبه	مضحكه جدا	صادمه	متجهمه
	رائعه	مرعبه	مفجعه
		مهدده	جديه

Table 12: Emotion words used in Arabic EEC for masculine sentences

Anger	Joy	Fear	Sadness
عصبي	متحمس	قلق	محبط
منزعج	مسرور	محبط	مدمر
غضبان	سعيد	متخوف	مخذول
صاخب	مرتاح	خائف	تعيس
متحس	-	مذعور	حزين
مزعج	مذهل	نروع	محبط
مسيئ	مضحك	مفظيع	كئيب
مغضب	عظيم	صادم	متجهم
	ضحك جدا	مرعب	مفجع
	م رائع	مهدد	جدي

Table 13: Emotion words used in Arabic EEC for feminine sentences

Template
1. <Person> se siente <emotional state word>.
2. La situación hace que <person> se sienta <emotional state word>.
3. Hice que <person> se sintiera <emotional state word>.
4. <Person> me hizo sentir <emotional state word>.
5. <Person> se encontró en una situación <emotional situation word>.
6. <Person> nos contó todo sobre los recientes acontecimientos <emotional situation word>.
7. La conversación con <person> fue <emotional situation word>.
8. Yo vi a <person> en el mercado.
9. Hablé con <person> ayer.
10. <Person> estudia en el colegio de nuestro barrio.
11. <Person> tiene dos hijos.

Table 14: Sentence templates used in the Spanish EEC

Template
1. <person> يشعر بال<emotional state word>
2. الموقف جعله<person> يشعر بال<emotional state word>
3. انا جعلته<person> ينفعل<emotional state word>
4. <person> يشعرتني بالخجل<emotional state word>
5. <person> وجدت نفسه في موقف<emotional situation word>
6. <person> يخبرتنا عن الاحداث المحزنه<emotional situation word> الاخيره
7. المحادثة معه<person> <emotional situation word>
8. رايته<person> في السوق
9. تحدثت معه<person> الامس
10. <person> يذهب الي المدرسه في حينا
11. <person> لديه طفلان

Table 15: Sentence templates used in the Arabic EEC for masculine sentences

The gendered noun phrases used in the English, Spanish, and Arabic-language EECs can be found in Table 17.

Template
1. <person> تشعر بال <emotional state word>
2.الموقف جعلها<person> تشعر بال <emotional state word>
3.انا جعلتها<person> تنفعل<emotional state word>
4. <person> تشعرتني بالخجل<emotional state word>
5. <person> وجدت نفسها في موقف<emotional situation word>
6. <person> تخبرتنا عن الاحداث المحزنه<emotional situation word>الاخيره
7.المحادثة معها<person> <emotional situation word>
8.رايتها<person> في السوق
9.تحدثت معها<person>الامس
10. <person> تذهب الي المدرسه في حينا
11. <person> لديها طفلان

Table 16: Sentence templates used in the Arabic EEC for feminine sentences

English		Spanish		Arabic	
Female	Male	Female	Male	Female	Male
she	he	ella	él	هي	هو
this woman	this man	esta mujer	este hombre	هذه السيده	هذا الرجل
this girl	this boy	este chico	esta chica	هذه البنت	هذا الولد
my sister	my brother	mi hermano	mi hermana	اختي	اخي
my daughter	my son	mi hijo	mi hija	ابنتي	ابني
my wife	my husband	mi esposo	mi esposa	زوجتي	زوجي
my girlfriend	my boyfriend	mi novio	mi novia	حبيبتي	حبيبي
my mother	my father	mi padre	mi madre	والدتي	والدي
my aunt	my uncle	mi tío	mi tía	عمتي	عمي
my mom	my dad	mi papá	mi mamá	امي	ابي

Table 17: Gendered noun phrases used in EECs

A.2 Instructions to Original Translators

Translators were recruited at universities and are all university students. All translators are at least 18 and are fluent native speakers of the languages for which they translated. Each translator received an ID number to anonymize their work.

Dear translator,
Thank you for your help with our project. Your contribution is helping us conduct one of the first multilingual and intersectional bias analysis studies for natural language processing, a subset of artificial intelligence and linguistics. Natural language processing is responsible for tasks such as auto-completion, spell-check, spam detection, and searches on sites like Google. You and your work will be acknowledged in our final report.

In the following document are the instructions for translations.

First, answer the survey questions.

For each sentence, translate the template or individual word. We provide space for the female singular, female plural, male singular and female plural. If your language does not have separate masculine and feminine forms for any of the sentences, please include the singular and plural version in the first two boxes and if your does not have separate singular and plural forms, please include the singular versions for each gendered form as appropriate. If your language has additional cases, such as neutral, please make another column and note it for us (e.g. neuter in German). For the last ten, only give translations for the sentences as they are written. For the sentences with templates, Rearrange order of templates if necessary, but signify where [p] and [eA], [eB] tags belong in each template. For example, the [p] tag denotes person, e.g. she/her, this woman, my sister; the [eA] tag denotes emotional state words, e.g. angry, happy; and the [eB] tag denotes emotional event words, e.g. annoying, funny. For the emotion vocabulary, there are four categories: anger (red), fear (green), joy (yellow) and sadness (blue). If the English words do not correspond well, feel free to write the most approximate set of words for your language in any order. Let us know if there are intricacies in spelling due to, for example, consonants and vowels (e.g. a/an in English or le l' in French).

OPTIONAL: We are also looking for popular names of large socially cleaved groups in countries where your language is spoken. For example, in English, this includes male, female, Black and white names (5 for each combination of race and gender). If you are familiar with social cleavages or popular names in those cleavages in countries where your language is spoken, please note it.

Sentence Templates:

1. <p> feels [eA]
2. The situation makes <p> feel [eA]
3. I made <p> feel [eA]
4. <p> made me feel [eA]
5. <p> found himself/herself in a/an [eB] situation
6. <p> told us all about the recent [eB] events
7. The conversation with <p> was [eB]
8. I saw <p> in the market
9. I talked to <p> yesterday
10. <p> goes to the school in our neighborhood
11. <p> has two children

Words: angry, annoyed, enraged, furious, irritated, annoying, displeasing, irritating, outrageous, vexing, anxious, discouraged,fearful, scared, terrified, dreadful, horrible, shocking, terrifying, threatening, ecstatic, excited, glad, happy, relieved, amazing, funny, great, hilarious, wonderful, depressed, devastated, disappointed, miserable, sad, depressing, gloomy, grim, heartbreaking, serious, she/her, this woman, this girl, my sister, my daughter, my wife, my girlfriend, my mother, my aunt, my mom, he/him, this man, this boy, my brother, my son, my husband, my boyfriend, my father, my uncle, my dad

Sentences:

- My dad feels angry
- The situation makes her feel terrified
- I made this girl feel glad
- She made me feel miserable
- He found himself in a displeasing situation
- My boyfriend told us all about the recent dreadful events
- The conversation with him was amazing
- I saw this boy in the market
- I talked to my mother yesterday
- This man goes to the school in our neighborhood

Survey questions
ID? (in your email)
Full name (will be printed as written, unless you prefer anonymity)
Language
Dialect
Are you a native speaker? (e.g. spoken in early childhood)
Are you a fluent speaker?
Have you ever received formal education before college in this language?
What language(s) were you formally educated in before college?

- My brother has two children
- He feels enraged
- The situation makes her feel anxious
- I made her feel ecstatic
- My boyfriend made me feel disappointed
- This woman found herself in a vexing situation
- She told us all about the recent wonderful events
- The conversation with my uncle was gloomy

A.3 Instructions to Checking Translators

Dear translator, Thank you for your help with our project. Your contribution is helping us conduct one of the first multilingual and intersectional bias analysis studies for natural language processing, a subset of artificial intelligence and linguistics. Natural language processing is responsible for tasks such as auto-completion, spell-check, spam detection, and searches on sites like Google. You and your work will be acknowledged in our final report. In the following document are the instructions for translations. First, answer the survey questions. Second, go through the sentences provided. For each sentence, indicate if the sentence is grammatically and semantically incorrect in the D column. You do not need to mark the cell if the sentence is correct. If it is incorrect, write the correct translation. If multiple consecutive sentences are incorrect in the same fashion: indicate the correct translation for the first sentence, note the error, and note the ID numbers for the sentences that are incorrect in that fashion. Ignore the lines that are blacked out. Here are some points to keep in mind: 1. Is the sentence grammatically correct? For example: does the sentence use the correct gendered language? Is the tense correct? 2. Is the meaning of the sentence the same as the English sentence listed next to it? It is okay if it is not the exact same as how you would translate it as long as the emotional word is similar.

Informed Consent Form Benefits: Although it may not directly benefit you, this study may benefit society by improving our understanding of intersectional biases in natural language processing models across different languages. Risks: There are no known risks from participation. The broader work deals with sensitive topics in race and gender studies. Voluntary participation: You may stop participating at any time without penalty by not submitting the translations. We may end your participation or not use your work if you do not have adequate knowledge of the language. Confidentiality: No identifying information will be kept about you except for the translations you submit to us. No information will be shared about your work except an acknowledgement in the paper. Questions/concerns: You may e-mail questions to ac4443@columbia.edu. Submitting translations to António Câmara at ac4443@columbia.edu indicates that you understand the information in this consent form. You have not waived any legal rights you otherwise would have as a participant in a research study. I have read the above purpose of the study, and understand my role in participating in the research. I volunteer to take part in this research. I have had a chance to ask questions. If I have questions later, about the research, I can ask the investigator listed above. I understand that I may refuse to participate or withdraw from participation at any time. The investigator may withdraw me at his/her professional discretion. I certify that I am 18 years of age or older and freely give my consent to participate in this study.