

Inherent Discriminability of BERT Towards Racial Minority Associated Data

Maryam Ramezanzadehmoghadam[1], Hongmei Chi[1(✉)], Edward L. Jones[1], and Ziheng Chi[2]

[1] Department of Computer and Information Sciences, Florida A&M University, Tallahassee, FL 32307, USA
{maryam1.ramezanzade,hongmei.chi,edward.Jones}@famu.edu

[2] Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong, China

Abstract. AI and BERT (Bidirectional Encoder Representations from Transformers) have been increasingly adopted in the human resources (HR) industry for recruitment. The increased efficiency (e.g., fairness) will help remove biases in machine learning, help organizations find a qualified candidate, and remove bias in the labor market. BERT has further improved the performance of language representation models by using an auto-encoding model which incorporates larger bidirectional contexts. However, BERT's underlying mechanisms that enhance its effectiveness, such as tokenization, masking, and leveraging the attention mechanism to compute vector score, are not well understood.

This research analyzes how BERT's architecture and its tokenization protocol affect the low number of occurrences of the minority-related data using the cosine similarity of its embeddings. In this project, by using a dataset of racially and gender-associated personal names and analyzing the interactions of transformers, we present the unfair prejudice of BERTs' pre-trained network and autoencoding model. Furthermore, by analyzing the distance of an initial word's token and its MASK replacement token using the cosine similarity, we will demonstrate the inherent discriminability during pre-training. Finally, this research will deliver potential solutions to mitigate discrimination and bias in BERT by examining its geometric properties.

Keywords: BERT · Machine learning fairness · Natural language processing · Racial minority · Pre-trained model · Recruitment

1 Introduction

Our last names and given names are our first social labels that correspond with our gender identity, race and ethnicity, or country of origin. However, with AI being increasingly adopted in screening resumes in the human resources

Supported by organization x.

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12951, pp. 256–271, 2021.
https://doi.org/10.1007/978-3-030-86970-0_19

(HR) industry for purposes such as predicting employee attrition and background verification, women and individuals from minority ethnic groups (compared to majority groups) are facing unfavorable labor market outcomes in many economies [13]. In addition, AI-based recruitment platforms are widely adopted in some industries recruiting new employees. However, the potential discrimination in recruitment algorithms is hard to assess [28]. In this paper, we check on name discrimination only and review BERT's inherent bias.

Machine learning is being used in a wide range of applications. With machines being trained on different data sets to make decisions, suggestions, and recommendations, it is genuinely essential to take the "fairness" of these decisions into account. In the context of decision-making, fairness is the "absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics" [18]. A model's output, if not fair, will result in decisions that can be biased against certain groups with a set of characteristics.

With the widespread use of AI systems and their human and computer interaction applications processing natural language, it is crucial to consider system fairness while designing and engineering them. Unfortunately, there are many different sources of bias when dealing with Big Data, leading to unfair biases towards other groups. [18] in their work have introduced and discussed some of the most general and essential sources of bias introduced in [29] and [21]. Even though all of the introduced bias sources can be applied to the field of natural language processing, but Historical Biases, Representational Biases, Popularity Biases, and Social Biases are the main focus of this work [19].

Historical Biases are the already existing bias and socio-technical issues in the world which are unrelated to the policy under study and can permeate the data generation process even given a perfect sampling and feature selection which can have a significant effect on the policy's hoped for outcome [29]. Representational Biases are the cause of wrongfully defining and sampling from a population. Popularity Biases come from the following underlying issue: More popular Items tend to be exposed more. When studying natural language, popularity metrics are ubiquitous to be manipulated based on the existing historical biases towards minority groups available in linguistic corpora. Social bias happens when other people's actions or content coming from them affect systems judgment. Considering the existing historical biases towards minority-related groups and ethnicities (African America, Hispanic, Asian, Middle Eastern), Representational, Popularity, and Social biases are easy and probable to happen in linguistic corpora, leading to Indirect systematic discrimination.

Indirect discrimination happens when individuals appear to be treated based on seemingly neutral and non-protected attributes; however, protected groups or individuals still get to be treated unjustly as a result of implicit effects from their protected attributes [31]. Systematic discrimination happens as patterns of behavior, policies, or practices that are part of the culture or structures of an organization, and which create or perpetuate disadvantage for racialized persons [10].

In various applications, pre-trained natural language processing (NLP) models run the risk of exploiting and reinforcing the societal biases that are present in the underlying data [8]. Therefore, it is crucial to understand the unjust bias inherited by them and ensure that the decisions do not reflect discriminatory behavior toward certain groups or populations.

BERT has further improved the performance of deep learning and neural network architecture by using an auto-encoding model which incorporates larger bidirectional contexts [24]. However, the underlying mechanisms of BERT that are improving its effectiveness, such as tokenization, masking, and leveraging the attention mechanism to compute vector scores, are not well understood. This paper analyses how BERT's architecture and its tokenization protocol affect the low occurrences of the minority-related data via analyzing BERT's vocabulary, tokenization, and embeddings using WEAT and cosine similarity test.

The rest of this paper is organized as follows. Section 2 presents an overview of the Natural Language Processing, context-free and contextual pre-trained language models, as well as BERT. Related works will be discussed in the Sect. 3. A detailed description of the datasets that have been used is provided in Sect. 4. Details on the methodology and performed analysis is provided in Sect. 5. Preliminary results are presented in Sect. 6. Finally, Sect. 7 provides the conclusion and future works.

2 Background

Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis to achieve human-like language processing for a range of tasks or applications [17]. Measurable property, in machine learning, is a piece of information that can be observed as a characteristic when solving a problem via selecting relative, quantitative and qualitative. Selected feature, is a critical step in model construction since it can simplify the model, shorten the training time, deliver a practical algorithm, and enhanced results.

2.1 Language Representations

Word embeddings are used to converting each word in text data into vectors which are a numerical representation of words, and vectors are used to extract the best features from the content of texts. These vector representations capture the underlying words in relation to the other words present in a sentence. Words with similar meanings will be grouped closer together in a hyperplane that distinct them from words that are position further in that hyperplane. This step could be done by either creating a corpus-specific word embedding based and trained on the selected features that will be introduced from the training data set, or it could be done by using a pre-trained word embedding. The problem with the embedding layer is that its features will be relevant for only a specific issue. There won't be enough words introduced to it to map the encoded categorical variables. Pre-trained representations can either be context-free or contextual:

- Context-free models such as Word2Vec [9] or GloVe [22] generate a single "word embedding" representation for each word in the vocabulary no matter the position of the word in the text.
- Contextual models such as ELMO [23], Universal Language Model Fine-tuning (ULMFit) [14] and BERT, generate a representation of each word based on the other words in the text.

Contextual representations can further be unidirectional or bidirectional. ELMO, ULMFit, and previously pre-trained models were unidirectional, which means that each word is only contextualized using the words to its left (or right). However, BERT has a bidirectional architecture.

2.2 BERT

BERT is designed to pre-train deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications [12].

BERT with it's bidirectional architecture claims to obtain state-of-the-art results on eleven natural language processing tasks [12].

3 Related Works

Several studies on identification and mitigation of bias in statistical context-free word embeddings such as GloVE and Word2Vec by [5] and Skip-Gram embeddings trained on various data sets and FastText embeddings by [7] and [15]. However, there have been only a few studies [32] and [2] on the fairness of contextual language models (CLMs) which are being leveraged in a variety of domains. Studies on BERT's fairness are mainly focused on gender bias, and there have been only a few studies that analyzed fairness of BERT's word embedding towards racial minority groups.

Zhao et al. (2019) [32] have quantified, analyzed and mitigated gender bias exhibited in ELMo's contextualized word vectors, and demonstrated that training data for ELMo contains significantly more male than female entities, ELMo's embeddings systematically encode gender information, and its co-reference system inherits its bias and presents a significant bias on the WinoBias probing corpus.

Bhardwaj, Majumder and Poria (2020) [4] analyze the gender bias that CLMs induce in different downstream tasks related to emotion and sentiment intensity prediction. They have trained a simple regressor utilizing BERT's word embeddings for each task and evaluated the gender bias in regressors using an equity evaluation corpus. Using this method, they proved a significant dependence of BERT's predictions on gender-particular words and phrases.

Kurita et al. (2019) [16] argued that contextual word embeddings, such as BERT are optimized to capture the statistical properties of training data. They also tend to pick up and amplify social stereotypes present in the data. They proposed a template-based method to quantify bias in BERT, which is generalizable to unveiling other biases in multi-class settings, such as racial and religious biases. Our work leverages their approach to demonstrate the racial bias in BERT.

Bartl, Nissim and Gatt [1] measured BERT's gender bias by studying associations between gender-denoting target words and names of professions in English and German, comparing the findings with real-world workforce statistics. Their results demonstrated that male person terms seem to have a more stable position in BERT, which could cause their probabilities in the model to not vary much depending on the context and make them less susceptible to change through fine-tuning. Our work uses the list of professions provided by them to demonstrate the racial bias in BERT.

Sharma, Dey and Sinha (2020) [27] demonstrated in their work that three models (BERT, RoBERTa, and BART) trained on MNLI and SNLI data-sets are significantly prone to gender-induced prediction errors and using a debiasing technique, such as augmenting the training data set to ensure that it is a gender-balanced data can help reduce such bias in some instances.

Podkorytov, et al. (2020) [24] presented experimental results illustrating that BERT does not produce "effective" contextualized representations for words. Instead, its improved performance may mainly be due to fine-tuning or classifiers that model the dependencies explicitly by encoding syntactic patterns in the training data.

4 Dataset

Bias in pre-trained contextualized word representation models is commonly measured via sentence templates. For this purpose, we have used two different corpora (EEC and BEC-Pro) developed by [15] and [1] and a data set from the Field Experiment on Labor Market Discrimination (LMD) designed by [3].

The Equity Evaluation Corpus (EEC) data set analyzes bias in NLP systems using the context's connection with emotions. The BEC-Pro data set analyzes bias in NLP systems using the connection of the context with professions. The LMD data set provides a list of names associated with two different races of African American and European American and their sex. The details of the mentioned data set will be discussed in Sect. 4.1, 4.2 and 4.3.

4.1 Equity Evaluation Corpus

The Equity Evaluation Corpus (EEC) is a data set of thousands of sentences to determine whether automatic systems consistently give higher (or lower) sentiment intensity scores to sentences involving a particular race or gender [15]. This corpus includes 11 sentence templates divided into two groups of emotion and

non-emotion words with two variables of <person> and <emotion word>. First, we have used this template to generate sentences by instantiating the <person> noun phrases with the African American and European American first names associated with both genders of female and male. Next, we have instantiated the <emotion word> with the emotional state words correspond to pleasantness/unpleasantness, rich/poor, and happy/sad following the work of [6] and [15].

4.2 BEC-Pro

The BEC-Pro data set includes sentence templates containing a gender-denoting noun phrase <person> and a <profession>. The professions on this template were formed by three groups of 20 professions from the U.S. Bureau of Labor Statistics (2020) [20] with the following categories:

- Group 1 with highest female participation (88.3%–98.7%)
- Group 2 with lowest female participation (0.7%–3.3%)
- Group 3 with a roughly 50-50 distribution of male and female employees (48.5%–53.3%)

We have used the five-sentence templates of [1] that include the <person> noun phrases and the carries explicit gender information and a profession term. In addition, we have instantiated the <person> noun phrases with the African American and European American first names associated with both genders.

4.3 Labor Market Discrimination

The Field Experiment on Labor Market Discrimination (LMD) data set was constructed by [3] when studying race in the labor market. They posted fictitious resumes to help-wanted ads in Boston and Chicago newspapers. Marianne Bertrand and Sendhil Mullainathan's [3] research used name frequency data calculated from birth certificates of all babies born in Massachusetts between 1974 and 1979. They tabulated these data by race to determine which names are distinctively European American and distinctively African-American. Distinctive names were those that had the highest ratio of frequency in one racial group to the frequency in the other racial group. A survey was also conducted to check the distinctiveness of their data. We have used their data set of racially distinct names and combined it with the two previously mentioned corpora to perform a racial bias analysis on the BERT algorithm.

5 Methodology

We have performed two different methods (WEAT and Cosine Similarity) to analyze bias in the BERT algorithm, which will be discussed in more detail in the following sections. Section 5.1 gives an overview of BERT's pre-trained language model. Section 5.2 discusses the vocabulary data set of BERT that is used

for word embeddings. Section 5.3 explains the Word Embedding Association Test (WEAT), one of the methods used to analyze bias in BERT. Finally, Sect. 5.4 presents the cosine similarity method used to analyze bias towards word embeddings generated by BERT. We have used BERT pre-trained auto tokenizer and auto model for Tensorflow from the Hugginface transformer library to perform these tests. **BERT Base Cased** has been used for all tests since it was relevant with using names as a noun in sentence pairs.

5.1 BERT Algorithmic Discriminability

BERT's architecture for the training phase first forms a bidirectional representation of the sentence and then masks the target words with a special [MASK] token. This makes the prediction of the target word more difficult for the model. The masked language model (MLM) randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context. Even though this approach produces a bidirectional pre-trained model but introducing the [MASK] token during training creates a mismatch between the pre-training and fine-tuning input, which is mitigated by choosing only 15% of the token positions at random for prediction. [16] estimated the probability of a masked, gendered target word's association with an attribute word in a sentence using the MLM. However, the results in [24] show that the BERT models do not produce "effective" contextualized representations for words, and its improved performance can be the cause of fine-tuning or classifiers that model the dependencies explicitly by encoding syntactic patterns in the training data. Therefore if BERT is not introduced to certain words, such as Racial minority data (African American Names in our case) during its pre-training or training phase, it can not provide not only an effective but also a correct contextualized representation of the words.

5.2 BERT Vocabulary

BERT's pre-training corpus is developed by the BooksCorpus (800M words) [33] and English Wikipedia (2,500M words). BERT uses WordPiece [30] embeddings with a 30,000 token vocabulary. Our work used BERT-Base cased model, which includes a vocab file to map WordPiece to word id. After pursuing the vocab, we have looked for the names of African Americans along with the European Americans of both male and female genders. Section 6 will discuss the Notable results of this analysis. When BERT is introduced with a new word in a sentence that is not included in its pre-training vocabulary, it will use a word piece model that breaks the word embedding down into multiple sub-words, and looks up their embeddings in WordPiece, and then treats the word as separate embeddings of the subwords. However, if the introduced word does not have any sub-words, BERT will choose each alphabetical character of the word as a subword and finds its embedding from the WordPeice to create a representation of that word. We have analyzed the word embedding of all African American, and European

American names are using BERT's tokenization method. The notable results of this analysis are discussed in Sect. 6.

5.3 WEAT Test

Implicit Association Test (IAT) is a measure of the strength of associations between concepts, evaluations, and stereotypes [25]. The Word-Embedding Association Test (WEAT) is a statistical test analogous to the IAT, which is widely applied to the semantic representation of words in AI, and termed word embedding. Word embeddings represent each word as a vector in a vector space of about 300 dimensions, based on the textual context in which the word is found. In WEAT, the distance between a pair of vectors (their cosine similarity score) is used as analogous to reaction time in the IAT to compare a set of target concepts (e.g., Caucasian and African American names) denoted as **X** and **Y** (each of equal size N), with a set of attributes to measure bias over social attributes and roles (e.g., career/family words) denoted as **A** and **B**. The degree of bias for each target concept **t** is calculated as shown below, where **sim** stands for cosine similarity:

$$s(t, A, B) = [mean_{a \epsilon A} sim(t, a) - mean_{b \epsilon B} sim(t, b)] \quad (1)$$

To measure bias using this method, target noun phrases are the names of Racially classified groups and the attributes are professions and sentimental phrases. In this method, Model aim's in masking tokens to create unbiased setting and then predict them. An example of a masking sentence is shown in Table 1.

Table 1. Masked sentence set example for WEAT

Type	Example
Original sentence	Keisha is a scientist
Target masked	[MASK] is a scientist
Attribute masked	Keisha is a [MASK]
Target & attribute masked	[MASK] is a [MASK]

Following [1] and [16] who use WEAT, we measure the influence of the attribute **A**, which can be a profession or emotion, on the likelihood of the target (T), which denotes people of Caucasian and African American racial groups person: P(T—A). Overall assumptions to perform such analysis on BERT and expect valid results is as follow:

- likelihood of a token is influenced by all other tokens in the sentence
  - Thus likelihood of Target will be different based on presence of the Target:

$$P(T) \neq P(T—A)$$

– Therefore, likelihood of Target will be different depending on different racially classified groups:

$$P(T_{female}|A) \neq P(T_{male}|A)$$

Details of calculating the WEAT probability scores following [1] and [16] is as follow:

1. Take the Original sentence with Target and Attribute words (Demonstrated in **Original Sentence** row of Table 1)
2. Mask the Target word (Demonstrated in **Target Masked** row of Table 1)
3. Obtain the probability of target word in the sentence:

$$p_T = P(Keisha = [MASK]|sentence)$$

4. Mask both target and attribute word (Demonstrated in **Target and Attribute Masked** row of Table 1)
5. Obtain the probability of target word when the attribute is masked:

$$p_{prior} = P(Keisha = [MASK]|maskedsentence)$$

6. Calculate the association by dividing the target probability by the prior and take the natural logarithm

5.4 Cosine Similarity Test

Contextual embeddings are mostly transformations of embedding at layer 1 (having no contextual information) to layer 12 (pick up more contextual information with each layer once deeper into the network). We have used the output of the last layer as the word embedding in a sentence with Female and Male Names of the Caucasian and African American racial groups as the noun phrase and different attribute sets of Actor/Actress, Pleasant/Unpleasant, Rich/Poor and Wrong/Write. These sentences' cosine similarity has been measured between the gender and racial groups, and the results are demonstrated in Fig. 2 and 3.

6 Preliminary Results

The result of each of the tests ran on the data sets will be discussed in the following sections.

6.1 BERT's Vocabulary

As it has been discussed before, BERT decomposes each unknown word into subwords and character tokens that it can generate embeddings for so that it can average the sub word embedding vectors to generate an approximate vector and retain the contextual meaning of an original word. BERT's vocabulary contains:

- Whole words
- Sub-words occurring at the front of a word or in isolation ("em" as in "embeddings" is assigned the same vector as the standalone sequence of characters "em" as in "go get em")
- Sub-words not at the front of a word, which are preceded by '##' to denote this case
- Individual characters

The result of our analysis on the BERT's vocabulary and LMD's set of names indicates that Caucasian Names for both Female and Male genders are 100% included in BERT's vocabulary, and Only 11% of Female and 33% of Male African American Names are included in BERT's vocabulary shown in Fig. 1. In [16] study, the Words that are absent in the BERT vocabulary were removed from the targets, and Lack of racial minority-related names, i.e., "Targets" in BERT's vocabulary, can lead to not only social bias but also algorithmic bias in the architecture.

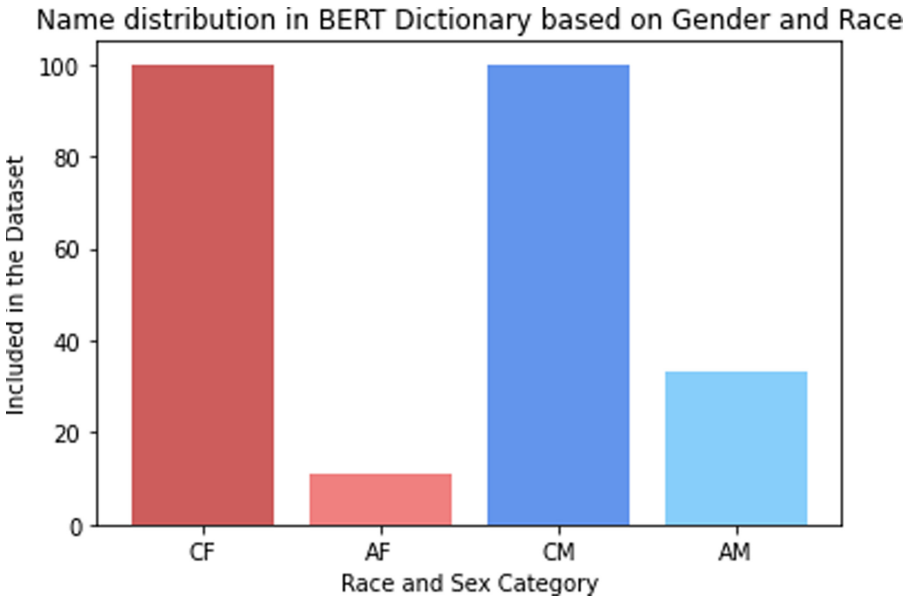


Fig. 1. Distribution of Caucasian and African American Names in BERT's vocabulary. C (Caucasian), A (African American), F (Female) and M (Male)

6.2 BERT's Algorithmic Discriminability

As shown in Table 1, you can observe that, since BERT does not include the African American names, it's treating them as unknown words and decomposing them into sub-words. The sub word embedding vectors to retain the contextual meaning of an original word which in the case of names it does not

bring the proper context by having single character's and words like "ish" in "Keisha", therefore this will change the overall meaning and context of a sentence that includes a name that's not included in BERT's vocabulary like Social Minority groups name. Table 2 represents the sub-word decomposition of African American names. This result accords with [24] work, proving that the complex interaction of "context dependent representation" and "reconstruction of input sentence" via transformers lead to geometric properties of the embeddings and subsequently affect the inherent discriminability of the resulting representations.

Table 2. Sub word decomposition of African American names

Original name	BERT's tokenization
Aisha	['Ai', '##sha']
Hakim	['Ha', '##kim']
Tamika	['Tam', '##ika']
Darnell	['Dar', '##nell']
Ebony	['E', '##bon', '##y']
Keisha	['Ke', '##ish', '##a']
Latoya	['La', '##to', '##ya']
Kareem	['Ka', '##ree', '##m']
Tanisha	['Tan', '##ish', '##a']
Rasheed	['Ra', '##she', '##ed']
Latonya	['La', '##tony', '##a']
Lakisha	['La', '##kis', '##ha']
Jermaine	['Je', '##rma', '##ine']
Tremayne	['T', '##rem', '##ayne']

6.3 WEAT

We have used the BERT pre-trained auto tokenizer and auto model for Tensorflow from the Hugginface transformer library to perform this test. The sentence templates BEC-Pro provided by [1] which we have used and replaced the gender-related pronouns to African American and Caucasian related names, were used to measure the association of target and attributes in sentences. The associations were measured using the templates shown in Table 1. As shown in the Second row of Table 1, we masked the "Target", applied the softmax function to the unscaled output of earlier layers predicted for the "Target" position in the sentence to obtain the probability distribution of the "Target" over the BERT vocabulary which using its vocabulary index the assassination probability value was obtained.

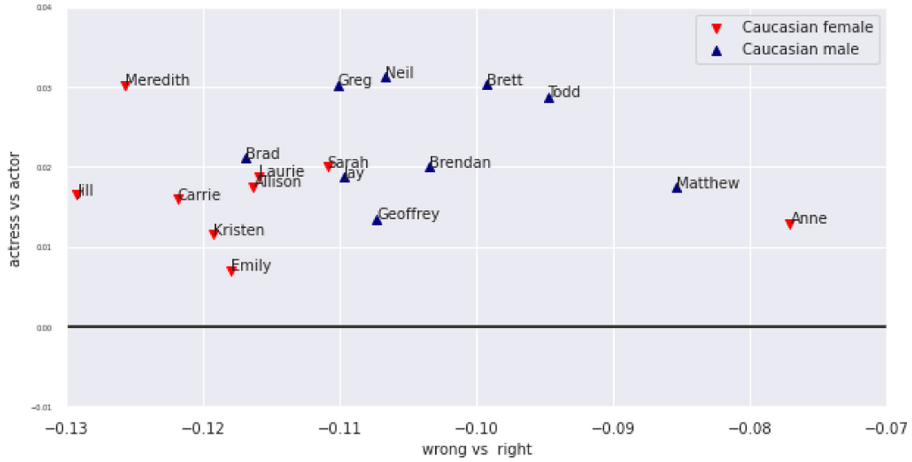
As shown in Table 3 there is a negative association between a target and an attribute for African American associated names, which means the probability of the target decreased through the combination with the attribute. However, there is a positive association value between the target and attribute for Caucasian associated names, indicating that the probability of the target increased through the combination with the attribute, with respect to its association probability. For both racial groups, the female gender has less increment in the probability of target combined with the attribute. This indicates that not only BERT is biased towards Racial minority groups but also it is biased towards gender minority groups, as it was illustrated in [1] work.

Table 3. WEAT results for classified racial and gender groups

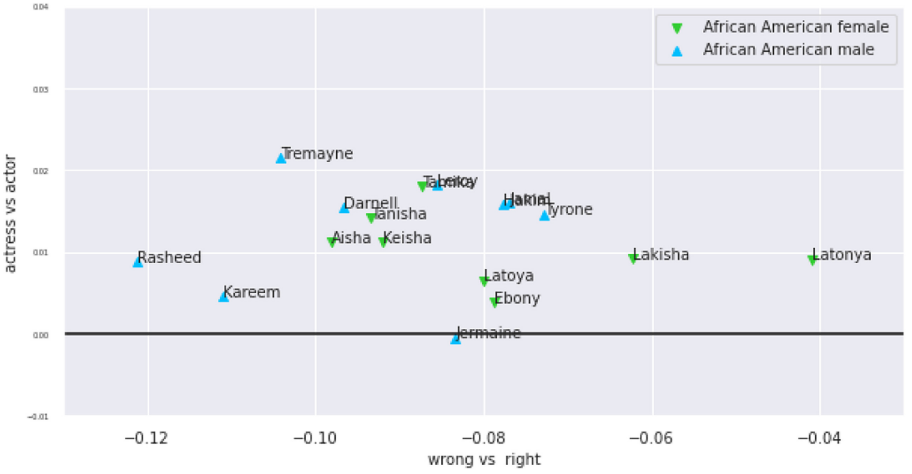
Race	Gender	Mean
African American	Male	−0.09
	Female	−0.12
Caucasian	Male	0.10
	Female	0.04

6.4 Cosine Similarity

It has been discussed by [11] that BERT was not optimized to get a well-shaped sentence embedding from it for unsupervised tasks, and it has the best performance on a supervised task where the classifier learns which dimension contribute and what the "scale" of those are. It has been further explained by [26] that BERT was not designed to produce helpful word/sentence embeddings that can be used with cosine similarities Because Cosine-similarity treats all dimensions equally, which puts high requirements for the created embed dings. We have analyzed the similarity of not just the sentences including both African American and Caucasian Targets and a sentimental Attribute such wrong/write, Actor/Actress, Pleasant and Unpleasant and the results demonstrated that for Caucasian names, two classes of female and male genders could be classified from each other using a boundary divider shown in Fig. 2 (A). Still, for the African American associated names, the two genders can not be classified into two different groups as if there is no similarity in their embeddings shown in Fig. 2 (B). Results also indicate that the word embedding for sentences including Caucasian male names is more biased towards positive sentiments such as pleasant and correct, shown in Fig. 3. This result is along with [4] observation that indicates having directional components from the embeddings that keep gender/racial features, causes into obtaining directions with a low magnitude of cosine similarity with actual gender/racial directions.



A



B

Fig. 2. Cosine similarity of Caucasian (A) and African American (B) associated names with sentimental attributes

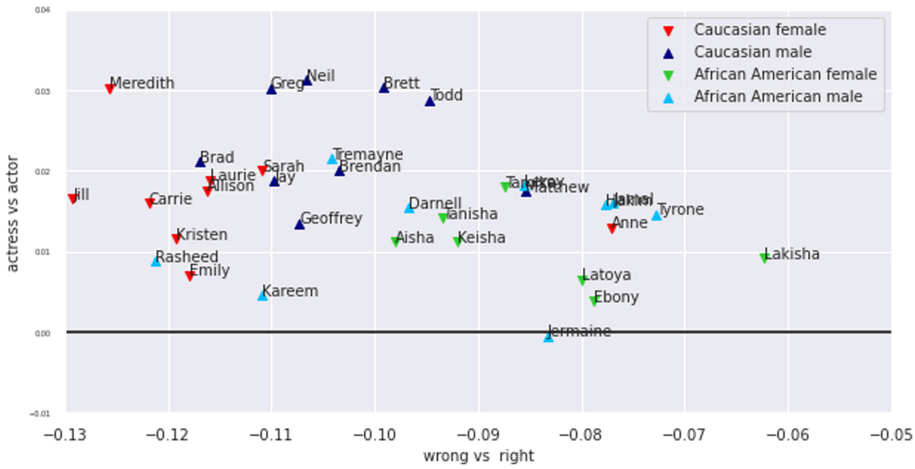


Fig. 3. Combined cosine similarity of both Caucasian and African American associated names with sentimental attributes

7 Conclusions

Among African Americans born in the last two decades, names provide a strong signal of socioeconomic status. AI recruitment platforms are widely adopted in industrial employee recruitment. The discrimination in recruitment algorithms affects these inequalities, which are difficult to assess. In this paper, we analyzed racial discrimination that can occur on such systems due to employee names. The results of our work illustrated that there exists an inherent discriminability that occurs during pre-training of BERT's algorithm. Future works and research on this study can analyze some mitigation strategies that can be applied to BERT and its architecture, such as having a more compiled vocabulary for the algorithm, training BERT with unique tokens for named entities and making them out from the raw text, using named entity recognition techniques to filter the names and replace them with an unknown token before feeding them into the pre-trained model.

References

1. Bartl, M., Nissim, M., Gatt, A.: Unmasking contextual stereotypes: measuring and mitigating Bert's gender bias. arXiv preprint arXiv:2010.14534 (2020)
2. Basta, C., Costa-Jussà, M.R., Casas, N.: Evaluating the underlying gender bias in contextualized word embeddings. arXiv preprint arXiv:1904.08783 (2019)
3. Bertrand, M., Mullainathan, S.: Replication data for: are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination (2019)
4. Bhardwaj, R., Majumder, N., Poria, S.: Investigating gender bias in Bert. arXiv preprint arXiv:2009.05021 (2020)

5. Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., Zemel, R.: Understanding the origins of bias in word embeddings. In: International Conference on Machine Learning, pp. 803–811. PMLR (2019)
6. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
7. Chaloner, K., Maldonado, A.: Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In: Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing, pp. 25–32 (2019)
8. Chang, K.-W., Prabhakaran, V., Ordonez, V.: Bias and fairness in natural language processing. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts (2019)
9. Church, K.W.: Word2vec. Nat. Lang. Eng. **23**(1), 155–162 (2017)
10. Ontario Human Rights Commission: Racism and racial discrimination: Systemic discrimination (fact sheet). Ontario Human Rights Commission. Accessed 17 July 2018
11. Devlin, J.: Question: why Bert underperforms issue #80 ukplab/sentence-transformers (November 2018)
12. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Hangartner, D., Kopp, D., Siegenthaler, M.: Monitoring hiring discrimination through online recruitment platforms. Nature **589**(7843), 572–576 (2021)
14. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018)
15. Kiritchenko, S., Mohammad, S.M.: Examining gender and race bias in two hundred sentiment analysis systems. arXiv preprint arXiv:1805.04508 (2018)
16. Kurita, K., Vyas, N., Pareek, A., Black, A.W., Tsvetkov, Y.: Measuring bias in contextualized word representations. arXiv preprint arXiv:1906.07337 (2019)
17. Liddy, E.D.: Natural Language Processing (2001)
18. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019)
19. Mozafari, M., Farahbakhsh, R., Crespi, N.: Hate speech detection and racial bias mitigation in social media based on Bert model. PLoS ONE **15**(8), e0237861 (2020)
20. Bureau of Labor Statistics: Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity. U.S. Bureau of Labor Statistics (January 2021)
21. Olteanu, A., Castillo, C., Diaz, F., Kıcıman, E.: Social data: biases, methodological pitfalls, and ethical boundaries. Front. Big Data **2**, 13 (2019)
22. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
23. Peters, M.E., et al.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
24. Podkorytov, M., Biś, D., Cai, J., Amirizirtol, K., Liu, X.: Effects of architecture and training on embedding geometry and feature discriminability in Bert. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
25. Harward Project Implicit. https://www.projectimplicit.net
26. Reimers, N.: Why cosine similarity of BERT, ALBERT, Robert is so big, almost near 1.0? · issue #2298 · huggingface/transformers (December 2019)
27. Sharma, S., Dey, M., Sinha, K.: Evaluating gender bias in natural language inference (2021)

28. Steed, R., Caliskan, A.: Image representations learned with unsupervised pre-training contain human-like biases. arXiv preprint arXiv:2010.15052 (2020)
29. Suresh, H., Guttag, J.V.: A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002 (2019)
30. Wu, Y., et al.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
31. Zhang, L., Wu, Y., Wu, X.: A causal framework for discovering and removing direct and indirect discrimination. arXiv preprint arXiv:1611.07509 (2016)
32. Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.-W.: Gender bias in contextualized word embeddings. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, pp. 629–634. Association for Computational Linguistics (2019)
33. Zhu, Y., et al.: Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 19–27 (2015)