Roadblocks in Gender Bias Measurement for Diachronic Corpora

Saied Alshahrani Esma Wali Abdullah R Alshamsan Yan Chen
Jeanna Matthews
Department of Computer Science
Clarkson University, Potsdam, NY, USA
alshahsf,walie,alshamar,cheny3,jnm@clarkson.edu

Abstract

The use of word embeddings is an important NLP technique for extracting meaningful conclusions from corpora of human text. One important question that has been raised about word embeddings is the degree of gender bias learned from corpora. Bolukbasi et al. (2016) proposed an important technique for quantifying gender bias in word embeddings that, at its heart, is lexically based and relies on sets of highly gendered word pairs (e.g., mother/father and madam/sir) and a list of professions words (e.g., doctor and nurse). In this paper, we document problems that arise with this method to quantify gender bias in diachronic corpora. Focusing on Arabic and Chinese corpora, in particular, we document clear changes in profession words used over time and, somewhat surprisingly, even changes in the simpler gendered defining set word pairs. We further document complications in languages such as Arabic, where many words are highly polysemous/homonymous, especially female professions words.

Keywords

word embedding, gender bias, NLP, Arabic, Chinese, profession words, diachronic

TLR

We document hurdles in applying a popular gender bias measurement technique using word embeddings of profession words and highly gendered word pairs for diachronic corpora in Arabic and Chinese.

1 Introduction

Natural Language Processing (NLP) plays a significant role in many powerful applications such as speech recognition, text translation, and autocomplete and is at the heart of many critical automated decision systems making crucial recommendations about our future world. Word embedding systems are widely used to represent text data as vectors and enable NLP computation. Systems such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and BERT (Devlin et al., 2018) ingest large corpora of human text and can be used to learn semantic and syntactic relationships between words.

At the same time, it has been demonstrated that these systems learn a wide variety of societal biases embedded in human text including racial bias, gender bias, and religious bias (Caliskan et al., 2017; Abid et al., 2021). In a widely cited paper, Bolukbasi et al. (2016) demonstrated that a system trained with a corpora of Google News would complete the word comparison "man is to computer programmer as woman is to what?" with the response "homemaker" suggesting an alarming level of gender bias when used in tasks such as sorting resumes for computer programming jobs. Chen et al. (2021) extended these techniques beyond English to eight other languages (Chinese, Spanish, Arabic, German, French, Farsi, Urdu, and Wolof) and applied them to Wikipedia corpora in each of these languages. They documented persistent gender bias and lack of representation in the modern NLP pipeline.

NLP research often uses large, modern datasets like Google News and Wikipedia. Developers of a wide variety of NLP-based applications begin with large pre-trained models that are also based on large corpora of human text (Bender et al., 2021). These pre-trained models also largely reflect the speech/writing of modern English speakers producing digital text. The speech/writing of speakers of the more than 7,000 languages spoken worldwide is often under-represented (Wali et al., 2020). Similarly, historical speech/writing is often under-represented despite the fact that historical speech/writing is often considered foundational to cultural identity. Investments in multilingual NLP

Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 140 - 148

and processing of diachronic corpora are essential if we want our NLP-based automated decision making systems to more widely reflect foundational cultural norms and identity from around the world.

The inspiration for this paper was to re-examine Bolukbasi et al.'s popular NLP-technique for quantifying gender bias from the perspective of applying it to diachronic corpora in Arabic and Chinese. Specifically, Bolukbasi et al.'s method begins with identifying a set of profession words and a set of highly gendered word pairs (defining set). In this paper, we explore the degree to which these words might change over time. We document ways in which this method is fundamentally fragile for diachronic corpora because of the way these sets of words would change over time.

In Section 2, for background, we elaborate on Bolukbasi et al. and Chen et al.'s multilingual extensions and some other relevant related work. Section 3 describes our experience with two different diachronic Arabic corpora, especially the impact on changes in profession set words over time. In Section 4, we discuss changes in some defining set words in Chinese using the Google Ngram Viewer. We conclude and discuss future work in Section 5.

2 Background and Related Work

Bolukbasi et al. (2016) pioneered a method for quantifying the amount of gender bias learned in by word embedding systems and many researchers have built on their techniques including Chen et al. (2021) who observed substantial hurdles in extending the techniques beyond English. In this paper, we build on both Bolukbasi et al. and Chen et al.'s work to examine additional hurdles that would arise when attempting to apply these techniques to diachronic corpora.

Bolukbasi et al.'s original method is based on two sets of words. The first set (the defining set) consists of 10 highly gendered word pairs (she-he, daughter-son, her-his, mother-father, woman-man, gal-guy, Mary-John, girl-boy, herself-himself, and female-male) and the second (profession set) consists of 327 profession words such as nurse, teacher, writer, engineer, scientist, manager, driver, banker, musician, artist, and chef. They used the difference between the defining set word pairs to define a gendered vector space and then evaluated the relationship of the profession words relative to this gendered vector space. Ideally, profession words would not reflect a strong gender bias. However, in practice, they often do. According to such a metric, the word doctor might be male biased or the word nurse female biased based on how these words are used in the corpora from which the word embedding model was produced.

Bolukbasi et al. (2016) uses these two sets of words to compute a gender bias metric for each word and from there to express the gender bias of a corpora. Specifically, each word is expressed as a vector by Word2Vec and then the center of the vectors for each defining set pair is calculated. For example, to calculate the center of the definitional pair woman/man, they average the vector for "woman" with the vector for "man". Then, they calculate the distance of each word in the definitional pair from the center by subtracting the center from each word in the pair (e.g., "woman" - center). They then apply Principal Component Analysis (PCA) to the matrix of these distances. PCA is an approach that compresses multiple dimensions into fewer dimensions, ideally in a way that the information within the original data is not lost. Usually, the number of reduced dimensions is 1-3 as it allows for easier visualization of a dataset. Bolukbasi et al. (2016) used the first eigenvalue from the PCA matrix (i.e. the one that is larger than the rest). Because the defining set pairs were chosen to be highly gendered, they expected this dimension to be related primarily to gender and therefore called it the gender direction or the g direction. Finally, the g direction is a vector, and there is a vector representing each word. Therefore, they used cosine similarity between the vector for each word, w, and the g direction vector as the measure of gender bias for that word. For a corpora or other collection of words, one can average the gender bias of words contained in the corpora as a measure of gender bias in the corpora using the equation of Bolukbasi et al. (2016) for the direct gender bias of an embedding:

$$\text{DirectBias}_c = \frac{1}{|N|} \sum_{w \in N} |\cos(\vec{w}, g)|^c$$

where N is the given gender neutral words, and c is a parameter that determines the strictness in measuring gender bias.

Chen et al. (2021) extended the Bolukbasi et al.' method to eight languages besides English - Chinese, Spanish, Arabic, German, French, Farsi, Urdu, and Wolof. In order to do so, they first made modifications to the defining set to make it more translatable across the 9 languages. For

example, they dropped pairs like she-he, her-his, gal-guy, Mary-John, herself-himself, and female-male because of problems in translation for some languages and adding pairs like queen-king, wife-husband, and madam-sir. Second, they observed that the Bolukbasi et al.'s method cannot be applied directly to languages such as Spanish, Arabic, German, French, and Urdu that primarily use grammatically gendered nouns (e.g., escritor/escritora in Spanish vs. writer in English). They solved this problem using a weighted average of the number of occurrences of each variant of the professional word (male, female, or neutral) multiplied by the gender bias score for that variant.

In this work, we build on both (Bolukbasi et al., 2016; Chen et al., 2021) and focus on the unique challenges that arise when applying these techniques to diachronic corpora. Specifically, we examined changes in both the profession set and defining set over time in Arabic and Chinese. Certainly, professions have changed drastically over that amount of time and so a method based on profession set words like Bolukbasi et al.'s method will have substantial challenges. We explored this using corpora including a database of Arabic poems spanning 11 eras from the Pre-Islamic period (before 610) to modern day. While we saw less change over time in the usage of the simpler defining set words than in the profession set words, we did observe some interesting changes in even the defining set words over time, especially in Chinese. In the process of this work, we also documented further complications in languages such as Arabic, where many words are highly polysemous/homonymous, especially female professions words.

Wevers (2019) also used word embeddings to examine gender bias over time. They used a collection of Dutch Newspaper articles spanning over four eras (1950-1990), training four embedding models per newspaper, one per era, using the Gensim implementation of Word2Vec to demonstrate how word embeddings can be used to examine historical language change. They observed clear differences in gender bias and changes within and between newspapers over time. Slight shifting of bias was observed in some themes like shifting towards female bias in themes related to sexuality and leisure (mostly seen in newspapers with religious background). Shifting towards male bias in themes related 'money', 'grooming', and negative emotions, especially in newspapers with a liberal background, was also observed.

Rudolph and Blei (2018) developed dynamic embeddings building on exponential family embeddings to capture the language evolution or how the meanings of words change over time. They used three datasets of the U.S. Senate speeches from 1858 to 2009, the history of computer science ACM abstracts from 1951 to 2014, and machine learning papers on the ArXiv from 2007 to 2015. They demonstrated how words like Intelligence, Iraq, computer, Bush, data change their meaning over time. They observed that the dynamic embeddings provided a better fit than classical embeddings and captured interesting patterns about how language changes. For example, a word's meaning can change (e.g., computer); its dominant meaning can change (e.g., values); or its related subject matter can change (e.g., Iraq).

Xu et al. (2019) demonstrated the characterization of the semantic weights of subword units in the composition of word meanings. They used a subword-incorporated or a word embedding model variant for the evaluation and revealed interesting patterns change in multiple languages. Their training datasets consist of Wikimedia dumps for 6 Languages (up until July 2017) consisting of Chinese and other Indo-European languages like English, French, German, and Italian. The results revealed major differences in the long-term temporal patterns of semantic weights between Chinese and five Indo-European languages. For example, in Chinese, the weights on subword units (characters) show a decreasing trend, i.e., individual characters play less semantic roles in newer words than older ones whereas the opposite trend was observed in other languages. Therefore, Chinese words are treated more as a whole semantic unit "synthetically", while words in Indo-European languages require more attention into the subword units "analytically". These results provide evidence towards word formations to the linguistic theories. For example, the notion of "word" in Chinese is always changing: Modern Chinese has multiple characters as a whole semantic unit opposite to its older counterpart. The semantic weight carried by a single character is decreasing over time. This is strong evidence in support of the claim that Chinese has been evolving towards more detailed multisyllabic words from concise and monosyllabic words.

Time Periods	Number of Books	Vocab Size	Token Size
Books Before Islam	3	16,460	39,255
Books Before 1900	2,820	2,075,505	566,366,883
Books After 1900	773	1,335,027	136,870,579
Duplicate Books	11	-	-
Unknown Books	2,931	-	-
All Shamela's Books	6,527	2,520,372	703,276,717

Table 1: Measurements of Shamela Library dataset in terms of the number of books, vocabulary size (unique words), and token size (all words) for each time period. We did not train a GloVe model on the unknown books alone or the duplicate books and therefore are not reporting vocab size and token size.

3 Changes in Arabic Over Time

Building on both Bolukbasi et al. (2016) and Chen et al. (2021), we consider how the sets of profession words required by the Bolukbasi et al.'s method would need to change over time in Arabic. We begin by describing two diachronic datasets that we used and how we processed these datasets, then we describe the changes in the profession word usage over time.

3.1 Datasets and Methodology

In this paper, we use two Arabic datasets: Shamela Library (المكتبة الشاملة) that is released by Shamela Library Foundation (2012), and Arabic Poem Comprehensive Dataset (APCD) by (Yousef et al., 2018). Shamela Library is a free project that collects thousands of Islamic religious and other related sciences books. APCD is a collection of Arabic poems spanning 11 eras, from the Pre-Islamic (before 610) to the Modern age (1924 - Now). Arabic NLP researchers commonly use these two datasets to study Arabic classics.

We processed the Shamela Library dataset version of 6,538 Arabic books (6,527 unique books after removing duplicates) in Microsoft Word format (1997-2004).[1] The books in this corpora were not labeled according to the publication dates. Thus, to study the language change over time in the Arabic language, we further classified Shamela's Arabic books into three different time periods based either on their publication date or the authors' date of death when publication date was not available. We identified books written before Islam or before 610 (only three books), books written before 1900 (2,820 books), and books written on or after 1900 (773 books). We were not able to identify publication dates or the authors' dates of death of the remaining 2,931 books due to not having any; Table 1 summarizes some key attributes of this dataset.

We also processed the APCD, an Arabic poetry dataset that is collected mainly from the Poetry Encyclopedia (الموسوعة الشعرية) that is released by Abu Dhabi Department of Culture and Tourism (2016) and Diwan (الديوان) (Diwan, 2013). Unlike Shamela, this dataset was already labeled by era, making it a good choice for studying language change over time. It has, before preprocessing, approximately 1,831,770 poetic verses labeled by their meter, the poet's name, and the era they were written in. One drawback of this corpora is that it is relatively small. Table 2 summarizes some key attributes of this dataset.

We then produced a total of 16 GloVe models (Pennington et al., 2014) from the three time periods of Shamela, the 11 eras of APCD, all Shamela, and all APCD.[2] Each GloVe model is a context-independent model that produces a one-word vector (word embedding) for each word even if that word appears in the context a few times unlike BERT and ELMo (Devlin et al., 2018; Peters et al., 2018). Each GloVe model provides vocabulary size, token size, and word vectors. It is important to note that before training GloVe models, it was necessary to preprocess the two datasets using Linux/Unix command-line utilities like `tr` (for translating or

[1] We contribute the scripts we wrote to process these corpora and overcome several challenges with the data. For example, one challenge we faced was correctly converting back and forth between the Arabic Windows-1256 to the Unicode (UTF-8) encoding schemes. The Arabic books were written in an old version of Microsoft Word (1997-2004), which caused encoding scheme conversion errors, resulting in unreadable characters by native Arabic speakers or even NLP tools. Scripts can be found here: https://github.com/Clarkson-Accountability-Transparency/gBiasRoadblocks

[2] Bolukbasi et al. (2016) used Word2Vec to generate word embeddings, and in this paper, we chose GloVe instead because GloVe performs better than Word2Vec in the Arabic language (Naili et al., 2017)

Eras	Poetic Verses	Vocab Size	Token Size
Pre-Islamic (before 610)	21,907	60,082	204,450
Islamic (610-661)	2,942	12,388	24,461
Umayyad (661–750)	63,776	119,533	610,563
Between Umayyad and Abbasid	24,077	65,220	221,058
Abbasid (750–1258)	234,494	252,339	2,156,195
Andalusian (756–1269)	111,011	151,503	1,024,653
Fatimid (909–1171)	124,129	172,460	1,171,842
Ayyubid (1174–1252)	112,350	152,165	1,061,503
Mamluk (1250–1517)	164,780	198,748	1,550,669
Ottoman (1517–1924)	159,576	186,795	1,492,132
Modern (1924 - Now)	778,723	462,478	7,146,135
All APCD's eras	1,797,765	736,576	16,663,658

Table 2: Measurements of Arabic Poem Comprehensive Dataset in terms of number of poetic verses, vocabulary size (unique words), and token size (all words) for each era.

deleting characters), `sed` (for filtering and transforming text), `iconv` (for converting between encoding schemes), and `awk` (for pattern scanning and language processing), along with CAMeL tools (Obeid et al., 2020), an open-source python toolkit for Arabic NLP, to dediacritize the Arabic diacritical marks and remove unnecessary characters.

3.2 Modern and Historical Professions

We began with a consideration of how the profession sets used in Bolukbasi et al. (2016) and Chen et al. (2021) would need to change over time. First, we identified 50 modern profession words that we expect would simply not exist in the older time periods/eras in Shamela and APCD datasets.[3] For example, the profession of electrician would not have existed before the advent of electricity. Second, we identified 50 historical profession words that we think exist in older time periods/eras in Shamela and APCD datasets but which are much less common in modern times.

As in Chen et al. (2021), we further categorized each word based on gender. In Arabic, most profession words have a male variant and a female variant in which the spelling is changed slightly based on gender, for example female pilot (طيَّارة) and male pilot (طيَّار). Linguistically, many professions that would be extremely uncommon for men or women do have a male or female version of the word (e.g., it is rare for a woman to have the profession chamberlain/head of staff (حَاجِب), but there is a female word for that profession). However, in some cases, either the male or female version does not even exist linguistically (e.g., there is no male word of midwife (قَابِلة) profession). There are also more rare neutral words, like musician (موسِيقَار), that is used for both genders with no spelling changes.

In the APCD dataset, we found, as expected, that there are some modern professions that occur noticeably only in the modern era of the Arabic poems, but do not appear at all in the previous historical eras, such as the male engineer (مُهَندِس) that occurs 17 times, and the neutral profession of an electrician (كَهرَبائِي) that occurs only four times in the modern age, indicating that those modern professions are increasingly appearing in the modern age of the Arabic poems and confirming that Arabic native speakers (i.e., Arabs) still use the poems as an effective way to document the Arabic language changes over time.

On the other side of history, in the Shamela dataset, we found that a few historical professions frequently occur in the time periods before 1900 but not significantly after 1900. Some professions reflect essential shifts in legality. For example, one profession that is fortunately no longer legal or acceptable is male slaver (نَخَّاس). Fortunately, the male slaver profession appears much less often (only 12 times) in the time period after 1900, while it appears unpleasantly 118 times before the 1900 time periods. As another example, male chamberlain/head of staff (حَاجِب) appears 9,518 before the 1900 time periods, but only appears 914 times in

[3] We point to an expanded technical report with the full list of used modern and historical profession words. The report can be accessed here: https://lin-web.clarkson.edu/~jmatthew/LChange2022/

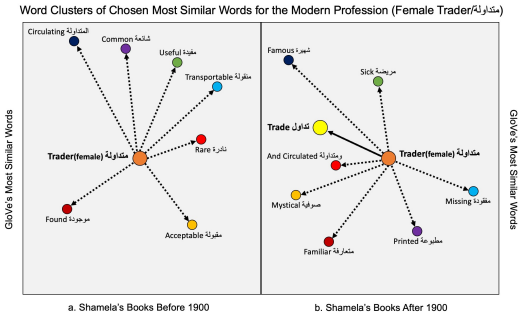

Figure 1: a. A word cluster of chosen GloVe's most similar words of the female profession trader (مُتَدَاوِلَة) in Shamela Library dataset in the time period before 1900, demonstrating that its word cluster is including different words with different meanings due to its homonymy. b. A word cluster of chosen GloVe's most similar words of the female profession trader (مُتَدَاوِلَة) in Shamela Library dataset in the time period after 1900, illustrating that a new related-trading activity word joining the profession word cluster, (trade/تَدَاوُل)[4]

the time period after 1900, showing that this male profession/position is on its way to extinction.

3.3 Polysemous/Homonymous Professions

The Arabic language is one of the most morphologically rich languages, with a high level of orthographic ambiguity, causing native speakers to use the optional diacritical marks to differentiate between two words (Grosvald et al., 2019).[5]

We noticed in the Shamela Library dataset that a few modern profession words change their connotations over time, and many profession words have alternate meanings due to the Arabic's orthographical ambiguity. We also found that this was especially true of female profession words. For example, the word (مُدَرِّسَة) for female teacher also means a school building (مَدْرَسَة), another word (طَيَّارَة) for a female pilot also means an airplane (طَيَّارَة). In all these cases, this complicates the use of both word counts and word embeddings in tracking the relative uses of profession words over time.

One homonymous example is the female trader (مُتَدَاوِلَة) profession. The same word (مُتَدَاوِلَة) also means common, famous, familiar, or circulating to describe a current news event. We see this alternate meaning dominate the usage of the word, complicating any attempt to study the prevalence of females engaged in this profession. Interestingly, we see evidence of change over time in the usage of this word. To investigate the semantic meaning of related words to the trading activity, we studied GloVe's most similar words (calculated based on the cosine similarity between two word vectors) for this profession word in two time periods of the Shamela Library dataset: before 1900 and after 1900. As shown in Figure 1a, before 1900, none of most similar words reflect the trading profession word (مُتَدَاوِلَة). However, in Figure 1b, after 1900, we see a word related to trading activity (trade/تَدَاوُل) appear in the most similar words of GloVe model. Thus, the connotation of the female trader (مُتَدَاوِلَة) profession is changing over time to more often reflect the actual profession of female trader (مُتَدَاوِلَة) and not just the alternate meaning of current news events.

3.4 Illegal Professions

In the religion of Islam, some professions are forbidden, for example, all types of usury, and serving, selling, or drinking alcohol. We examined a set of illegal/religiously forbidden profession words in Islam across the 11 ages of the Arabic poems, such as male usurer (مُرَابِي), female usurer (مُرَابِية), male bartender (سَاقِي), and female bartender (سَاقِيَة). Specifically, we closely focused on the diachronic semantic meaning change of the bartending profession words in the parallel eras of the APCD dataset. Interestingly, we found that bartending profession words in the early ages of the Arabic poems like Pre-Islamic, Islamic, and Umayyad only point to providing water to people but not serving wine even though the wine does exist. Those bartending profession words are polysemous and could carry other meanings like the male bartender (سَاقِي) could have a meaning of the phrase 'my leg' (سَاقِي), while the female bartender (سَاقِيَة) could have as well the

[4]English translations of the word clusters are automatically generated using Google Translator API that is included in the deep-translator Python model (https://deep-translator.readthedocs.io).

[5]In our preprocessing, we removed the optional diacritical marks as is generally recommended for Arabic NLP as a first step to reducing some data sparsity (Obeid et al., 2020). Unfortunately, removing diacritical marks increases the orthographic ambiguity, but retaining them would lead to a high degree of variance for the same word because the placement of diacritical marks varies with the grammatical placement of the word in a sentence. It is a difficult tradeoff for Arabic NLP that other researchers are attempting to tackle with advanced techniques, such as stemming and lemmatization (Kadri and Nie, 2006; Mubarak, 2017).

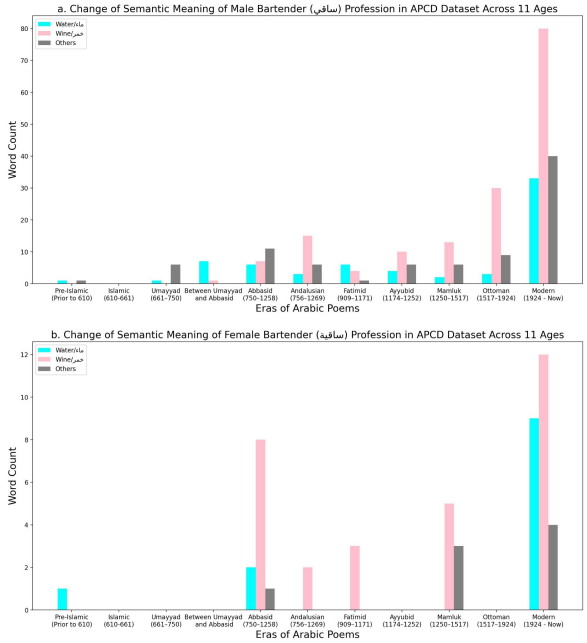

Figure 2: a. A word count of the occurrence of the male bartender (سَاقِي) across the 11 ages of the Arabic poems in the APCD dataset, showing the related meanings of the profession word like serving water, wine, or could be entirely meaning something that entirely unrelated to the profession word's meaning of serving drinks. b. A word count of the occurrence of the female bartender (سَاقِيَة) across the 11 ages of the Arabic poems in the APCD dataset, showing the related meanings like serving water, wine, or could be entirely meaning something that entirely unrelated to the profession word's meaning of serving drinks.

meaning of 'a water creek or an aqueduct' (سَاقِيَة).

To thoroughly investigate the occurrence of those profession words regarding their correlation with water – the allowed/halal drink, and the wine — the forbidden/haram drink in Islam, we manually analyzed the Arabic poems of each age and decided whether that word occurrence is a water-related meaning, wine-related meaning, or other unrelated meanings to both of the drinks. Figure 2a shows that the male bartender (سَاقِي) profession word started to appear in the Arabic poems as a profession of serving alcohol generally, wine exclusively, as a symbol of love, passion, and adoration for women from the age of between Umayyad and Abbasid until the Modern age.

One example of that is when the Abbasid Arabic poet, Abu Bakr Al-Sanobi (أبو بكر الصنوبري), said in his famous poem, the Pole of Pleasure in the Descriptions of Wines (قطب السرور في أوصاف الخمور): "O bartender of wine, do not forget us, O Goddess of Oud, spur singing (أيَا سَاقِي الخَمرِ لا تَنسنا – ويا ربَّة العُودِ حُثِّي الغِنَا)." Another example of that in another age, the Ottoman age, is for the Arabic poet, Abdul Ghani Al-Nabulsi (عبد الغني النابلسي), said in this romantic poem, Bartender O Bartender (ساقي يا ساقي): "Bartender O bartender, Give me some of his remaining wine (سَاقِي يَا سَاقِي – اسقِينِي مِن خَمرِهِ البَاقِي)

Similarly, in Figure 2b, the female bartender (سَاقِيَة) started to appear as a profession of serving wine from the age of between Umayyad and Abbasid until the Modern age as same as the male bartender (سَاقِي) profession word, except they did not appear in the two ages of Ayyubid and Ottoman. While the female and male bartender (سَاقِي و سَاقِيَة) surprisingly appeared in correlation with wine in the Arabic poems despite its religious forbiddance, both of the two profession words also refer to water-related words. For example, the female bartender (سَاقِيَة) refers to the 'water creek or aqueduct.' One example to show that is when the Modern Arabic poet, Rashid Ayoub (رشيد أيوب), said in his poem: "I sat in the meadow alone at the water creek, in which the water echoed the sound of my melodies",

جَلَستُ فِي الرَّوضِ وَحدِي عِندَ سَاقِيَةً

يُرَدِّدُ المَاء فِيهَا صَوت أَلحَانِي

4 Changes in Chinese Over Time

Although our primary focus in this study has been on Arabic, we found interesting evidence of change over time in Chinese as well. Classical Chinese (before 1900) uses a vocabulary and grammar that differs significantly from modern Chinese. We were surprised to find evidence not just of changes in professions over time, but also changes in defining set words. As we found in the diachronic corpora in Arabic, we expected changes in profession words over hundreds of years, but thought that the more fundamental defining set words like woman/man, girl/boy and madam/sir would not change substantially.

In Chinese, the word 'woman' can be translated in many ways, including "女子", "女人", and "妇女". The word "女子" was popularly used in an-

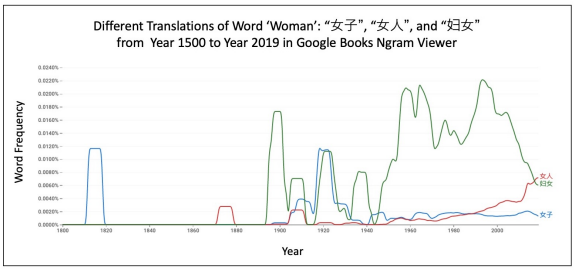

Figure 3: A timeline of word frequencies of different translations of word 'woman': "女子", "女人", and "妇女" that were found in multi-sources printed between 1500 and 2019 using Google Books Ngram Viewer.

cient times, but its usage has decreased in modern writing. In Figure 3, we used Google Books Ngram Viewer to chart the word frequencies of the different translations of the word 'woman': "女子", "女人", and "妇女" found in sources printed between 1500 and 2019 in Google's Books corpora in English, Chinese, French, German, Hebrew, Italian, Russian, or Spanish (Karch, 2021). This shows us that as languages evolve over time, defining sets, like profession sets, may also have to evolve to measure gender bias using methods like the Bolukbasi et al. (2016)'s method.

Besides using Google Books Ngram Viewer, we also assembled a small collection of works that might be considered "classics" in Chinese spanning the period 475 BC - 1992, for example 司马迁 (Records of the Grand Historian) by Qian Sima, 萧红 (Tales of Hulan River) by Hong Xiao, and 论语 (The Analects). We found that roughly half of the profession words used by Chen et al. (2021) did not appear, and that also two of the defining set words "boy" and "madam" used did not appear. Interestingly, Google Books Ngram Viewer showed that the word 'madam' was used very frequently between 1905 and 1910, but our small classics corpora did not include texts written in that time period. Again, these results indicate that as languages evolve over time, profession sets and even defining set words would have to evolve to measure gender bias.

5 Conclusion and Future Work

In order for NLP to reflect the rich multilingual, multicultural, and historical heritage of human text, it is essential that NLP techniques be extended beyond modern digital English text to multilingual and diachronic corpora. In this paper, we have explored the challenges of applying an important technique for measuring the gender bias learned by word embedding systems to diachronic corpora. We also have shown how techniques like those pioneered by Bolukbasi et al. (2016) and extended by Chen et al. (2021) have fundamental limitations when analyzing corpora spanning large periods of time. We showed that their technique based on analyzing the gender bias of profession words would have difficulty because professions change drastically over hundreds of years. Interestingly, we also documented changes in defining and profession set words over time and also challenges with polysemous/homonymous profession words especially female profession words in Arabic.

In this paper, we have focused mostly on identifying the problems with techniques applied successfully to measure gender bias in modern corpora like Google News or Wikipedia. In the future work, we plan to focus more on modifying profession sets and defining sets over time to overcome these problems. Our results indicate that as languages evolve over time, defining sets and profession sets would have to evolve to measure gender bias.

In this study, we focused on Arabic and Chinese, but we would like to extend our work to more languages. Adding an English corpora may be our next step. Although we like to actively focus on languages besides English, English can serve as an important comparison point because so much of the modern NLP tool chain has been optimized for English. We may be able to study the impact of changes in profession sets and defining sets over time with fewer complicating factors. We would also like to experiment with different advanced Arabic NLP techniques like stemming and lemmatization (Kadri and Nie, 2006; Mubarak, 2017) and see how applying such techniques could improve the results and reduce Arabic's orthographical ambiguity or even other Arabic NLP-related current issues like correcting spelling errors, especially in Arabic dialects, where there are no official orthography rules (Habash et al., 2018).

6 Acknowledgments

We'd like to thank the Clarkson Open Source Institute for their help and support with infrastructure and hosting of our experiments. We'd like to thank Abigail Matthews and Thomas Middleton for their help and support in writing and reviewing the manuscript.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? New York, NY, USA. Association for Computing Machinery.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

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

Yan Chen, Christopher Mahoney, Isabella Grasso, Esma Wali, Abigail Matthews, Thomas Middleton, Mariama Njie, and Jeanna Matthews. 2021. Gender bias and under-representation in natural language processing across human languages. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 24–34.

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

Diwan. 2013. Poetry dataset. *https://www.aldiwan.net/*.

Shamela Library Foundation. 2012. Shamila library dataset. *https://shamela.ws/page/download*.

Michael Grosvald, Sarah Al-Alami, and Ali Idrissi. 2019. Word reading in arabic: Influences of diacritics and ambiguity. In *36th West Coast Conference on Formal Linguistics*, pages 176–181. Cascadilla Proceedings Project.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, et al. 2018. Unified guidelines and resources for arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Youssef Kadri and Jian-Yun Nie. 2006. Effective stemming for arabic information retrieval. In *Proceedings of the International Conference on the Challenge of Arabic for NLP/MT*.

Marzieh Karch. 2021. How to use the ngram viewer tool in google books. In *https://www.lifewire.com/google-books-ngram-viewer-1616701*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Hamdy Mubarak. 2017. Build fast and accurate lemmatization for arabic. *arXiv preprint arXiv:1710.06700*.

Marwa Naili, Anja Habacha Chaibi, and Henda Hajjami Ben Ghezala. 2017. Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science*, 112:340–349. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the 12th language resources and evaluation conference*, pages 7022–7032.

Abu Dhabi Department of Culture and Tourism. 2016. Poetry encyclopedia. *https://poetry.dctabudhabi.ae*.

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

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

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, pages 1003–1011.

Esma Wali, Yan Chen, Christopher Mahoney, Thomas Middleton, Marzieh Babaeianjelodar, Mariama Njie, and Jeanna Neefe Matthews. 2020. Is machine learning speaking my language? a critical look at the nlp-pipeline across 8 human languages. *arXiv preprint arXiv:2007.05872*.

Melvin Wevers. 2019. Using word embeddings to examine gender bias in dutch newspapers, 1950-1990. *arXiv preprint arXiv:1907.08922*.

Yang Xu, Jiasheng Zhang, and David Reitter. 2019. Treat the word as a whole or look inside? subword embeddings model language change and typology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 136–145.

Waleed A. Yousef, Omar M. Ibrahime, Taha M. Madbouly, Moustafa A. Mahmoud, Ali H. El-Kassas, Ali O. Hassan, and Abdallah R. Albohy. 2018. Arabic poem comprehensive dataset. *https://hci-lab.github.io/ArabicPoetry-1-Private/PCD*.