Gender Bias in Natural Language Processing Across Human Languages

Abigail Matthews
University of
Wisconsin-Madison

Isabella Grasso
Christopher Mahoney
Yan Chen
Esma Wali
Thomas Middleton
Jeanna Matthews
Clarkson University
jnm@clarkson.edu

Mariama Njie
Iona College

Abstract

Natural Language Processing (NLP) systems are at the heart of many critical automated decision-making systems making crucial recommendations about our future world. Gender bias in NLP has been well studied in English, but has been less studied in other languages. In this paper, a team including speakers of 9 languages - Chinese, Spanish, English, Arabic, German, French, Farsi, Urdu, and Wolof - reports and analyzes measurements of gender bias in the Wikipedia corpora for these 9 languages. We develop extensions to profession-level and corpus-level gender bias metric calculations originally designed for English and apply them to 8 other languages, including languages that have grammatically gendered nouns including different feminine, masculine, and neuter profession words. We discuss future work that would benefit immensely from a computational linguistics perspective.

1. Introduction

Corpora of human language are regularly fed into machine learning systems as a key way to learn about the world. Natural Language Processing 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 (Yordanov 2018)(Banerjee 2020)(Garbade 2018). Systems are taught to identify spam email, suggest medical articles or diagnoses related to a patient's symptoms, sort resumes based on relevance for a given position, and many other tasks that form key components of critical decision making systems in areas such as criminal justice, credit, housing, allocation of public resources and more.

In a highly influential paper "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings", Bolukbasi et al. (2016) developed a way to measure gender bias using word embedding systems like Word2vec. Specifically, they defined a set of gendered word pairs such as ("he","she") and used the difference between these word pairs to define a gendered vector space. They then evaluated the relationship of profession words like doctor, nurse, or teacher relative to this gendered vector space. They demonstrated that word embedding software trained on a corpus of Google news could associate men with the profession computer programmer and women with the profession homemaker. Systems based on such models, trained even with "representative text" like Google news, could lead to biased hiring practices if used to, for example, parse resumes and suggest matches for a computer programming job. However, as with many results in NLP research, this influential result has not been applied beyond English.

In some earlier work from this team, "Quantifying Gender Bias in Different Corpora", we applied Bolukbasi et al.'s methodology to computing and comparing corpus-level gender bias metrics across different corpora of the English text (Babaeianjelodar 2020). We measured the gender bias in pre-trained models based on a "representative" Wikipedia and Book Corpus in English and compared it to models that had been fine-tuned with various smaller corpora including the General Language Understanding Evaluation (GLUE) benchmarks and two collections of toxic speech, RtGender and IdentityToxic. We found that, as might be expected, the RtGender corpora produced the highest gender bias score. However, we also found that the hate speech corpus, IdentityToxic, had lower gender bias scores than some of more representative corpora found in the GLUE benchmarks. By examining the contents of the IdentityToxic corpus, we found that most of the text in Identity Toxic reflected bias towards race or sexual orientation, rather than gender. These results confirmed the use of a corpus-level gender bias metric as a way of measuring gender bias in an unknown corpus and comparing across corpora, but again was only applied in English.

Here we build on the work of Bolukbasi et al. and our own earlier work to extend these important techniques in gender bias measurement and analysis beyond English. This is challenging because unlike English, many languages like Spanish, Arabic, German, French, and Urdu, have grammatically gendered nouns including feminine, masculine and, neuter or neutral profession words. We translate and modify Bolukbasi et al.'s defining sets and profession sets in English for 8 additional languages and develop exten-

sions to the profession-level and corpus-level gender bias metric calculations for languages with grammatically gendered nouns. We use this methodology to analyze the gender bias in Wikipedia corpora for Chinese (Mandarin Chinese), Spanish, English, Arabic, German, French, Farsi, Urdu, and Wolof. We demonstrate how the modern NLP pipeline not only reflects gender bias, but also leads to substantially over-representing some (especially English voices recorded in the digital text) and under-representing most others (speakers of most of the 7000 human languages and even writers of classic works that have not been digitized).

In Section 2, we describe modifications that we made to the defining set and profession set proposed by Bolukbasi et al. in order to extend the methodology beyond English. In Section 3, we discuss the Wikipedia corpora and the occurrence of words in the modified defining and profession sets for 9 languages in Wikipedia. In Section 4, we extend Bolukbasi's gender bias calculation to languages, like Spanish, Arabic, German, French, and Urdu, with grammatically gendered nouns. We apply this to calculate and compare profession-level and corpus-level gender bias metrics for Wikipedia corpora in the 9 languages. We conclude and discuss future work in Section 5. Throughout this paper, we discuss future work that would benefit immensely from a computational linguistics perspective.

2. Modifying Defining Sets and Profession Sets

Word embedding is a powerful NLP technique that represents words in the form of numeric vectors. It is used for semantic parsing, representing the relationship between words, and capturing the context of a word in a document (Karani 2018). For example, Word2vec is a system used to efficiently create word embeddings by using a two-layer neural network that efficiently processes huge data sets with billions of words, and with millions of words in the vocabulary (Mikolov 2013).

Bolukbasi et al. developed a method for measuring gender bias using word embedding systems like Word2vec. Specifically, they defined a set of highly gendered word pairs such as ("he", "she") and used the difference between these word pairs to define a gendered vector space. They then evaluated the relationship of profession words like doctor, nurse or teacher 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, doctor might be male biased or nurse female biased based on how these words are used in the corpora from which the word embedding model was produced. Thus, this gender bias metric of profession words as calculated from the Word2Vec model can be used as a measure of the gender bias learned from corpora of natural language.

In this section, we describe the modifications we made to the defining set and profession set proposed by Bolukbasi et al. in order to extend the methodology beyond English. Before applying these changes to other languages, we evaluate the impact of the changes on calculations in English. In this section, we also describe the Wikipedia corpora we used across 9 languages and analyze the occurrences of our defining set and profession set words in these corpora. This work is also described, but with a different focus in Wali et al. (2020) and Chen et al. (2021).

2.1. Defining Set

The defining set is a list of gendered word pairs used to define what a gendered relationship looks like. Bolukbasi et al's original defining set contained 10 English word pairs (she-he, daughter-son, her-his, mother-father, woman-man, gal-guy, Mary-John, girl-boy, herself-himself, and female-male) (Bobluski et al. 2016). We began with this set, but made substantial changes in order to compute gender bias effectively across 9 languages.

Specifically, we removed 6 of the 10 pairs, added 3 new pairs and translated the final set into 8 additional languages. For example, we removed the pairs she-he and herself-himself because they are the same word in some languages like Wolof, Farsi, Urdu, and German. Similarly, we removed the pair her-his because in some languages like French and Spanish, the gender of the object does not depend on the person to which it belongs.

We also added 3 new pairs (queen-king, wife-husband, and madam-sir) for which more consistent translations were available across languages. Interestingly, as we will discuss, the pair wife-husband introduces surprising results in many languages. Our final defining set for this study thus contained 7 word pairs and Table 1 shows our translations of this final defining set across the 9 languages included in our study.

2.2 Professions Set

We began with Bolukbasi et al's profession word set in English, but again made substantial changes in order to compute gender bias effectively across 9 languages. Bolukbasi et al. had an original list of 327 profession words (2016), including some words that would not technically be classified as professions like saint or drug addict. We narrowed this list down to 32 words including: nurse, teacher, writer, engineer, scientist, manager, driver, banker, musician, artist, chef, filmmaker, judge, comedian, inventor, worker, soldier, journalist, student, athlete, actor, governor, farmer, person, lawyer, adventurer, aide, ambassador, analyst, astronaut, astronomer, and biologist. We tried to choose a diverse set of professions from creative to scientific, from high-paying to lower-paying, etc. that occured in as many of the 9 languages as we could. As with Bolukbasi et al.'s profession set, one of our profession words, person, is not technically a profession, but we kept it because, unlike many professions, it is especially likely

English	Chinese	Spanish	Arabic	German	French	Farsi	Urdu	Wolof
woman	女人	mujer	النساء	Frau	femme	زن	عورت	Jigéen
man	男人	hombre	رجل	Mann	homme	مرد	آدمی	Góor
daughter	女儿	hija	ابنة	Tochter	fille	دختر	بیٹی	Doom ju jigéen
son	儿子	hijo	ولد	Sohn	fils	پسر	بیٹا	Doom ju góor
mother	母亲	madre	ام	Mutter	mère	مادر	ماں	Yaay
father	父亲	padre	اب	Vater	père	پدر	باپ	Baay
girl	女孩	niña	ابنة	Mädchen	fille	دختر	لڑکی	Janxa
boy	男孩	niño	صبي	Junge	garçon	پسر	لڑکا	Xale bu góor
queen	女王	reina	ملكة	Königin	reine	ملکہ	ملکہ	Jabari buur
king	国王	rey	ملك	König	roi	پادشاہ	بادشاہ	Buur
wife	妻子	esposa	زوجة	Ehefrau	épouse	همسر	بیوی	Jabar
husband	丈夫	esposo	الزوج	Ehemann	mari	شوهر	شوہر	jëkkër
madam	女士	señora	سيدتي	Dame	madame	خانم	محترمہ	Ndawsi
sir	男士	señor	سيدي	Herr	monsieur	آقا	جناب	Góorgui

Table 1: Final defining set translated across languages. Note: Wolof is primarily a spoken language and is often written as it would be pronounced in English, French and Arabic. This table shows it written as it would be pronounced in French.

to have a native word in most human languages.

The primary motivation for reducing the profession set from 327 to 32 was to reduce the work needed to translate and validate all of them in 9 languages. Even with 32 words, there were substantial complexities in translation. As we mentioned, languages with grammatically gendered nouns can have feminine, masculine, and neuter words for the same profession. For instance, in Spanish, the profession "writer" will be translated as "escritora" for women and "escritor" for men, but the word for journalist, "periodista", is used for both women and men.

Profession words are often borrowed from other languages. In this study, we found that Urdu and Wolof speakers often use the English word for a profession when speaking in Urdu or Wolof. In some cases, there is a word for that profession in the language as well and in some cases, there is not. For example, in Urdu, it is more common to use the English word "manager" when speaking even though there are Urdu words for the profession manager. In written Urdu, manager could be written directly in English characters (manager) or written phonetically as the representation of the word manager using Urdu/Arabic characters (منىجر) or written as an Urdu word for manager (منتظم /منتظمه).

A similar pattern occurs in Wolof and also in Wolof there are some additional complicating factors. Wolof is primarily a spoken language that when written is transcribed phonetically. This may be done using English, French, or Arabic character sets and pronunciation rules. Thus, for the same pronunciation, spelling can vary substantially and this complicates NLP processing such as with Word2Vec significantly. After making these substantial changes to the defining sets and profession sets, the first thing we did was analyze their impact on gender bias measurements in English. Using both Bolukbasi et al's original defining and professions sets and our modified sets, we computed the gender bias scores on the English Wikipedia corpus. With our 7 defining set pairs and 32 profession words, we conducted a T-test and even with these substantial changes the T-test results were insignificant, inferring that the resulting gender bias scores in both instances have no statistically significant difference for the English Wikipedia corpus. This result was an encouraging validation that our method was measuring the same effects as in Bolukbasi et al. even with the modified and reduced defining set and profession set.

While our goal in this study was to identify a defining set and profession set that could more easily be used across many languages and for which the T-test results indicated no statistically significant difference in results over the English Wikipedia corpus, it would be interesting to repeat this analysis with additional variations in the defining set and profession set. For example, we considered adding additional pairs like sister-brother or grandmother-grandfather. In some languages like Chinese, Arabic, and Wolof, there are different words for younger and older sister or brother. We also considered and discarded many other profession words such as bartender, policeman, celebrity, and electrician. For example, we discarded bartender because it is not a legal profession in some countries. We would welcome collaborators from the computational linguistics community to help identify promising defining set pairs and profession set words which which to experiment.

3. Wikipedia Corpora Across Languages

Bolukbasi et al. applied their gender bias calculations to a Word2Vec model trained with a corpus of Google news in English. In Babaeianjelodar et al., we used the same defining and profession sets as Bolukbasi et al. to compute gender bias metrics for a BERT model trained with Wikipedia and a BookCorpus also in English. In this paper, we train Word2Vec models using our modified defining and profession sets and the Wikipedia corpora for 9 languages. Specifically, we use the Chinese, Spanish, Arabic, German, French, Farsi, Urdu, and Wolof corpora downloaded from Wikipedia on 2020-06-20. We would like to examine more languages in this way and would welcome suggestions of languages to prioritize in future work.

3.1. Differences in Wikipedia across Languages

While there are Wikipedia corpora for all 9 of our languages, they differ substantially in size and quality. Wikipedia is a very commonly used dataset for testing NLP tools and even for building pre-trained models. However, for many reasons, a checkmark simply saying that a Wikipedia corpus exists for a language hides many caveats to full representation and participation. In addition to variation in size and quality across languages, not all speakers of a language have equal access to contributing to Wikipedia. For example, in the case of Chinese, Chinese speakers in mainland China have little access to Wikipedia because it is banned by the Chinese government (Siegel 2019). Thus, Chinese articles in Wikipedia are more likely to have been contributed by the 40 million Chinese speakers in Taiwan, Hong Kong, Singapore, and elsewhere (Su 2019). In other cases, the percentage of speakers with access to Wikipedia may vary for other reasons such as access to computing devices and Internet access.

Using Wikipedia as the basis of pre-trained models and testing of NLP tools also means that the voices of those producing digital text are prioritized. Even authors of classic works of literature that fundamentally shaped cultures are under-represented in favor of writers typing Wikipedia articles on their computer or even translating text written in other languages with automated tools.

3.2. Word Count Results

One critical aspect of our process was to examine the number of times each word in our defining set (7 pairs) and 32 profession words occurs in the Wikipedia corpus for each language. This proved an invaluable step in refining our defining and profession sets, understanding the nature of the Wikipedia corpora themselves, catching additional instances where NLP tools were not designed to handle the complexities of some languages, and even catching simple errors in our own translations and process. For example, when our original word count results for German showed a count of zero for all words, we discovered that even though all nouns in German are capitalized, in the Word2vec processed Wikipedia corpus for German, all words were in lowercase. This was an easy problem to fix, but illustrates the kind of "death by a thousand cuts" list of surprising errors that can occur for many languages throughout the NLP pipeline.

One important limitation to note is that for many languages, if a word is expressed with a multi-word phrase (e.g. astronomer(عالم الفلك) in Arabic), the word count reported by Word2Vec for this phrase will be zero. For each language, there is a tokenizer that identifies the words or phrases to be tracked. In many cases, the tokenizer identifies words as being separated by a space. The Chinese tokenizer however attempts to recognize when multiple characters that are separated with spaces should be tracked as a multi-character word or concept. This involves looking up a string of characters in a dictionary. Once again this demonstrates the types of surprising errors that can occur for many languages throughout the NLP pipeline. It is also possible to add the word vectors for component words together as a measure of the multi-word pair, but this is not always ideal. In this study, we did not attempt this, but it would be interesting future work.

Another important factor is that the Wikipedia corpora for some languages are quite small. In Wolof, for example, only two of our profession words occurred ("nit", the word for person, occurred 1401 times and waykat, the word for musician, occurred 5 times). This is partly because of multi-word pairs and partly because of variants in spelling. However, we think it is especially due to the small size of the Wolof corpus because the percentage of profession words amongst the total words for Wolof is similar to that of other languages. Across the 9 languages, the percentage of profession words varied from 0.014% and 0.037%. Wolof actually had one of the higher percentages at 0.026%. However, its overall Wikipedia corpus is tiny (1422 articles or less than 1% of the number of articles even in Urdu, the next smallest corpora) and that simply isn't a lot of text with which to work. Even so, Wolof is still better represented in Wikipedia than the vast majority of the over 7000 human languages spoken today! This is another clear illustration of how the gap in support for so many languages leads directly to the under-representation of many voices in NLP-guided decision-making.

We do not have room to include the word counts for the defining sets and profession sets for all 9 languages here, but an expanded technical report with this data is available at http://tinyurl.com/clarksonnlpbias.

4. Extending Profession and Corpora Level Gender Bias Metrics

We have already described how we established a modified defining set and profession set for use across 9 languages and then evaluated the use of these sets of words in Wikipedia. We also described how we used the Wikipedia corpora of these 9 languages to train

Word2Vec models for each language. In this section, we describe how we extend Bolukbasi et al.'s method for computing the gender bias of each word.

We begin with Bolukbasi et al.'s method for computing a gender bias metric for each word. Specifically, each word is expressed as a vector by Word2Vec and we calculate the center of the vectors for each definitional pair. For example, to calculate the center of the definitional pair she/he, we average the vector for "she" with the vector for "he". Then, we 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. "she" - center). We 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. 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 the highly gendered, they expect this dimension to be relate primarily to gender and therefore call it the gender direction or the g direction. (Note: The effectiveness of this compression can vary and in some cases, the first eigenvalue may not actually be much larger than the second. We see cases of this in our study as we will discuss.) Finally, we use Bolukbasi et al.'s formula for direct gender bias:

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

where N represents the list of profession words, g represents the gender direction calculated, w represents each profession word, and c is a parameter to measure the strictness of the bias. In this paper, we used $c = 1$; c values and their effects are explained in more detail in Bolukbasi et al. We examine this gender bias score both for the individual words as well as an average gender bias across profession words as a measure of gender bias in a corpus.

To apply this methodology across languages, some important modifications and extensions were required, especially to handle languages, like Spanish, Arabic, German, French, and Urdu, that have grammatically gendered nouns. In this section, we describe our modifications and apply them to computing and comparing both profession-level and corpus-level gender bias metrics across the Wikipedia corpora for 9 languages.

4.1. Evaluating the Gender Bias of Defining Sets Across Languages

To begin, in Figure 1, we present the gender bias scores, calculated as described above according to Bolukbasi et al.'s methodology, for each of our 14 defining set words (7 pairs) across 9 languages. Female gender bias is represented as a positive number (red bar) and male gender bias is represented as a negative number (blue bar). Not all defining set words occur in the Wikipedia corpus for Wolof. Some because they are translated in multi-word phrases and some simply because of the same size of the corpora.

The defining set pairs were specifically chosen because we expect them to be highly gendered. In most cases, the defining set words indicated male or female bias as expected, but there were some exceptions. One common exception was the word husband. Husband, somewhat surprisingly, has a female bias in a number of languages. We hypothesize that "husband" may more often be used in relationship to women (e.g. "her husband"). One might guess that the same pattern would happen for wife then but it does not appear to be the case. We hypothesize that it may be less likely for a man to be defined as a husband outside of a female context, where women may often be defined by their role as a wife even when not in the context of the husband. This is an interesting effect we saw across many languages.

In Figure 2, we aggregate the gender bias for all the male words (sir, husband, king, etc.) and all the female words (madam, wife, queen, etc.) This presentation emphasizes several key aspects of the results. For example, we can see that for Spanish, English, Arabic, German, French, Farsi, and Urdu, that the female words are female leaning and that most male words and male leaning as one might expect,with the exception of husband in all of these languages and also man in Farsi. We can also see that female words have more female bias than male words have male bias.

We can also see problems with both Chinese and Wolof. We have discussed some of the problems in Wolof with the size of the corpora and the difficulty of matching phonetically transcribed words. However, for Chinese, we have a sizable corpora and many occurrences of the defining set words. After much investigation, we isolated an issue related to the Principal Component Analysis (PCA) in Chinese. As we described at the beginning of this section, Bolukbasi et al.'s methodology calls for using the largest eigenvalue and in their experience the first eigenvalue was much larger than the second and they analyzed their results using only this dominant dimension. However, we found that this was not always the case. In particular for the Chinese Wikipedia corpus, the largest eigenvalue of the PCA matrix is not much larger than the second.

In Figure 3, we report the difference in PCA scores between the dominant component and the next most dominant component across 9 languages in our study. We also add a bar for the value Bolukbasi et al. reported for the Google News Corpora in English that they analyzed. Chinese has the lowest. Wolof has the highest with 1.0, but only because there were not enough defining pairs to meaningfully perform dimension reduction into 2 dimensions. We repeated our analysis without the wife-husband pair and found that the difference in PCA scores improved for all languages except for

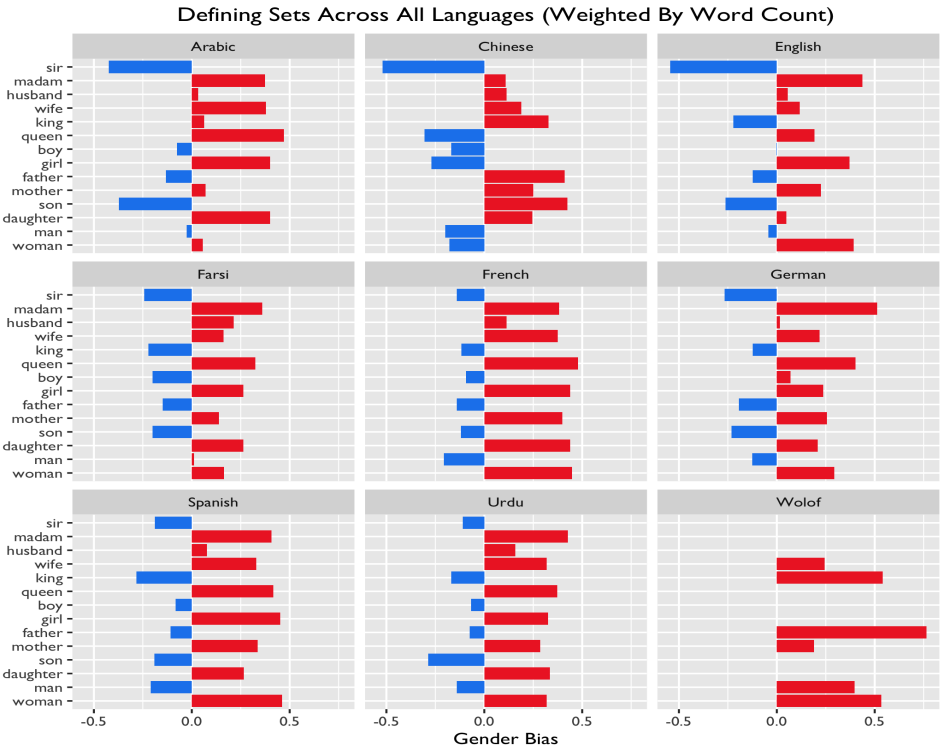

Figure 1: Defining Sets Across Languages The x-axis represents per-word gender bias scores as proposed by Bolukbasi et al. Female gender bias is represented as a positive number (red bar) and male gender bias is represented as a negative number (blue bar). Not all defining set words occur in the small Wikipedia corpus for Wolof. We note that boy in English has a gender bias of -0.002 which is such a small blue line that it is difficult to see.

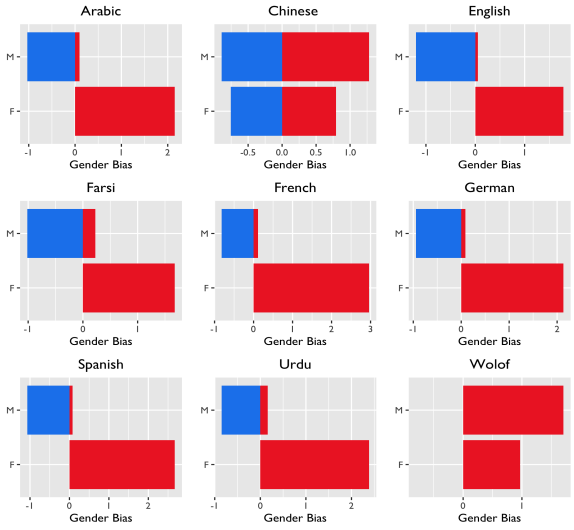

Figure 2: Defining Set Summary For each language, we aggregate the gender bias scores of male defining set words (the M bar) and female defining set words (the F bar).

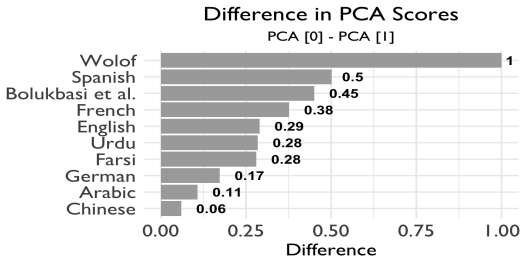

Figure 3: Difference in importance between first and second principal components by language. A larger difference increases the confidence we've isolated the gender direction.

Wolof. Wolof remains 1.0 because we didn't find any defining pairs. We have been experimenting with modifications to the defining set in Chinese including isolating the contribution of each individual defining set pair and adding many pairs that while meaningful specifically in Chinese would not have worked well across all languages (e.g. different pairs for paternal and maternal grandmother and grandfather). However, we have yet to find a defining set that works well and would welcome collaborations from linguists with expertise in Chinese.

The word boy in German, Junge, also highlights some important issues. Junge can also be used as an adjective such as in "junge Leute" (young people) and it is also a common surname. Since these different uses of the word are not disambiguated, it is likely that the token "junge" encompasses more meaning than simply boy. We also saw this with the defining set word "fille" in French which means both girl and daughter. This

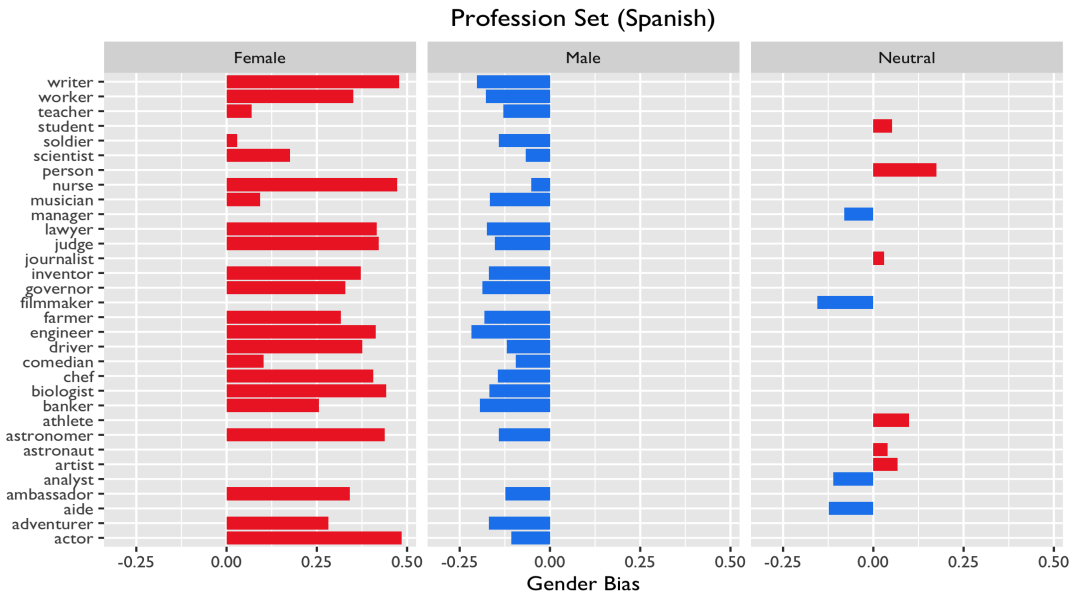

Figure 4: Per Profession Gender Bias for Spanish. Broken down into female only variants, male only variants and neutral variants.

Figure 5: (LEFT)Per-Profession Gender Bias Metrics for Languages Without Grammatically Gendered Nouns (RIGHT)Per-Profession Gender Bias Metrics for All Languages Weighting by Word Count

problem of disambiguation occurs in many languages and multiple meanings for words should be considered when selecting terms. We would appreciate the insight of linguists in how to handle disambiguation of terms more generally.

4.2. Evaluating the Gender Bias of Profession Sets Across Languages

Having analyzed the defining set results where there is a clearly expected gender for each word, we move on to the question of computing the gender bias scores for each of our 32 profession words. Bolukbasi et al.'s methodology can be applied directly in English and

also in other languages which, like English, do not have many grammatically gendered nouns. Of the 9 languages, we studied, Chinese, Farsi and Wolof are also in this category.

The situation is more complicated in languages with grammatically gendered nouns. Five of the languages we are studying fall into this category: Spanish, Arabic, German, French, and Urdu. In these languages, many professions have both a feminine and masculine form. In some cases, there is also a neutral form and in some cases there is only a neutral form. In Section 2.2, we discussed how Urdu also often uses English words directly. Thus there are neutral Urdu words and neutral English words used in Urdu. To form a per-profession bias metric, we averaged the bias metrics of these various forms in several different ways. First, we averaged them, weighting each different form of a profession equally. However, we found that this overestimated the female bias in many cases. For example, in German the male form of scientist, Wissenschaftler, has a slight male gender bias (-0.06) and the female form, Wissenschaftlerin, has a strong female gender bias (0.32). When averaged together evenly, we would get an overall female gender bias of 0.13. However, the male form occurs 32,467 times in the German Wikipedia corpus while the female form occurs only 1354 times. To take this difference into account, we computed a weighted average resulting in an overall male gender bias of -0.04. With this weighted average, we could observe intuitive patterns across languages with grammatically gendered nouns and languages without. This increases our confidence in the usefulness of these profession level metrics and in particular the weighted average.

In Figure 4, we show an example breakdown of the gender bias scores for the Spanish profession words. We show female only variants, male only variants and neutral only variants. At http://tinyurl/clarksonnlpbias, we provide a technical report with a breakdown like this for all 5 of the gendered languages in our study. Notice that the gender bias for all female words is indeed female and that the gender bias for all male words is indeed male. Neutral words show a mix of male and female bias. This an intuitive and encouraging result that further supports the use of per-word gender bias calculations across languages. This is often true in other languages, but not exclusively so.

In Figure 5, we compare these profession-level gender bias scores across languages. On the left, we show results for the languages without grammatically gendered nouns. On the right, we show results across all languages using the weighted average (weighted by word count).

In Figure 6, we use Pearson's correlation for cluster analysis to examine 7 languages, omitting Chinese and Wolof because of problems with PCA and corpora size. This exploratory analysis provokes a number of questions for future work including: How do linguistics inform bias outputs (e.g. If English is a mixture of Germanic and Latin languages, is that why it's clustered with those languages even though it's not gendered?) and How does a language being inherently gendered affect the resulting bias of a NLP model in that language?

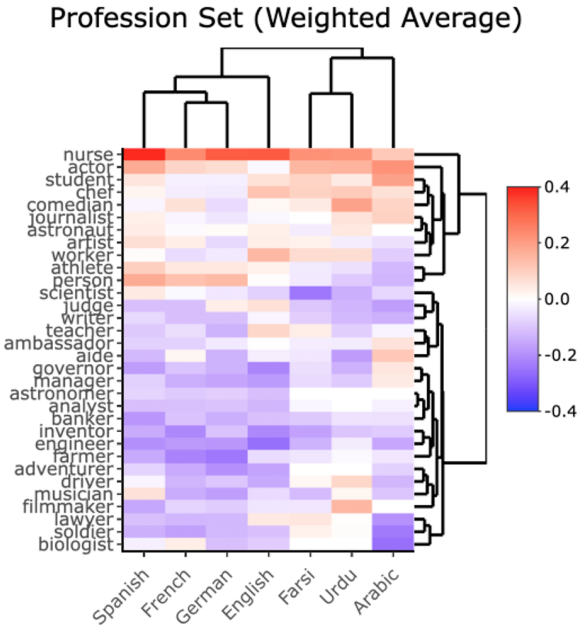

Figure 6: Pearson's correlation for cluster analysis across 7 languages

6. Conclusion and Future Work

We have extended an influential method for computing gender bias from Bolukbasi et al., a technique that had only been applied in English. We made key modifications that allowed us to extend the methodology to 8 additional languages, including languages with grammatically gendered nouns. With this, we quantified how gender bias varies across the Wikipedia corpora of 9 languages and discuss future work that could benefit immensely from a computational linguistics perspective.

Specifically, we would like to explore additional languages as well as understand better how variations in defining sets and profession sets can highlight differences among languages. We would like to compare gender bias across different corpora of culturally important texts written by native speakers. Even within one language, we would like to examine collections with different emphasis such as gender of author, different time periods, different genres of text, different country of origin,etc. Our work is an important first step toward quantifying and comparing gender bias across languages - what we can measure, we can more easily begin to track and improve, but it is only a start. The majority of human languages need more useful tools and resources to overcome the barriers such that we can build NLP tools with less gender bias.

Acknowledgements

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 Golshan Madraki, Marzieh Babaeianjelodar, and Ewan Middelton for help with language translations as well as our wider team including William Smialek, Graham Northup, Cameron Weinfurt, Joshua Gordon, and Hunter Bashaw for their support.

References

Babaeianjelodar, M.; Lorenz, S.; Gordon, J.; Matthews, J.; and Freitag, E. 2020. Quantifying gender bias in different corpora. In Companion Proceedings of the Web Conference 2020, WWW '20, page 752–759, New York, NY, USA, 2020. Association for Computing Machinery. DOI: https://doi.org/10.1145/3366424.3383559.

Banerjee, D. 2020. Natural Language Processing (NLP) Simplified: A Step-by-step Guide. Datascience foundation. Retrieved from https://datascience.foundation/sciencewhitepaper/natural-language-processing-nlp-simplified-a-step-by-step-guide

BERT. 2020. BERT Pretrained models. Github. Retrieved from https://github.com/google-research/bert#bert

Bolukbasi, T.; Chang, K.; Zou, J. Y.; Saligrama, V.; and Kalai, A.T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems, pp. 4349-4357.

Buolamwini, J; and Gebru T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91, 2018.

Bussieck, J. 2017. Demystifying Word2Vec. Retrieved from https://www.deeplearningweekly.com/blog/demystifying-word2vec/

Chen, Y., Mahoney, C., Grasso, I., Wali, E., Matthews, A., Middleton, T., Njie, M., and Matthews, J. Gender Bias and Under-Representation in Natural Language Processing Across Human Languages Proceedings of the 2021 AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES), May 19-21 2021.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805

Garbade, M. J. 2018. A Simple Introduction To Natural Language Processing. Retrieved from https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32

Holley, R. 2009. How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine, March/April 2009, Volume 15 Number 3/4 ISSN 1082-9873. Retrieved from http://www.dlib.org/dlib/march09/holley/03holley.html

Karani, D. 2018. Introduction to Word Embedding and Word2Vec. Retrieved from https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

Karch, M. 2020. How to Use the Ngram Viewer Tool in Google Books. Retrieved from https://www.lifewire.com/google-books-ngram-viewer-1616701

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781. Retrieved from https://arxiv.org/abs/1301.3781

Mithe, R.; Indalkar, S.; and Divekar, N. 2013. Optical character recognition. International journal of recent technology and engineering. ISSN: 2277-3878, Volume-2, Issue-1, March 2013.

NLTK. 2005. Natural Language Toolkit. Retrieved from http://www.nltk.org/

Nosek, B. A.; Banaji, M. R.; and Greenwald, A. G. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. Group Dy- namics: Theory, Research, and Practice, 6(1):101, 2002. Ohio University. Wolof Language. Retrieved from https://www.ohio.edu/cis/african/languages/wolof

Rong, X. 2016. word2vec Parameter Learning Explained. arXiv:1411.2738. Retrieved from https://arxiv.org/abs/1411.2738

Siegel, R. 2019. Search result not found: China bans Wikipedia in all languages. Retrieved from https://www.washingtonpost.com/business/2019/05/15/china-bans-wikipedia-all-languages/

Su, Q, G. 2019. Which Parts of the World Speaks Mandarin Chinese?. Retrieved from https://www.thoughtco.com/where-is-mandarin-spoken-2278443

Wali, E., Chen, Y., Mahoney, C., Middleton, T., Babaeianjelodar, M., Njie, M., and Matthews, J. Is Machine Learning Speaking my Language? A Critical Look at the NLP-Pipeline Across 8 Human Languages Participatory ML Workshop,Thirty-seventh International Conference on Machine Learning (ICML 2020), July 17 2020.

WikipediaA. 2020. Wikipedia: German language. Retrieved from https://en.wikipedia.org/wiki/German_language

WikipediaB. 2020. Wikipedia: List of languages by total number of speakers. Retrieved from https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

WikipediaC. 2020. Wikipedia: List of Wikipedias. Retrieved from https://en.wikipedia.org/wiki/List_of_Wikipedias

WikipediaD. 2020. Wikipedia: Wolof Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Wolof_Wikipedia

Williams, A., Nangia, N., Bowma, S. 2020. MultiNLI, The Multi-Genre NLI Corpus. Retrieved from https://cims.nyu.edu/~sbowman/multinli

Yordanov, V. 2018. Introduction To Natural Language Processing For Text. Medium. Retrieved from https://towardsdatascience.com/introduction-to-natural-language-processing-for-text-df845750fb63

Yamada, I.; Asai, A.; Sakuma, J.; Shindo, H.; Takeda, H.; Takefuji, Y.; and Matsumoto Y. 2018. Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. arXiv:1812.06280. Retrieved from https://arxiv.org/abs/1812.06280