Understanding Social Biases Behind Location Names in Contextual Word Embedding Models

Fangsheng Wu, Mengnan Du, Chao Fan, Ruixiang Tang, Yang Yang, Ali Mostafavi,
and Xia Hu, *Member, IEEE*

***Abstract*—Embeddings of textual data containing location names (e.g., social media posts) have essential applications in various contexts such as marketing and disaster management. In these downstream implementations, social biases behind location names are highly prone to introduce unfair results through their embeddings; for example, emergent text messages with swapped location names might result in varied rescue responses. Hence, it is critical to address social biases encoded in location names and to seek its mitigation. Prevalent works addressing biases in embeddings mainly focus on individual attributes like gender or ethnicity. Yet, a large number of social attributes behind location names (e.g., income level and population density) makes it challenging to originate the source of biases. Existing mitigation methods based on finding attribute subspaces cannot be simply applied to address social biases. Moreover, bias mitigation tends to simultaneously remove necessary semantics from embeddings, making it difficult to achieve a balance between mitigation performance and semantics retention. In this article, we first employ the concept of counterfactual fairness to investigate the social biases encoded in training data. Then, we quantify the biases in the contextual embeddings (BERT and ELMo). We report a high correlation between biases in the training data and embeddings. Next, we introduce a novel bias mitigation algorithm that customizes bias representations for any location names. The method yields debiased location name vectors for various social attributes simultaneously. The proposed algorithm achieves a better mitigation performance on overall attributes compared with a prevalent postprocessing method, while maintaining correctness by retaining semantic information.**

***Index Terms*—Contextual word embeddings, fairness, mitigation, social attributes.**

Manuscript received February 7, 2021; revised July 7, 2021; accepted August 11, 2021. This work was supported in part by the National Science Foundation under Grant CMMI-1846069 (CAREER), in part by the Amazon Web Services (AWS) Machine Learning Award, and in part by the National Academies' Gulf Research Program Early-Career Research Fellowship. *(Corresponding author: Fangsheng Wu.)*

Fangsheng Wu, Mengnan Du, and Yang Yang are with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, USA (e-mail: wufangsheng@tamu.edu; dumengnan@tamu.edu; yangyangsandy@tamu.edu).

Chao Fan and Ali Mostafavi are with the Department of Civil Engineering, Texas A&M University, College Station, TX 77843, USA (e-mail: chfan@tamu.edu; amostafavi@civil.tamu.edu).

Ruixiang Tang and Xia Hu were with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, USA. They are now with the Department of Computer Science, Rice University, Houston, TX 77005, USA (e-mail: ruixiang.tang@rice.edu; xia.hu@rice.edu).

Digital Object Identifier 10.1109/TCSS.2021.3106003

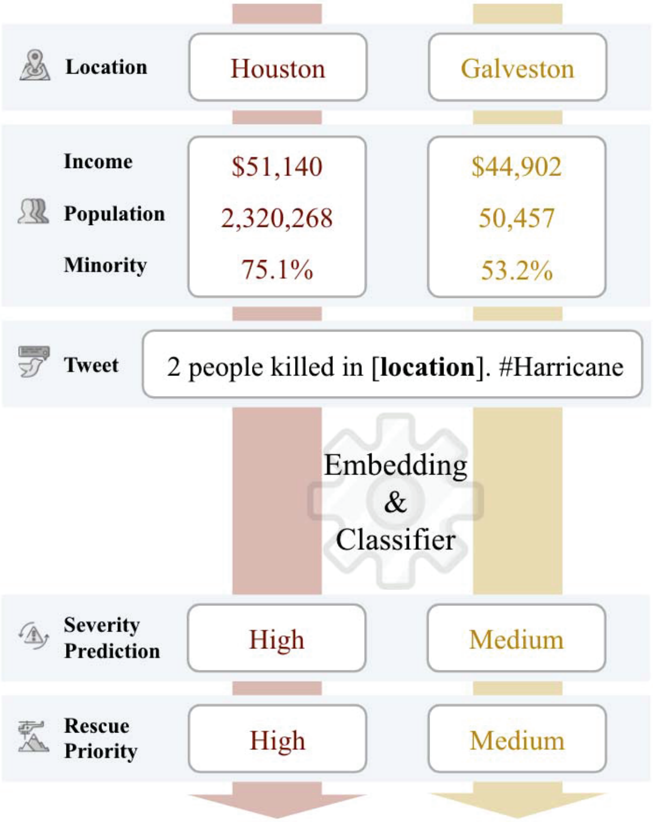


Fig. 1. Demonstration of potential unfair outcomes for location name bias in embeddings.

I. Introduction

Contextual embedding models [1]–[4] have been utilized as the backbones of a vast majority of natural language processing (NLP) techniques to achieve better performance.

In the implementations of contextual word embedding models, social biases in location names bring serious threats to the fairness of a large amount of related downstream tasks. For instance, neighborhoods can be classified to have less or greater severity of damages in disasters based on social media posts just for their different names even when occurring in the same damage report, causing unfair situation awareness and relief operations [5]–[8]. A social-media-based disaster awareness predictor might discriminate locations with certain values of social attributes [5]. As shown in Fig. 1, the probability predicted to have people in trouble in different locations might vary significantly when the input only varies in the location names. Other numerous sentiment-analysis-based applications relying on word embeddings, e.g., product recommendation, decision making, intelligent customer services, may face the same biases and produce biased outcomes [9]–[12].

Current efforts related to originating biases in contextual embeddings mainly focus on gender biases and ethnic biases [13]–[17]. Accordingly, prevalent bias mitigation algorithms are mostly built based on the outcomes of finding fundamental subspaces for each gender or ethnic group [13], [18], [19]. Social attributes behind locations that are potential to bring biases are numerous [20] (e.g., population, income level, education level, or aging level). While some existing algorithms bring solutions to finding subspaces for a single binary attribute like gender or a single multilabel attribute like ethnicity, they cannot be simply implemented in mitigating biases related to social attributes.

The challenges in addressing the bias problems encoded in location names are threefold. First, there are numerous social attributes among which we need to originate the attributes introducing social biases to the embeddings of location names. However, prevalent bias mitigation algorithms are mostly built based on the outcomes of finding subspaces for each attribute [13], [18]. Subsequently, the second challenge is that the work is arduous and time-consuming to determine the subspaces for the numerous attributes in order to employ the existing bias mitigation methods. Third, an ideal bias mitigation method should efficiently remove biases for various attributes, and at the same time retain the necessary semantic information [21]–[24] for downstream tasks. However, the semantics of embeddings could be a tradeoff for bias mitigation. Therefore, it is challenging to achieve the balance between mitigation performance and semantic retention.

In this article, we propose a unified framework for detecting and mitigating social biases in location names, which are represented by separable embeddings for location names from different social groups. First, in order to provide a more formalized way to characterize and categorize social attributes, we employ the definitions of social attributes determined by Centers for Disease Control and Prevention (CDC) Social Vulnerability Index [20]. Based on the definitions, we employ the concept of counterfactual fairness to untangle the relationships between the embeddings and the extent of the bias for all the chosen attributes. Second, to relieve the time-consuming process for finding subspaces for each attribute, we develop a novel mitigation method that mitigates social biases of location names simultaneously for all determined social attributes. The proposed method modifies embeddings on dimensions with respect to the embedding results of a perturbed sentence corpus based on posts obtained through Twitter PowerTrack application programming interface (API). This approach significantly reduces statistical biases in the embeddings in order to yield fair outcomes for downstream tasks depending on location names. Third, in order to examine the maintenance of semantic information after embeddings are debiased, we employ benchmark datasets for concept categorization tasks. Evaluations conducted on bias removal and semantics retention indicate that the proposed method achieves significant debiasing performance for various social attributes (lower WEAT, i.e., Word-Embedding-Association-Test [25] test effect sizes and lower support vector machines (SVMs) classifier performances) as well as semantic retention performance (stable purity scores for concept categorization tasks). We conduct the experiments on BERT (BERT-Base, Cased) [26] and ELMo (Original) [2] as two state-of-the-art models. The main contributions of this article can be summarized as follows.

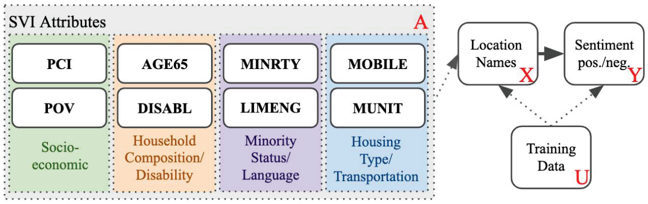


Fig. 2. Counterfactual fairness diagram. Each colored rectangle represents the summary theme. The solid arrow represents the prediction directions. Red letters correspond to variables in (1).

1) *Evaluation of social biases* encoded by the two contextual embedding models in location names.
2) *A novel customized mitigation method* that simultaneously mitigates biases for all attributes.
3) *A case study* conducted on the core classifier module of a recently developed pipeline for disaster situation awareness utilizing tweets during crisis [5].

The rest of this article is organized as follows. Section II presents the definition of the social biases and the methodologies to detect them in both the training data and the embedding output of two popular word embedding models. In Section III, we introduce the proposed bias mitigation approach. In Section IV, we conduct experiments to verify the effectiveness and the correctness of the proposed approach. Section V includes the implementation of the proposed approach on a state-of-the-art disaster detection model. Related works are presented in Section VI. Concluding remarks and future works are presented in Section VII.

II. Bias Detection

In this section, we first formalize the definition of the social biases discussed in this article and the problem of social biases detection in location names. Based on the definition of social biases, we measure the occurrence disparity of location names in training datasets for BERT and ELMo. Next, we quantitatively measure the biases exhibiting in the output of embeddings. We observe a positive correlation between the occurrence disparity of location names in training data and the biases with regard to their social attributes in the embeddings.

A. Problem Statement

1) Definition of Social Biases: We define a social biased embedding as an embedding that leads to high variation in the probabilities of a classification result for the same label (e.g., positive or negative sentiments in this article), given location names from different social groups (with varying

TABLE I

Socio-Demographic Attributes Selected for Analysis in This Article. These Attributes Are Commonly Used for Determining Social Vulnerability and Hence Are Appropriate for Examining Location Name Biases

Summary Theme	Abbreviation	Meaning
Socioeconomic	**PCI**	Per capita income estimate
	POV	Percentage of persons below poverty estimate
Household Composition/Disability	**AGE65**	Percentage of persons aged 65 and older estimate
	DISABL	Percentage of civilian noninstitutionalize d population with a disability estimate
Minority Status/Language	**MINRTY**	Percentage of minority (all persons except white, nonHispanic) estimate
	LIMENG	Percentage of persons (age 5+) who speak English ”less than well” estimate
Housing Type/Transportation	**MOBILE**	Percentage of mobile homes estimate
	MUNIT	Percentage of housing in structures with 10 or more units estimate

compositions of social attributes). The mathematical expression is presented in the following paragraph.

2) Counterfactual Fairness: Counterfactual fairness focuses on the counterfactual of situations with different settings of social attributes [27]. We model social biases based on the disparity of prediction results between the different situations. For instance, one may be concerned whether embedding results of the name of a neighborhood at below-average income level would be different if, in counterfactual settings, it is a neighborhood at above-average income level. The evaluation of counterfactual fairness provides evidence of the group of social attributes that are introducing biases and the extent to which these biases are introduced.

We obtain eight social attributes for all counties over the United States from the Social Vulnerability Index (SVI) dataset [20]. Two attributes are selected from each of the four themes provided by the dataset as shown in Table I.

With the perspective from counterfactual fairness [28], we arrange these social attributes together as the protected attribute set A. The location names form the input set X and the sentiment polarities form the output set Y to be predicted. We define the polarities or the polarity groups of a certain attribute as the set of samples that have extreme (within top or bottom certain percentile) values under this attribute. The training data of the model is considered as the set of latent background variables U. As shown in Fig. 2, the social bias can be represented as location names (X) showing disparity in sentiment output (Y), due to biases of sensitive attributes (A) presented in training data (U). Countering this definition of social biases, the counterfactual fairness of the predictor $\hat{Y}$ can be claimed true when

$$\begin{aligned}&\mathrm{P}(\hat{\mathrm{Y}}_{A\leftarrow a}(U) = y \mid X = x, A = a)\\ &\qquad = \mathrm{P}(\hat{\mathrm{Y}}_{A\leftarrow a'}(U) = y \mid X = x, A = a) \quad (1)\end{aligned}$$

where y is the sentiment polarity prediction and x is a specific location name. a represents situations of the locations being in a certain polarity group of one certain SVI attribute, and a' represents the counterfactual situation of locations being in the opposite polarity group of the same SVI attribute.

B. Word Frequency Disparity in Training Data

Recent studies show that models adopt and amplify biases from the training data [29]. In this part, we examine the distribution of the protected attribute set A represented by the sensitive attributes within the background variables U represented by the training data. We quantify the biases in training data by measuring the word frequency disparity. Based on the definition of social bias and the determined social attributes, we report the frequency disparity of location names from different social groups.

1) Datasets for Experiments: The training of BERT (BERT-Base, Cased) includes two steps: pretraining and fine-tuning. Downstream tasks utilize embedding models with pretrained parameters to save the computational cost for pre-training. The BooksCorpus dataset (800M words) [30] and the English Wikipedia dataset (2500M words) are used by the research team to pretrain the models.

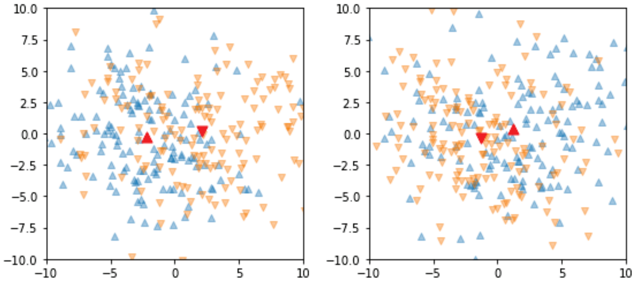
ELMo (Original) also provides pretrained embeddings for users, which are trained on the One-Billion Word Benchmark [31] corpus.

2) Methodology: All 3142 counties over the United States are ranked in terms of their attribute values. For each of the determined social attribute, we gather the top 5% and bottom 5% counties into polarity groups, and consider the counties in the top groups as having high values of respective attributes, and those in bottom groups as having low values. Counties belonging to different social attribute groups are likely to have varied occurrence frequencies in public textual materials. As a result, the occurrences of county names may also show a similar disparity in the training data (e.g., book or webpage content) of word embedding models.

We first preprocess the BooksCourpus and the One-Billion by removing all the punctuation and transforming words into lower case. Then we record the occurrences of county names by counting the occurrences of the phrases consisting of the county name and “county” (e.g., “harris county”). We denote the occurrence of words from the top polarity group as c_t and those from the bottom polarity group as c_b. In order to provide an intuitive measurement of the inequality in name occurrences from polarity groups, the results are normalized in the form of $c_t/(c_t + c_b)$ and $c_b/(c_t + c_b)$.

C. Social Biases in Word Embeddings

In this part, we evaluate the correlation between the sentiment output of $\hat{y}$ (the predictor) and the social demographic values within A (the protected attribute set) implied by location names in X (the input set). This correlation quantifies the extent of inequity between the two sides of (1), hence quantifying the extent of counterfactual unfairness. In addition to the biases in training datasets, we further examine how the social biases are encoded in the embeddings of words. We seek to find the correlation between biases in embeddings and biases in training datasets.



(a) (b)

Fig. 3. 2-D PCA of polarity groups for the example attributes MOBIL and MUNIT. Blue triangles represent vectors from the top groups and orange ones are from the bottom groups. The red triangles pointing upward and downward represent the centroids of the top and bottom clusters. (a) 2-D PCA of MOBIL. (b) 2-D PCA of MUNIT.

1) Geometry of Attributes: The embedding vectors of location names from the polarity groups are down scaled to 2-D space based on the first two principle components of principal component analysis (PCA) as shown in Fig. 3. The centroids of each pair of groups are highlighted. Embeddings from two polarity groups exhibit as clusters that subtly dissociate from each other. The cluster centroids are segregated from each other. Considering that such separation may occur in numerous other social attributes, we hypothesize that this bias could be proliferated in implementations. This hypothesis is examined based on the results of the bias detection in Section II-D2.

2) Word-Embedding-Association-Test: WEAT [25] is conducted on the embeddings of the polarity groups and the ground truth mean vectors, respectively, as two target sets U, V (corresponding to a, a') and two attribute sets P, Q (corresponding to polarized values of y). The effect size d measures how distinctive target and attribute sets are. p-Value measures the significance of the null hypothesis that two target word sets have no correlation with two attribute word sets. A higher p-value (over 0.05) indicates that the polarity groups are less distinctive and the bias is less significant.

The target sets U, V are the embeddings of location names from polarity groups concatenated together with neutral phrases as sentences (e.g., "People are" + "in Harris County"). The attribute sets P, Q are sentences synthesized for ground truth positive and negative sentiments. We collect a set of positive and negative adjectives provided by Sentiwordnet [32] and the same collection of neutral phrases in synthesized sentences (e.g., "People are" + "desperate" or "People are" + "delighted"). These sets of ground truth sentimental sentences are employed as the attribute sets.

The test statistics can be represented as

$$S(U, V, P, Q) = \Sigma_{u\in U} s(u, P, Q) - \Sigma_{v\in V} s(v, P, Q) \quad (2)$$

where

$$s(\vec{w}, P, Q) = \frac{\Sigma_{\vec{p}\in P}\phi(\vec{w}, \vec{p})}{|P|} - \frac{\Sigma_{\vec{q}\in Q}\phi(\vec{w}, \vec{q})}{|Q|}. \quad (3)$$

$\phi()$ represents the cosine similarity of two vectors. The denominators are the length of the sets, yielding mean cosine similarities. The one-sided p-value of the test can be represented as the probability of

$$s(U_i, V_i, P, Q) > s(U, V, P, Q) \quad (4)$$

for all permutation of possible X_i and Y_i. The effect size d can be obtained with

$$\left(\frac{\Sigma_{\vec{u}\in U} s(u, P, Q)}{|X|} - \frac{\Sigma_{\vec{v}\in V} s(v, P, Q)}{|Y|}\right)\Big/ \text{std}_{w\in U\cup V}(s(w, P, Q)) \quad (5)$$

where std() represents the standard deviation.

In outputs, a higher effect size indicates more bias between the name sets and severity. The lower the p-value is (lower than 0.05), the more confident we can be that the bias exists.

3) SVM Classification: In an ideally fair embedding space, the representations of location names from different social groups should be an inseparable cluster of vectors. Our intuition to introduce SVM classification is based on this assumption. The performance of SVM classifiers measures the separability of such sets of embedding vectors to measure the existence of the biases.

To determine the existence of subspaces corresponding to social attributes of location names, we train binary SVMs. SVMs are flexible and computationally efficient classifiers to predict the values of social attributes of location name in the embedding space. The values of attributes are polarized into binary classes: high and low. Accordingly, the SVMs are employed as binary classifiers, categorizing names from the two polarity groups. The more capable the trained classifiers are to predict the protected values, the more information of those social attributes can be indicated as encoded in the embedding model. This capability is measured by the $f-1$ scores of the predictions, which are calculated by $2 * (\text{recall}^{-1} + \text{precision}^{-1})^{-1}$.

For each attribute, we train 1000 SVMs on arbitrarily partitioned 80% of embeddings from each polarity group as the training data and report their average performance. Then we test them on the remaining 20% embeddings.

D. Bias Detection Results

1) Bias Detection in Training Data: The statistics of polarities vary significantly with respect to which polarity the locations belong to (the first two rows of Table II). The training data is strongly unbalanced for most of the eight attributes, in terms of the occurrence ratio between top and bottom group (mostly over 4:1 or below 1:4). Also, biases in different training data employed by different models vary. Take "**AGE65**" attribute as an example; the dominating group of BERT and ELMo alters from the top group to the bottom group. For other attributes, the disparity of bias detection results between datasets lands only on the word frequencies. The biases in embeddings introduced by models trained on the biased datasets are expected to be correlated with the bias in the data. In order to clearly illustrate the connection between training data and the corresponding embeddings, we report the correlation with a correlation symbol for each attribute and model, which is expected to be identical to the correlation between polarity groups of the embeddings.

2) Bias Detection in Word Embeddings: The results of the WEAT tests and SVM classification are shown in the bottom half of Table II. The detection results for BERT and ELMo are given as follows:

TABLE II

LEFT AND RIGHT COLUMNS UNDER EACH ATTRIBUTE CORRESPOND TO THE MEASURES OF BERT AND ELMO. THE FIRST AND SECOND ROWS PRESENT THE OCCURRENCES OF NAMES IN BOOKSCORPUS FROM THE TOP AND BOTTOM POLARITY GROUPS OF EACH ATTRIBUTES. THE FIRST AND THE SECOND ± OF EACH COLUMN PRESENT POSITIVE OR NEGATIVE CORRELATION BETWEEN THE OCCURRENCE AND POLARITY GROUPS, AND BETWEEN THE EMBEDDINGS OF THE NAMES FROM POLARITY GROUPS AND THE SENTIMENT CENTROIDS (Top-Pos. AND Btm.-Neg. AS "+"; Top-Neg. AND Btm.-Pos. AS "−"). d AND p REPRESENTS THE EFFECT SIZES AND THE p-VALUES OF WEAT TESTS. THE BOTTOM ROW SHOWS THE f-1 SCORES OF 1000 SVMS TRAINED IN THE VECTOR SPACES

		PCI		POV		AGE65		DISABL		MINRTY		LIMENG		MOBILE		MUNIT	
		B.	E.	B.	E.	B.	E.	B.	E.	B.	E.	B.	E.	B.	E.	B.	E.
Bias	top	**0.88**	**0.93**	**0.25**	0.16	0.61	0.15	0.19	0.06	0.80	0.76	0.92	**0.89**	**0.18**	0.04	**0.89**	**0.94**
in	btm.	**0.12**	**0.07**	**0.75**	0.84	0.39	0.85	0.81	0.94	0.20	0.24	0.08	**0.11**	**0.82**	0.96	**0.11**	**0.06**
Data	corr.	**+/+**	**+/+**	**-/-**	**-/-**	+/-	-/-	-/-	-/-	+/+	+/+	+/+	**+/+**	**-/-**	-/-	**+/+**	**+/+**
WEAT	d	**0.306**	**0.228**	**0.199**	0.141	0.095	0.134	0.110	0.554	0.001	0.346	0.009	**0.222**	**0.473**	0.146	**0.219**	**0.316**
	p	**0.001**	**0.017**	**0.043**	0.115	0.194	0.108	0.189	0.340	0.482	0.152	0.470	**0.029**	**0.000**	0.107	**0.021**	**0.004**
SVM	f-1	**0.70**	**0.63**	**0.66**	0.79	0.55	0.52	0.56	0.79	0.58	0.51	0.62	**0.66**	**0.72**	0.70	**0.70**	**0.66**

1) *BERT:* The location names belonging to polarity groups related to "socioeconomic" and "housing or transportation" attributes show significant biases. The WEAT test yields effect sizes greater than 0.199, with relatively low p-values less than 0.05 (whose null hypothesis is that the difference between targets is not significantly related to the attributes). The high $f - 1$ scores of SVM classifiers trained on the embeddings of these attributes indicate that the embeddings are much more likely (above 0.66 compare with close to 0.50) to be separated than those from ideally unbiased embeddings. In terms of the correlations, the attributes with significant sentimental bias all show the same correlation with the occurrence disparities of the attribute sets.
2) *ELMo:* We observe that the WEAT and the $f-1$ scores are indicating significant biases for attributes **PCI**, **LIMENG**, and **MUNIT**. The correlations of biases all match the inequity of training data. This matched correlation indicates that location names with frequent exposure to the training data yield a more positive sentimental inclination for their embedding trained on such data. This issue might be because the exposure of locations' names in textual materials is in proportion to their degree of development and the positivity of their context.

III. Bias Mitigation

Based on the biases detected and their underlying social attributes revealed in Section II, in this section, we propose a customized methodology to mitigate biases for a variety of attributes simultaneously (Fig. 4). We first extract the bias content of location names from their embeddings, and then exclude the biases from the vectors in a customized way to reduce bias information contained in the embeddings while maintaining the semantics for downstream tasks.

A. Motivation

Location names should exhibit equality in embedding spaces so that they are not to be discriminated against in downstream tasks, such as sentiment analysis for disaster situation awareness. Prevalent embedding debiasing methods such as hard debias introduced by Bolukbasi *et al.* [18] or data augmentation introduced by Zhao *et al.* [13] determine subspaces for the debiased attributes. However, in such high dimensional spaces, it is arduous to define bias subspaces for numerous social attributes that introduce biases to all different tasks. In addition, for different models with different structures that are trained on different datasets, there should be a post-hoc debiasing algorithm to produce a customized statistical structure and an improved way to extract biases. To provide a generalized methodology of embedding bias mitigation, our goal in this section is to customize the definition of one uniform bias for each location name in each dimension of the embedding space in a real-time manner.

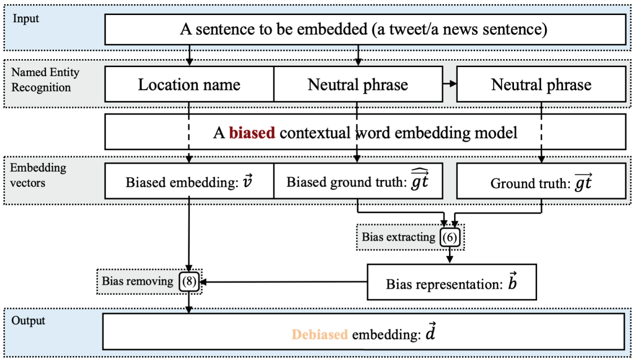

Input
A sentence to be embedded (a tweet/a news sentence)
Named Entity Recognition
Location name
Neutral phrase
Neutral phrase
A biased contextual word embedding model
Embedding vectors
Biased embedding: $\vec{v}$
Biased ground truth: $\widehat{\vec{gt}}$
Ground truth: $\vec{gt}$
Bias extracting (6)
Bias removing (8)
Bias representation: $\vec{b}$
Output
Debiased embedding: $\vec{d}$

Fig. 4. Overview of the proposed mitigation method. We employ the NER approach to divide an input sentence into a location name and neutral phrases. Then we extract the bias representation through the embeddings of neutral phrases with and without the location name in context. Bias is removed from the embedding of the location name according to the bias representation.

B. Data Collection and Preprocessing

We collect tweet posts with timestamps within the range of August 22 to September 30 in 2017 through the Twitter PowerTrack API. Using the Stanford named entity recognition (NER) tool [33], we identify county names mentioned in the posts and filter out the posts without any county names. We design a sentence template according to the most frequent contents in the filtered tweet posts to shape a testing word array corpus. In order to obtain a clear idea about the presentation of location name bias, we maintain the tweet bodies, considering them as neutral phrases in the form of word arrays. Accordingly, we substitute location names for those from various social attribute polarities, considering them as biased names. For example, we may have "People in" + "Harris County" +

"need help" as one of our word arrays, where "People in" and "need help" are the neutral phrases that are controlled within the same corpus, and "Harris County" is the biased name to be debiased.

We equally select county names according to the process for bias detection in Section II. The biased county names are from the top and bottom 5% polarity groups within the United States in terms of certain social vulnerability attributes. Each sentence in this set is represented as an array of words.

C. Bias Extraction

Both BERT and ELMo are contextual embedding models. The input of these models is sentences as arrays of word tokens. The resulting embeddings are vectors containing contextual information from adjacent words. Each word token is embedded with one vector of constant length. The bias extraction and removal process is based on this word-level structure of input and output. There might be multiple vectors for one word if a word is disassembled into multiple tokens. In this case, we mean-pool the vectors of tokens that belong to one word.

Contextual word embedding models bring contextual information to each word's embedding. The embeddings of the context of potentially biased words are straightforward sources to extract those encoded biases. As a reference or a ground truth, we first obtain the original embeddings of neutral phrases alone. Then, we extract the difference between the reference embeddings and the embeddings when neutral phrases are embedded with biased words (e.g., location names). As the ground truth vectors, the embeddings gt of the neutral phrases are collected from the model. Then, the word arrays are used as input into embedding models as well, with the output obtained as embedded vector arrays.

For each of the vector arrays, we select the vectors $\hat{\vec{\text{gt}}}_i$ corresponding to the neutral phrases in the biased sentence. The remaining vectors v_i are the name vectors to be debiased. On the jth dimension, we record the mean "bias rate" between all the i word vectors in the neutral phrases from the biased sentences $\hat{\vec{\text{gt}}}$ and their counterparts within the ground truth vectors $\vec{\text{gt}}$. The bias representation vector $\vec{b}$ is composed of the mean "bias rates" on all of the embedding space dimensions.

$$b_j = \frac{1}{|I_{\text{neu}}|}\Sigma_{i\in I_{\text{neu}}}\hat{\text{gt}}_{ij}/\text{gt}_{ij}, \quad 1 \le j \le n \tag{6}$$

$$\vec{b} = (b_1, b_2, \ldots, b_n) \tag{7}$$

where I_{neu} is the set of indices of each word in the neutral phrase, n is the dimensionality of the embedding space.

D. Bias Subtraction

Biased embeddings of location names are tailored on each dimension accordingly based on the obtained representation of "bias rate." We adjust each dimension of each word vector $\vec{v}_i$. The values of v_{ij} are shrunk or dilated in proportion to the "bias rate" b_j based on its corresponding dimension. Accordingly, we obtained the debiased word vector $\vec{d}$

$$d_{ij} = \frac{v_{ij}}{\alpha + \log_\beta b_j}, \quad 1 \le j \le n \tag{8}$$

$$\{\vec{d}_i = (d_{i1}, d_{i2}, \ldots, d_{\text{in}}) \mid i \in I_{\text{loc}}\} \tag{9}$$

where α and β are two damping factors to limit the extent of debiasing for a balance between removing biases and maintaining semantic information. In implementation, β is often determined as the form of an exponential of 10. In our experiments, α is fixed to 5 and β is been fixed to 10^3.

Algorithm 1 Customized Bias Mitigation Method

```
Input  : neutral word vectors: $\{\text{gt}_i \in \mathbf{R}^D | i \in I_{\text{neu}}\}$
         biased neutral word vectors: $\{\hat{\text{gt}}_i \in \mathbf{R}^D | i \in I_{\text{neu}}\}$
         location name vectors: $\{v_i \in \mathbf{R}^D | i \in I_{\text{loc}}\}$
         damping factors: $\alpha$, $\beta$
$b = [0] * D$
$\{d_i = [0] * |v_i| | i \in I_{\text{loc}}\}$
for $j$ from 1 to $D$ do
    $\text{sum} = 0$
    for $i \in I_{\text{neu}}$ do
        $\text{sum}+ = \hat{\text{gt}}_{ij}/\text{gt}_{ij}$
    $b_j = \text{sum}/|I_{\text{neu}}|$
for $i \in I_{\text{loc}}$ do
    for $j$ from 1 to $D$ do
        $d_{ij} = v_{ij}/(\alpha + log_\beta b_j)$
Output: $\{d_i \in \mathbf{R}^D | i \in I_{\text{loc}}\}$
```

E. Implementation

The pseudocode of the algorithm is shown in Algorithm 1. To implement this method in contextual word embedding tasks, location names are suggested to be retrieved through NER approaches (e.g., the Stanford NER tool [33]). With the determined location names, contextual words with and without those names serve as the biased and the ground truth neutral phrases for the following steps of the debiasing process.

IV. Experiments

In this section, we conduct experiments on two prevalent embedding models. We evaluate its mitigation and semantic retention performances, with comparison to a baseline.

A. Experimental Settings

In Section II, we observe that the major source of biases in pretrained embeddings is the word frequency disparity within the training dataset. We consider that the disaster-related dataset introduced for fine-tuning of embedding models is highly likely to share the same unbalanced trait on location names [5]. Thus the mitigation method is experimented on the pretrained models to produce more generalized results for various downstream tasks.

We maintain the same sentence sets of all polarity groups of 8 social attributes. For both BERT (BERT-Base, Cased) and ELMo (Original), the mitigation method is implemented upon their embeddings. The tailored embeddings are subsequently

TABLE III

Debiasing Results of Each SVI Attributes for Bolukbasi's Postprocessing Method (Bol.) and Our Customized Debiasing Method (Cus.). Bold Rows Are Those With Significant Bias in Training Data As Well As the Embeddings

	BERT							ELMo						
	Biased				De-biased			Biased				De-biased		
	d	p	f-1		d	p	f-1	d	p	f-1		d	p	f-1
PCI	**0.306**	**0.001**	**0.70**	**Bol.**	**0.066**	**0.712**	**0.47**	**0.228**	**0.017**	**0.63**	**Bol.**	**0.080**	**0.788**	**0.53**
				Cus.	**0.088**	**0.763**	**0.54**				**Cus.**	**0.096**	**1.000**	**0.55**
POV	**0.199**	**0.043**	**0.66**	**Bol.**	**0.168**	**0.069**	**0.55**	0.141	0.115	0.79	Bol.	0.071	0.738	0.52
				Cus.	**0.012**	**0.526**	**0.50**				Cus.	0.043	1.000	0.58
AGE65	0.117	0.194	0.55	Bol.	0.231	0.993	0.56	0.134	0.108	0.52	Bol.	0.036	0.623	0.46
				Cus.	0.159	0.906	0.55				Cus.	0.120	1.000	0.52
DISABL	0.110	0.180	0.56	Bol.	0.231	0.012	0.56	0.554	0.340	0.79	Bol.	0.118	0.840	0.55
				Cus.	0.127	0.866	0.49				Cus.	0.020	1.000	0.59
MINRTY	0.031	0.482	0.58	Bol.	0.106	0.181	0.51	0.346	0.152	0.51	Bol.	0.020	0.438	0.43
				Cus.	0.046	0.642	0.51				Cus.	0.012	0.854	0.55
LIMENG	0.028	0.470	0.62	Bol.	0.173	0.944	0.55	**0.222**	**0.029**	**0.66**	**Bol.**	**0.185**	**0.768**	**0.54**
				Cus.	0.224	0.964	0.58				**Cus.**	**0.092**	**1.000**	**0.54**
MOBILE	**0.473**	**0.000**	**0.72**	**Bol.**	**0.356**	**0.002**	**0.60**	0.146	0.107	0.70	Bol.	0.182	0.519	0.40
				Cus.	**0.201**	**0.953**	**0.55**				Cus.	0.058	1.000	0.58
MUNIT	**0.219**	**0.021**	**0.70**	**Bol.**	**0.273**	**0.988**	**0.57**	**0.316**	**0.004**	**0.66**	**Bol.**	**0.147**	**0.088**	**0.57**
				Cus.	**0.001**	**0.501**	**0.59**				**Cus.**	**0.012**	**0.920**	**0.49**

evaluated with the same WEAT test and SVM classification and are visualized with PCA. In PCA, we expect to see mingled vectors from the two sets with centroids close to each other. From a quantitative perspective, we expect to see relatively reduced effect sizes, insignificant p-values (lower than 0.05 as insignificant) and $f-1$ scores closer to 0.5. These metrics reflect a decent mitigation performance.

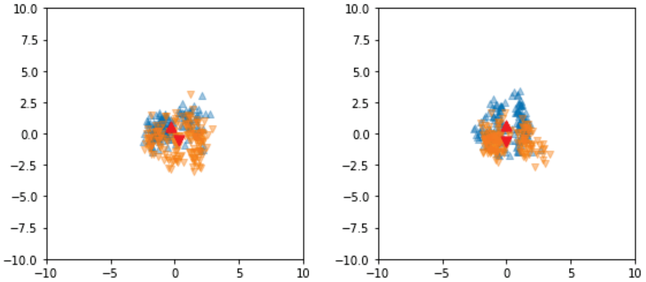
B. Baseline

We employ the postprocessing debiasing method proposed in [18]. In this method, first, attribute subspaces are determined, followed by hard debiasing and soft bias correction. We define bias subspaces of each attribute by providing two groups of definitive words. To compare the generality of our method with the baseline, for each pair of polarity groups, we control the mitigated direction to be wealth (corresponding to **PCI**) with a group of predetermined definitive word pairs (e.g., "wealth" and "poverty").

C. Bias Mitigation Evaluation

1) Geometric: In the embedding spaces, as shown in Fig. 5, we find that the vectors from two polarity groups are forming relatively denser clusters with centroids closer to each other. Also, for the two samples, we observe that the centroids rotate in the direction orthogonal to those before debiasing.

2) Quantitative Performance: The quantitative performance of debiasing is shown in Table III. For the bold attributes observed to have a significant bias in the training dataset of the two models, we also observe significant bias with this experiment setting. For BERT, on the bias-significant attributes, **PCI**, **POV**, **MOBILE**, and **MUNIT**, we observe that not only the effect sizes d are reduced to lower than 50% of the original size, but also the p-values indicating the insignificance of the biases (lower than 0.05 indicates a significant bias) rose from below 0.05 to over 0.5. Furthermore, for the other attributes that impose insignificant bias before the process, their p-values are further increased and results are significant. For ELMo, we observe significant biases for **PCI**, **LIMENG**, and **MUNIT**. For all the three attributes, the effect sizes are mostly reduced to less than 50% (except **PCI**). The p-values all increase significantly to close to 1.0. For the other attributes, the mitigation brings lower significance by reducing the effect sizes d and increase their p-values. For the debiased embeddings of both models, The SVM back-predictors have lower performance for classifying attribute polarity groups from their location name vectors. The $f-1$ scores have been reduced from over 0.6 to less than 0.6, much closer to random guess (0.5).



(a) (b)

Fig. 5. 2-D PCA of polarity groups of example attributes MOBIL and MUNIT after debiasing. See the caption of Fig. 1 for a detailed description on the legends. (a) 2-D PCA of MOBIL. (b) 2-D PCA of MUNIT.

The baseline postprocessing method is shown to perform relatively well in the controlled attribute **PCI**. For the other seven attributes that the method does not specifically debias, the effect sizes measured by WEAT and the $f-1$ scores of SVM classifications are observed to be less reduced than our customized method. This result indicates that our customized method has a more significant debiasing performance on overall attributes, rather than the only one specific attribute mitigated by the baseline method.

D. Mitigation Versus Semantics Retention Tradeoff

One of the challenges for developing word embedding debiasing methods is the balance between mitigation performance

TABLE IV

RESULTS OF BERT AND ELMO FOR CONCEPT CATEGORIZATION ON BENCHMARKS PRIOR (Org.) AND POSTERIOR TO BOLUKBASI'S POST-PROCESSING METHOD (Bol.) AND OUR CUSTOMIZED DEBIASING METHOD (Cus.). PERFORMANCE IS MEASURED IN PURITY SCORES. IN BOTH MODELS, THERE IS NO SIGNIFICANT DEGRADATION OF PERFORMANCE DUE TO APPLYING THE PROPOSED METHOD

		BLESS	BM	AP	ESSLLI
B.	Org.	0.80	0.38	0.41	0.71
	Bol.	0.75	0.27	0.35	0.69
	Cus.	0.75	0.30	0.42	0.73
E.	Org.	0.82	0.31	0.54	0.78
	Bol.	0.73	0.31	0.52	0.78
	Cus.	0.84	0.28	0.51	0.78

and retention of semantics for downstream tasks. In this section, we evaluate how our proposed method retains semantics after mitigation.

1) Evaluation Methodology: We conduct concept categorization tasks to evaluate semantic retention of the debiased embeddings. In concept categorization, a set of words are provided with labeled classes. The capability of an embedding to distinguish one class from the other is the test objective. Clustering methods such as KMeans are employed to cluster words by their embeddings. To quantitatively evaluate the retention of semantics, the purity scores of clustering are measured. The purity [34] of concept categorization has been evaluated upon several benchmark datasets, including the BLESS [21], BM [22], AP [23], and ESSLLI-2008 [24]. For all of these datasets, we record the words and attach them to the provided conceptual labels. The embeddings of both biased and debiased models are clustered with KMeans. The number of clusters for KMeans is fixed to the exact number of conceptual labels the datasets provide. We then assemble the words in the benchmark datasets with the same neutral phrases used in debiasing and process them with the proposed mitigation methods. The embeddings of the target words are recorded. The purity scores on clustering the biased and debiased vectors of target words

$$\mathrm{purity}(C, S) = \frac{1}{n}\Sigma_i \max_j (c_i \cup s_j) \tag{10}$$

where $C = \{c_1, c_2, \ldots, c_n\}$ denotes the clusters, n represents the number of clusters, and $S = \{s_1, s_2, \ldots, s_m\}$ denotes the semantic categories, are shown in Table IV.

2) Results: The original BERT achieves relatively high purity scores in categorizing BLESS and ESSLLI datasets and BM and AP under 0.5. The purity after being mitigated by the proposed method, compared to the original scores, does not decrease significantly. In fact, purity scores in half of the tests even increase. For ELMo, the same pattern of purity fluctuation occurs in the original model. After the embedding is debiased by the proposed method, scores of BM, AP, and ESSLI decrease insignificantly. The purity score of BLESS increases. Compared with the results of the baseline method, the purity has been less significantly impaired after debiasing the embeddings with our method, thus indicating that the model retains adequate semantics for downstream tasks.

V. Case Study: A Social Media Content Classifier for Disaster Situation Awareness

In this section, we evaluate our mitigation method on a BERT-based social media content classifier in a pipeline for disaster situation awareness and responses [5]. The pipeline includes three modules: input, learning, and output.

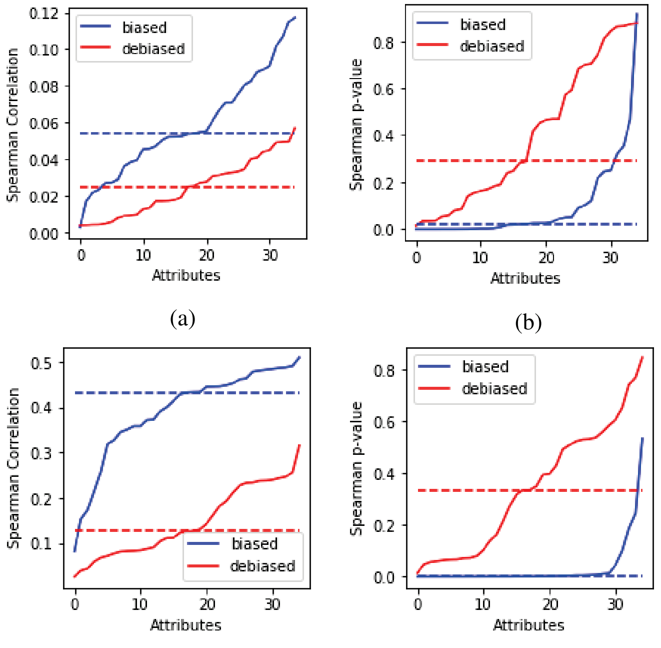
1) *The input module* filters social media data to retain necessary input information.
2) *The learning module* is composed of the location entity extraction and the fine-tuned BERT (BERT-Base, Cased) based classifier. The location entity extraction unit obtains multiscale location information from social media contents. The BERT classifier label contents with eight different predetermined humanitarian categories ("affected individuals," "injured people," "missing people," "infrastructure and utility damage," "vehicle damage," "rescue, volunteering or donation," "Other relevant" and "Nonrelevant").
3) *The output module* conducts further evaluations of the development of disaster events.

A. Task Objective

The case study focuses on the BERT classifier in the learning module. The classifier is developed to determine the humanitarian category related to tweets content. In the implementation, the fine-tuned model outputs predicted probabilities for each of the 8 humanitarian categories [5]. In the ideal nonbiased case, for a set of tweets consisting of various location names and identical phrases as the context, the prediction for a certain category should be identical. However, if biases are encoded in the embeddings of the location names, we expect to observe a correlation between the nonuniform predictions and the values of certain social attributes of the tweets with different location names.

1) Test Data: Following the study of the original pipeline [5], we adopt the humanitarian categories and labeled tweet posts for Hurricane Harvey in 2017 from the dataset CrisisMMD [35]. First, tweet posts labeled as "affected individuals" are selected. Using the Stanford NER tool [33], we identify county names mentioned among the selected posts and filter out the posts without county names. Next, we introduce perturbation to the sentences by controlling the sentence bodies and substituting the original location names with locations names from polarity groups of all 34 social attributes provided by SVI. This perturbed tweet corpus is then conveyed to the classifier as the input.

2) Evaluation Metrics: We compare the prediction rankings prior and posterior to debiasing with the determined metrics. The classifier outputs eight predicted probabilities for the eight humanitarian categories for each input sentence. We rank the counties by their predicted probabilities of being categorized 'affected individuals', in a descending order. This rank is denoted as R_0. We then rank the counties by the values under each SVI attribute. This rank is denoted as R_1. Each county will have its R_0 and one R_1 for each SVI attribute. The ranking R_1 of the counties' value under each SVI attribute is also saved in a descending order. Spearman correlations between



(a)

(b)

(c)

(d)

Fig. 6. Sorted correlations and p-values of Spearman test of all SVI attributes on classification predictions for all counties [(a) and (b)] and top/bottom 1% counties [(c) and (d)] prior and posterior to debiasing. Broken lines represents the median values among all attributes.

R_0 and all 34 R_1s are measured. ρ represents the Spearman correlations. The p-value is measured on the null hypothesis that two rankings are uncorrelated. If the embeddings of location names are not correlated with their social attributes, the prediction for sentences containing different locations should be identical ($\rho = 0$). The Spearman correlation is firstly measured on the original R_0 for all counties within the United States. Next, we retain only 1% top and 1% bottom counties according to the probability ranking R_0 and record social attribute rankings R_1 only for counties within the two polarity groups. Spearman correlation is measured accordingly.

B. Quantitative Results

Research studies have provided standards for interpreting the measures of Spearman correlations [36].

For all US counties, as shown in Fig. 6(a) and (b), the correlations of attributes have shown a decreasing order for over 50%. The p-value, on the other hand, increased from 0.021 (under 0.05) to 0.289, indicating less confidence for the correlations. For the selected top and bottom 1% counties, as shown in Fig. 6(c) and (d), the correlations between attributes and the probability ranking of the predictions have a median value decreased from 0.442 to 0.137, interpreted as from strong correlation to negligible according to correlation standard for politic area [36]. The tests yield p-values with a median value significantly increasing from 10^{-4} to 0.335, from below the 0.001 threshold of significance to insignificant. This result indicates that the correlations between the classifier's outputs and SVI attributes are less significant after debiasing. The distribution curves of ρ and p-values demonstrate that the biases are mitigated under the majority of attributes.

The debiased classifier yields an average probability of 74.95% as the prediction for the test sentences on "affected individuals," correctly as they were labeled in CrisisMMD. This result indicates that the debiased embeddings thoroughly retain the necessary semantics for the pipeline.

C. Discussion

The Spearman tests reveal the extent to which social biases encoded in embedding models could be propagated to the downstream tasks. These biases, as a consequence, produce predictions with prejudices. In this case of classifier implemented for disaster situation awareness, a prejudiced prediction is highly prone to yielding results shown in Fig. 1. In the implementation, this would result in an unbalanced location-wise awareness of disaster situations, and thus bring critical crisis responses to different locations with potential disparity in resource allocation and relief and rescue efforts.

After bias mitigation, the correlations between predictions of locations and their social attributes are largely reduced to an insignificant level. This case study demonstrates clearly how this approach could yield fairer outputs for social-related tasks based on contextual word embedding models.

VI. Related Work

Word embeddings are trained on large-scale corpora consisting of artificial text materials. Recent studies show the presence of biases related to gender and corresponding occupations [13], [18], [25], ethnicity, and corresponding human names [37], [38] in embedding models. Therefore, models utilizing these embeddings tend to involve undesired biases, and subsequently propagate them into downstream task results [39], [40].

The current literature on embedding bias mitigation mainly focuses on the prevalent attributes such as gender or ethnicity. A postprocessing debiasing method for word2vec model trained on the Google News corpus [41] is proposed in [18]. In these studies, gender biases are mitigated through three steps: gender space identification, hard debiasing, and soft debiasing. For contextual embeddings like BERT and ELMo, several researchers have proposed methods to detect as well as mitigate gender biases [13]. Biases are observed in unbalanced training data and the geometry of the embedding spaces. For ELMo, researchers introduced data augmentation and neutralization, respectively, for the training and the testing processes [13]. Some other studies also show strong evidence related to social and intersectional biases in various state-of-the-art contextual embeddings [42], with the debiasing methods remain to be purposed. Since this article is focusing on biases in location names, the social attributes behind locations are inevitable to be a source of bias. In this article, We adopt the definition of counterfactual fairness [28] to determine the statistical notions of fairness.

In addition to the debiasing process, the retention of semantics in the embeddings is necessary to be tested. Existing semantic retention tests mainly consist of two categories of benchmark evaluations, including word analogy and concept categorization. For word analogy, prevalent benchmark tasks include MSR [43] and Google word analogy.

For concept categorization, benchmark datasets such as the BLESS dataset [21], the Battig 1969 set [22], the Almuhareb-Poesio dataset [23] and the ESSLLI 2008 [24] are commonly employed by researches. Concept categorization tasks are tested and evaluated on debiased embeddings to examine the retention of semantics by the proposed bias mitigation methods.

As a major implementation of word embeddings, researchers use sentiment analysis techniques on social media data to increase disaster situation awareness [44]–[46]. Word embedding models are employed by different recent studies to help accelerate the process and enrich the semantic expression [5], [47], [48]. In these tasks, social media contents are first embedded into vector spaces and are then classified by various clustering methods. The names of locations shape an essential part of the training data of the embedding models, as well as the input for the models to predict on. Hence, biases related to location names from the embedding models should be mitigated before the downstream tasks.

VII. Conclusion

The biases under different social attributes encoded in the contextual models BERT and ELMo are originated and quantified in this study. We find a positive correlation between the occurrences of locations in training data and the positive sentimental inclination of their embeddings. The proposed bias mitigation method shows its versatility of reducing a major proportion of bias on different attributes from different embeddings. At the same time, the method well retains semantic information of embeddings for downstream tasks. With the case study, we demonstrate how those biases are potentially hampering the location-wise fairness of crisis situation awareness, and the subsequent relief efforts that impact potential human lives. In this case, our method help to produce a more objective and fairer severity evaluation across locations.

Acknowledgment

This material is based in part upon work supported by the National Science Foundation under Grant CMMI-1846069 (CAREER). Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] I. Iacobacci, M. T. Pilehvar, and R. Navigli, “SensEmbed: Learning sense embeddings for word and relational similarity,” in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics*, 2015, pp. 95–105.

[2] M. E. Peters *et al.*, “Deep contextualized word representations,” 2018, *arXiv:1802.05365*. [Online]. Available: http://arxiv.org/abs/1802.05365

[3] A. Akbik, D. Blythe, and R. Vollgraf, “Contextual string embeddings for sequence labeling,” in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 1638–1649.

[4] F. Viegas *et al.*, “CluWords: Exploiting semantic word clustering representation for enhanced topic modeling,” in *Proc. 12th ACM Int. Conf. Search Data Mining*, Jan. 2019, pp. 753–761.

[5] C. Fan, F. Wu, and A. Mostafavi, “A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters,” *IEEE Access*, vol. 8, pp. 10478–10490, 2020.

[6] M. Imran, P. Mitra, and C. Castillo, “Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages,” 2016, *arXiv:1605.05894*. [Online]. Available: http://arxiv.org/abs/1605.05894

[7] Q. Huang and Y. Xiao, “Geographic situational awareness: Mining tweets for disaster preparedness, emergency response, impact, and recovery,” *ISPRS Int. J. Geo-Inf.*, vol. 4, no. 3, pp. 1549–1568, Sep. 2015.

[8] A. Sen, K. Rudra, and S. Ghosh, “Extracting situational awareness from microblogs during disaster events,” in *Proc. 7th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2015, pp. 1–6.

[9] H. A. M. Hassan, G. Sansonetti, F. Gasparetti, A. Micarelli, and J. Beel, “Bert, elmo, use and infersent sentence encoders: The panacea for research-paper recommendation?” in *Proc. RecSys*, 2019, pp. 6–10.

[10] H. Xu, B. Liu, L. Shu, and P. S. Yu, “BERT post-training for review reading comprehension and aspect-based sentiment analysis,” 2019, *arXiv:1904.02232*. [Online]. Available: http://arxiv.org/abs/1904.02232

[11] H. Xu, B. Liu, L. Shu, and P. S. Yu, “Review conversational reading comprehension,” 2019, *arXiv:1902.00821*. [Online]. Available: http://arxiv.org/abs/1902.00821

[12] M. Mozafari, R. Farahbakhsh, and N. Crespi, “A bert-based transfer learning approach for hate speech detection in online social media,” in *Proc. Int. Conf. Complex Netw. Appl.* Cham, Switzerland: Springer, 2019, pp. 928–940.

[13] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang, “Gender bias in contextualized word embeddings,” 2019, *arXiv:1904.03310*. [Online]. Available: http://arxiv.org/abs/1904.03310

[14] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” 2019, *arXiv:1908.09635*. [Online]. Available: http://arxiv.org/abs/1908.09635

[15] C. Basta, M. R. Costa-Jussá, and N. Casas, “Evaluating the underlying gender bias in contextualized word embeddings,” 2019, *arXiv:1904.08783*. [Online]. Available: http://arxiv.org/abs/1904.08783

[16] K. Kurita, N. Vyas, A. Pareek, A. W Black, and Y. Tsvetkov, “Measuring bias in contextualized word representations,” 2019, *arXiv:1906.07337*. [Online]. Available: http://arxiv.org/abs/1906.07337

[17] R. Binns, “Fairness in machine learning: Lessons from political philosophy,” 2017, *arXiv:1712.03586*. [Online]. Available: http://arxiv.org/abs/1712.03586

[18] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, “Man is to computer programmer as woman is to homemaker? debiasing word embeddings,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4349–4357.

[19] T. Wang, X. Victoria Lin, N. F. Rajani, B. McCann, V. Ordonez, and C. Xiong, “Double-hard debias: Tailoring word embeddings for gender bias mitigation,” 2020, *arXiv:2005.00965*. [Online]. Available: http://arxiv.org/abs/2005.00965

[20] B. E. Flanagan, E. W. Gregory, E. J. Hallisey, J. L. Heitgerd, and B. Lewis, “A social vulnerability index for disaster management,” *J. Homeland Secur. Emergency Manage.*, vol. 8, no. 1, Jan. 2011, Art. no. 3.

[21] M. Baroni and A. Lenci, “How we blessed distributional semantic evaluation,” in *Proc. Workshop GEometrical Models Natural Lang. Semantics*, 2011, pp. 1–10.

[22] W. F. Battig and W. E. Montague, “Category norms of verbal items in 56 categories a replication and extension of the Connecticut category norms,” *J. Exp. Psychol.*, vol. 80, no. 3, pp. 1–46, Jun. 1969.

[23] A. Almuhareb, “Attributes in lexical acquisition,” Ph.D. dissertation, Dept. Comput. Sci., Univ. Essex, Colchester, U.K., 2006.

[24] M. Baroni, S. Evert, and A. Lenci, *Bridging the Gap Between Semantic Theory and Computational Simulations: Proceedings of the Esslli Workshop on Distributional Lexical Semantics*. Hamburg, Germany: FOLLI, 2008.

[25] A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” *Science* vol. 356, no. 1334, pp. 183–186, 2017.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[27] S. Mitchell, E. Potash, S. Barocas, A. D’Amour, and K. Lum, “Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions,” 2018, *arXiv:1811.07867*. [Online]. Available: http://arxiv.org/abs/1811.07867

[28] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, “Counterfactual fairness,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4066–4076.

[29] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Men also like shopping: Reducing gender bias amplification using corpus-level constraints,” 2017, *arXiv:1707.09457*. [Online]. Available: http://arxiv.org/abs/1707.09457

[30] Y. Zhu *et al.*, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 19–27.

[31] C. Chelba *et al.*, “One billion word benchmark for measuring progress in statistical language modeling,” 2013, *arXiv:1312.3005*. [Online]. Available: http://arxiv.org/abs/1312.3005

[32] A. Esuli and F. Sebastiani, “SentiWordNet: A publicly available lexical resource for opinion mining,” in *Proc. LREC*, vol. 6, 2006, pp. 417–422.

[33] J. R. Finkel, T. Grenager, and C. D. Manning, “Incorporating non-local information into information extraction systems by Gibbs sampling,” in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2005, pp. 363–370.

[34] *Introduction to Information Retrieval*, Univ. Cambridge, Cambridge, U.K., 2009.

[35] F. Alam, F. Ofli, and M. Imran, “CrisisMMD: Multimodal Twitter datasets from natural disasters,” 2018, *arXiv:1805.00713*. [Online]. Available: http://arxiv.org/abs/1805.00713

[36] H. Akoglu, “User’s guide to correlation coefficients,” *Turkish J. Emergency Med.*, vol. 18, no. 3, pp. 91–93, Sep. 2018.

[37] A. Romanov *et al.*, “What’s in a name? Reducing bias in bios without access to protected attributes,” 2019, *arXiv:1904.05233*. [Online]. Available: http://arxiv.org/abs/1904.05233

[38] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, “Word embeddings quantify 100 years of gender and ethnic stereotypes,” *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 16, pp. E3635–E3644, Apr. 2018.

[39] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Gender bias in coreference resolution: Evaluation and debiasing methods,” 2018, *arXiv:1804.06876*. [Online]. Available: http://arxiv.org/abs/1804.06876

[40] J. Escudé Font and M. R. Costa-jussa, “Equalizing gender biases in neural machine translation with word embeddings techniques,” 2019, *arXiv:1901.03116*. [Online]. Available: http://arxiv.org/abs/1901.03116

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781

[42] Y. C. Tan and L. E. Celis, “Assessing social and intersectional biases in contextualized word representations,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13230–13241.

[43] T. Mikolov, W.-T. Yih, and G. Zweig, “Linguistic regularities in continuous space word representations,” in *Proc. Conf. north Amer. Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2013, pp. 746–751.

[44] M. K. Torkildson, K. Starbird, and C. Aragon, “Analysis and visualization of sentiment and emotion on crisis tweets,” in *Proc. Int. Conf. Cooperat. Design, Vis. Eng.* Cham, Switzerland: Springer, 2014, pp. 64–67.

[45] Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta, “Tweedr: Mining Twitter to inform disaster response,” in *Proc. ISCRAM*, 2014, pp. 269–272.

[46] C. Caragea *et al.*, “Mapping moods: Geo-mapped sentiment analysis during hurricane sandy,” in *Proc. ISCRAM*, 2014, pp. 1–19.

[47] J. Ray Chowdhury, C. Caragea, and D. Caragea, “Keyphrase extraction from disaster-related tweets,” in *Proc. World Wide Web Conf.*, 2019, pp. 1555–1566.

[48] J. Liu, T. Singhal, L. T. M. Blessing, K. L. Wood, and K. H. Lim, “CrisisBERT: A robust transformer for crisis classification and contextual crisis embedding,” 2020, *arXiv:2005.06627*. [Online]. Available: http://arxiv.org/abs/2005.06627

Fangsheng Wu received the bachelor’s degree in electronic information engineering from the Electronic Information School, Wuhan University, Wuhan, China, in 2018. He is currently pursuing the master’s degree in computer engineering with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA.

He also serves as a Research Assistant with the Urban Resilience AI Lab, Department of Civil and Environmental Engineering, Texas A&M University. His research scope includes social network data analysis and artificial intelligence techniques. He is also highly interested in exploiting various topics in social computing.

Mengnan Du is currently pursuing the Ph.D. degree in computer science at the CSE Department, Texas A&M University, College Station, TX, USA. His advisor is Dr. Xia Hu.

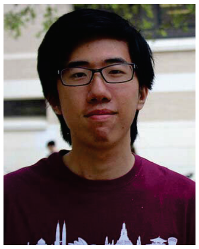
His research is on interpretable machine learning, with a particular interest in the areas of DNN interpretability. He is also interested in areas of fairness in deep learning, adversarial detection, fake news detection, and medical diagnosis.

Chao Fan received the M.S. degree from the University of California, Davis, CA, USA, in 2017 and the Ph.D. degree in civil engineering from Texas A&M University, College Station, TX, USA, in 2020.

He is currently a Post-Doctoral Researcher with the Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX, USA. He joined the Urban Resilience AI Laboratory in Fall 2017. His research interests include cutting-edge interdisciplinary research at the interface of engineering, science and policy for urban resilience in disasters using large-scale urban big data and computational methods. He received the Doctoral Fellow from the HICSS Conference.

Ruixiang Tang received the bachelor’s degree from the Department of Automation, Tsinghua University, Beijing, China, in 2018, advised by Prof. Jiwen Lu. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Rice University, Houston, TX, USA. His advisor is Prof. Xia Hu, who leads the DATA Lab.

His research interests lie in interpretable machine learning, with a particular interest in the areas of DNN interpretability. He is also interested in areas of fairness in deep learning, backdoor attack, and IP protection of ML pipelines.

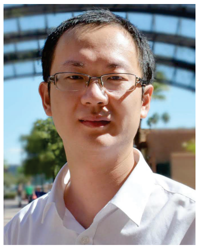


Yang Yang was born in Guangdong, China, in 1995. He received the B.S. degree in computer science from Texas A&M University, College Station, TX, USA, in 2019, where he is currently pursuing the M.S. degree in computer science.

After joining the Urban Resilience AI Laboratory, in May 2018, he worked on multiple interdisciplinary studies of applying data analytics and machine learning methods on social media analysis. His current research interests include fairness in artificial intelligence and data analysis on social media user behaviors.

Ali Mostafavi received the M.S. degree in industrial administration from the Krannert School of Management, West Lafayette, IN, USA, in 2011, and the Ph.D. degree in civil engineering from Purdue University, West Lafayette, in 2013.

He is currently an Associate Professor with the Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX, USA. He is also the Director of the Urban Resilience AI Laboratory. His research interests include creating transformative solutions for addressing the grand challenges pertaining to the nexus of humans, disasters, and infrastructure systems.



Xia Hu (Member, IEEE) received the B.S. and M.S. degrees in computer science from Beihang University, Beijing, China, in 2006 and 2009, respectively, and the Ph.D. degree in computer science and engineering from Arizona State University, Tempe, AZ, USA, in 2015.

He is currently an Associate Professor with the Department of Computer Science, Rice University, Houston, TX, USA. He has published nearly 100 papers in several major academic venues. His developed automated machine learning (AutoML) package, AutoKeras, has received more than 6000 stars on GitHub and has become the most rated open-source AutoML system.