# Discovering and Quantifying Bias in Arabic Contextualized Word Embedding

Fatime Houjaij

March 21, 2022

**Abstract**

Contextualized word embedding models, such as ELMo and BERT have become increasingly popular, replacing traditional context-free embeddings and attaining new state-of-the-art results in the majority of downstream NLP tasks. While BERT does provide a multilingual version, mBERT, it has been shown to not perform as well as models that are specifically trained on a single language. Therefore, dedicated language-specific models trained on BERT were developed. Arabic models include Antoun's et al. AraBERT and AraBERTv1 [2] and Abdul-Mageed's et al. ARBERT and MARBERT [1]. Bias is a recurring topics that pops out whenever we talk about word embeddings. Many studies were conducted on the bias in context-free embedding, a few were done for bias in contextualized word embedding. However, most of the studies done on bias in the contextualized sphere are done on the English models and, to the best of our knowledge, none has been done for bias in contextualized word embedding for Arabic models.

Our goal in this thesis is to uncover and quantify bias in these BERT-based Arabic models and if time permits find ways to combat these biases. To achieve this goal, we need to conduct our own study on the protected attributes in the Arab region that we want to test bias on, a data-set of these attributes and their respective stereotypes is needed. This data-set is going to be a considerable part of the research as some of them will be provided by academic sources, others will be translated from English sources when relevant and others we are going to have to generate ourselves. We are going to apply a range of recently introduced bias tests and choose a metric to quantify this bias in order to compare bias across the different Arabic models.

## 1 Background and Objective

Word embeddings is one of the most influential recent developments in NLP. Early pre-trained word representation models, while enabling significant advancement in Natural Language Processing NLP and Natural Language Understanding NLU tasks, neglected to consider the context in which the word appears in the embedding. Words were represented as static vectors in a continuous space which meant all forms of polysemous words had to have the same representation. Such issues led to the current replacing of static word embeddings with contextualized word representation models

like BERT and ELMO. Using word vectors that are sensitive to the context in which they appear, yielded massive improvements for the tasks at hand.

When we talk about NLP in downstream tasks we can not ignore the major issue of bias. Bias has been studied and shown to be present on several fronts in both context-free and contextualized word embeddings. Several studies were conducted on BERT to showcase the presence of bias when used in downstream tasks. Recognizing and combating these biases is crucial as word embeddings form the foundation of most language systems. Failing to recognize bias leads to unknowingly encoding these biases into language systems and therefore perpetuating and perhaps increasing dangerous cultural and racial stereotypes. Recognizing, measuring, and combating bias are topics of high importance nowadays.

Contextual models require massively large corpora and a high computational cost to train. These drawbacks restricted the availability of such models to English and a handful of other languages. To combat this, multilingual models like mBERT were trained. mBERT contains representations for over 100 languages, however, multilingual models' results fall short behind models trained on a single language. To remedy this some BERT-based Arabic models [1, 2] emerged. The aim is to set state-of-the-art results in Arabic and to achieve the same success that BERT had in English.

While there is evidence that models encode human biases, the amount of bias is not universal across all languages as shown by Lauscher and Glavaš in [7]. The focus of this thesis is on discovering and quantifying bias in BERT-based Arabic models. To the best of our knowledge, there are no studies done on bias in Arabic contextualized models. Lauscher et al. in [8] used a combination of bias tests to quantify and capture human-like biases in Arabic context-free word embeddings.

We aim to remedy the absence of studies for bias in contextualized Arabic word embeddings. In order to achieve this goal the following are the steps that need to be done:

- Decide on the protected binary, multidimensional, and intersectional attributes we want to check bias for.
- Develop the benchmarks needed to quantify bias in contextualized Arabic word embeddings.
- Develop appropriate bias tests and metrics that use the developed benchmarks to capture and quantify bias in currently available contextualized Arabic word embeddings.
- Compare the different benchmarks and different metrics and tests to quantify and compare bias across the currently-available BERT-based Arabic models and the multilingual ones.

To be able to achieve all of these goals, we have to address many challenges. The first challenge that we will tackle is building appropriate benchmarks. For that, we will be using and building and building on the translated specifications of the Word Embedding Association Tests (WEAT), AraWEAT, created by Lauscher et al. [8] and available on github. AraWEAT translates the dataset in the WEAT test developed by

Caliskan et al. [3] into Arabic using Google translate and makes sure to include the feminine and masculine variations of the word in order not to introduce a new bias. AraWEAT however, does not include the WEAT tests done on proper names since it does not make sense to translate them. However, it has been shown by Hall Maudslay et al. [10] that proper names are a good proxy for identifying bias towards specific groups of people. It is important to thus extend such benchmark with proper names that would represent different groups of people in the Arabic language, namely the male and female genders and people of different religions.

Moreover, we would like to also capture and quantify bias with respect to other protected attributes that are specific to the Arab region such as sects, religions, as well as combinations of protected attributes. We will thus need to extend our benchmark further to be able to use it to test for bias with respect to these protected attributes, instead of just with respect to gender. Finally, we need to adopt and extend bias metrics that are typically used to quantify bias in contextualized word embeddings to the case of Arabic and to extend them to handle more protected attributes and combinations of them, some of which might not be binary. For example, one such metric we will be adopting is Tan et al. modified the Sentence Encoder Association Tests SEAT [12], which is based on WEAT and developed by May et al. [11]).

# 2 Related Works

We review previous work done on uncovering and quantifying bias in contextualized word embedding and we introduce the different BERT-based Arabic models currently available.

## 2.1 Arabic BERT-based Models

AraBERT [2] was the first Arabic-specific BERT-based model and achieved state-of-the-art results in most Arabic NLP tasks that it was tested on. Later in [1] ARBERT and MARBERT collectively achieved new state-of-the-art results in comparison with AraBERT and mBERT.

In [2] Antoun et al. pre-trained BERT specifically for the Arabic language in the pursuit of achieving the same success that BERT did for the English language. They compared the performance of AraBERT to Google's multilingual BERT and other approaches. The results showed that AraBERT sets a new state-of-the-art performance for several Arabic NLP downstream tasks. AraBERT is pretrained on a collection of Arabic Corpus [5], OSIAN (Open Source International Arabic News Corpus), and manually scraped Arabic news websites. The dataset, after cleaning, consists of 70 million sentences (24GB of text). AraBERT also has a much larger vocabulary than multilingual models (60K) whereas mBERT and XLM-R have 5K and 14K, respectively.

In [1], Abdul-Mageed et al. improve on AraBERT by overcoming some of its limitations. First, it does not make use of only easily-accessible data. Second, AraBERT was only pre-trained on Modern Standard Arabic which limits its applications on tasks using dialects. Third, it has only been evaluated on three downstream tasks. Fourth,

it has not been compared to the current state-of-the-art multilingual model which is XLM-R [4]. [1] introduces two new models, ARBERT and MARBERT. ARBERT is pretrained on a large dataset (61 GB of text) and is focused on Modern Standard Arabic. ARBERT has an even larger vocabulary (than AraBERT) with over 100K words. MARBERT on the other hand, is based on the same network architecture as ARBERT, but is focused on both Dialectal Arabic and Modern Standard Arabic. The dataset consists of 1B Arabic tweets (128GB of text).

## 2.2 Detecting and Quantifying Bias

In [3], Caliskan et al. developed Word Embedding Association Test which measures bias by comparing the size of association between a concept X with attribute A and concept Y with attribute B, as opposed to concept X with attribute B and concept Y with attribute A. Lauscher et al. [8] created AraWEAT which contains the translated dataset of WEAT. In [11], May et al. proposed a simple generalization of WEAT from words to sentences and phrases: Sentence Encoder Association Test (SEAT). Now we talk about the paper, Assessing Social and Intersectional Biases in Contextualized Word Representations where Tan and Celis [12], adapted SEAT to use in contextual embeddings by using the contextual representation of the word of interest in the sentence instead of the sentence encoding when calculating associations. In this study, they started by showing the significant gender imbalance in the standard corpora that embeddings' models are pre-trained on. They showed that both sentence encoding and contextual world representation must be used in assessing bias as some biases are only uncovered by one and not the other. They also provided evidence on how racial bias is more strongly encoded than gender bias. And finally, they proved that least privileged groups (i.e., those with more than one minority attribute) suffer more bias than each of their constituent identity.

In the work, by Kurita et al. [6], the authors claim that the reason May et al. SEAT had inconsistent results was because the cosine method is not ideal for BERT. Instead of adapting SEAT as done in [12], they proposed a new method in capturing social bias for contextualized word embedding and show that it is more consistent than the traditional cosine-based method. They conducted a case study on the downstream task of Gendered Pronoun Resolution. This new method utilizes the fact that BERT is a masked language model and it calculates the increased log probability based on BERT's probability to predict certain target tokens versus others.

Although the work by Manzini et al. [9] mainly focuses on bias in context-free embeddings, however it also handles cases where the attribute we are testing for is not binary (example: the multi-class religion instead of the binary gender). Their approach is motivated by WEAT and it calculates the Mean Average Cosine Similarity (MAC) as a measure of bias. We have a set of target word embeddings $T$ containing terms that inherently contain some form of social bias and a set $A$ which contains sets of attributes that should not be associated with $T$. They defined a function that computes the mean cosine distance between a particular target $T_i$ and all terms in a particular attribute set $A_j$. We aim to apply the same logic used by Tan and Celis [5] to modify this WEAT-based approach in a multi-class setting and for contextualized word embeddings.

# 3 Time Table

| Task | May | Jun-Jul-Aug | Sep-Oct | Nov | Dec-Jan |
|---|---|---|---|---|---|
| Literature Review | | | | | |
| Data-set Collection and Translation | | | | | |
| Developing and Applying a Bias Metric | | | | | |
| Experiments & Comparisons | | | | | |
| Thesis Writing & Defense | | | | | |

# References


[1] Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*, 2020.

[2] Fady Baly, Hazem Hajj, et al. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, 2020.

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

[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[5] Ibrahim Abu El-Khair. 1.5 billion words arabic corpus. *arXiv preprint arXiv:1611.04033*, 2016.

[6] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*, 2019.

[7] Anne Lauscher and Goran Glavaš. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. *arXiv preprint arXiv:1904.11783*, 2019.

[8] Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. Araweat: Multidimensional analysis of biases in arabic word embeddings. *arXiv preprint arXiv:2011.01575*, 2020.

[9] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*, 2019.

[10] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. *arXiv preprint arXiv:1909.00871*, 2019.

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

[12] Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. *arXiv preprint arXiv:1911.01485*, 2019.