Gender Bias and Under-Representation in Natural Language Processing Across Human Languages

Yan Chen[1], Christopher Mahoney[1], Isabella Grasso[1], Esma Wali[1], Abigail Matthews[2], Thomas Middleton[1], Mariama Njie[3], Jeanna Matthews[1]
[1]Clarkson University, [2]University of Wisconsin-Madison, [3] Iona College
jnm@clarkson.edu

ABSTRACT

Natural Language Processing (NLP) systems are at the heart of many critical automated decision-making systems making crucial recommendations about our future world. However, these systems reflect a wide range of biases, from gender bias to a bias in which voices they represent. 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. In the process, we also document how our work exposes crucial gaps in the NLP-pipeline for many languages. Despite substantial investments in multilingual support, the modern NLP-pipeline still systematically and dramatically under-represents the majority of human voices in the NLP-guided decisions that are shaping our collective future. 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 like Spanish, Arabic, German, French and Urdu that have grammatically gendered nouns including different feminine, masculine and neuter profession words. We compare these gender bias measurements across the Wikipedia corpora in different languages as well as across some corpora of more traditional literature.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**.

KEYWORDS

bias, gender bias, natural language processing

ACM Reference Format:
Yan Chen[1], Christopher Mahoney[1], Isabella Grasso[1], Esma Wali[1], Abigail Matthews[2],, Thomas Middleton[1], Mariama Njie[3], Jeanna Matthews[1], . 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 (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462530

AIES '21, May 19–21, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462530

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 [20] [2] [7]. 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. Much like facial recognition systems are often trained to represent white men more than black women [5], machine learning systems are often trained to represent human expression in languages such as English and Chinese more than in languages such as Urdu or Wolof.

The degree to which some languages are under-represented in commonly used text-based corpora is well-recognized, but the ways in which this effect is magnified throughout the NLP-tool chain is less discussed. Despite huge and admirable investments in multilingual support in projects like Wikipedia (Wikipedia 2020C), BERT [6] [3], Word2Vec [9], Wikipedia2Vec [19], Natural Language Toolkit [10], MultiNLI [18], many NLP tools are only developed for and tested on one or at most a handful of human languages and important advancements in NLP research are rarely extended to or evaluated for multiple languages. For some languages, the NLP-pipeline is streamlined: large publicly available corpora and even pre-trained models exist, tools run without errors and there is a rich set of research results applied to that language [13]. However, for the vast majority of human languages, there is hurdle after hurdle. Even when a tool technically does support a given language, that support often comes with substantial caveats such as higher error rates and surprising problems. Also lack of representation at early stages of the pipeline (e.g. small corpora) adds to the lack of representation in later stages of the pipeline (e.g lack of tool support or research results not applied to that language).

In a highly influential paper "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings", Bolukbasi et al. developed a way to measure gender bias using word embedding systems like Word2vec [4]. 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 our own 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 [1]. 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.

The results in Babaeianjelodar et al. also demonstrated the difficulty in predicting the degree of gender bias in unknown corpora without actually measuring it. It is an increasingly common practice for application developers to start with pre-trained models and then add in a small amount of fine-tuning customized to their application. However, when the amount of gender bias learned from these "off-the-shelf" ingredients occurs unexpectedly, it can introduce unexpected learned gender bias in deployed applications in unpredictable ways, leading to significant problems when used to make critical decisions impacting the lives of individuals.

This common practice of application developers starting with a pre-trained model and then adding in a small amount of fine-tuning to customize their application has another important consequence. Pre-training from scratch using the large corpora necessary for meaningful NLP-results is expensive (i.e. days on a dozen CPUs). When a team can download a pre-trained model, they avoid this substantial overhead. Fine-tuning is much less expensive (i.e. hours on a single CPU). This makes NLP- based results accessible to a wider range of people, but only if such a pre-trained model is available for their language. When these easy-to-use pre-trained models exist for only a few languages, it steers more teams to languages with these pre-trained models and away from other languages and thus can further exacerbate the disparity in representation and participation among human languages.

In this paper, 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 modify and translate Bolukbasi et al.'s defining set and profession set in English for 8 additional languages and develop extensions 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 Mandarin Chinese, Spanish, English, Arabic, German, French, Farsi, Urdu, and Wolof.

In the process, we document many ways in which other sources of bias, beyond gender bias, are also being introduced by the modern NLP pipeline. Starting with the input data sets, Wikipedia is often used to train or test NLP-systems and there are substantial differences in the size and quality of the Wikipedia corpora across languages, even when adjusted for the number of speakers of each language. 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). As speakers of 9 languages who have recently examined the modern NLP toolchain, we highlight the difficulties that speakers of many languages face in having their thoughts and expressions included in the NLP-derived conclusions that are being used to direct the future for all of us.

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. In Section 5, we discuss the need to assess corpora beyond Wikipedia. We conclude in Section 6.

2 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 [8]. 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 [9].

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.

Word Pair	Reasons
she-he	They are the same in some languages like Wolof, Farsi, Urdu, and German relying on context for disambiguation. For example, in German, "sie" without additional context would be very difficult to determine whether it correlates to she, them, they, or you.
her-his	In some languages like French and Spanish, the gender of the possessive word refers to the object rather than to the person to whom the object belongs. For example, in German without context, "ihrer" could mean her, your, theirs or yours.
gal-guy	They are the same in some languages like Wolof. There aren't translations for these words in some languages like Urdu and Arabic.
Mary-John	Simply translating Mary and John to other languages is problematic, but so is trying to identify some alternate "typical" male-female names.
herself-himself	They are the same in some languages like Wolof, Farsi, Urdu, and German relying on context for disambiguation.
female-male	They can be both nouns or adjectives in many languages which introduces ambiguity.

Table 1: Removed Word Pairs

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 for 9 languages and analyze the occurrences of our defining set and profession set words in these corpora.

2.1 Defining Set

We call the list of gendered word pairs used to define what a gendered relationship looks like a defining set. 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) [4]. 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. Table 1 summarizes the reasons for the 6 removals. We also added 3 new pairs (queen-king, wife-husband, and madam-sir) for which more consistent translations were available across languages. Our final defining set thus contains 7 word pairs. Table 2 shows our translations of this final defining set across the 9 languages included in our study.

2.2 Professions Set

In English, most profession words do not fundamentally have a gender. Bolukbasi et al. evaluated the gender bias of a list of 327 profession words [4], including some words that would not technically be classified as professions like saint or drug addict. We call this the profession set. We began with Bolukbasi et al's profession set in English, but again made substantial changes in order to compute gender bias effectively across 9 languages. Specifically, 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 occurred 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 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 (منتظمه / منتظم).

A similar pattern occurs in Wolof, but 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

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 2: 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.

encouraging validation that our method was measuring the same effects in English as in Bolukbasi et al. had, even with the modified and reduced defining and profession sets.

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. In Chinese, there are different words for paternal and maternal grandmother and grandfather. 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. Additional experiments with defining sets in each individual language are ongoing future work.

Language	Number of Articles	Number of Speakers (thousand)	Articles/1000 Speakers
Chinese	1,149,477	921,500	1.25
Spanish	1,629,888	463,000	3.52
English	6,167,101	369,700	16.68
Arabic	1,067,664	310,000	3.44
German	2,485,274	95,000	26.16
French	2,253,331	77,300	29.15
Farsi	747,551	70,000	10.68
Urdu	157,475	69,000	2.28
Wolof	1,422	5,500	0.26

Table 3: Comparing the number of speakers of a language to the size of the Wikipedia Corpora for that language. For the number of articles and estimates of the number of speakers of Chinese, Spanish, English, Arabic, German, French, Farsi, Urdu, and Wolof. All of the numbers of speakers only accounted for native speakers. These statistics on the number of articles and the number of speakers are taken from Wikipedia itself [14] [15] [16] [17].

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.

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. Table 3 illustrates that even when accounting for differences in the number of speakers of each language worldwide, some languages have substantially more representation than others. It is interesting to note that the number of speakers of each language does not track evenly with the number of articles. French has the highest ratio of articles/1000 Speakers at 29.15 and Wolof the lowest at 0.26. English has the largest number of articles, even including languages not in our list, but its ratio of articles to speakers is lower than some other languages. In addition to the difference in the number of articles and articles per

1000 speakers, Wikipedia corpora for different languages also vary widely along many other dimensions including the total size of corpora in MB, total pages, percentage of articles that are simply stub articles with no content, number of edits, number of admins working in that language, total number of users and total number of active users.

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 [11]. 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 [12]. 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, as we described earlier, 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, the percentage of profession words amongst the total words for Wolof is still similar to that of other languages suggesting that it is the small size of the Wolof corpus that is the primary problem. In fact, the percentage of profession words varied from 0.014% and 0.037% across the 9 languages and Wolof had one of the higher percentages at 0.026%. On the other hand, Wolof's 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 much 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 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 such as she/he. 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 be highly gendered, they expected this dimension to be related primarily to gender and therefore called 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.)

We use Bolukbasi et al.'s formula for direct gender bias:

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

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 our knowledge, this is the first paper to apply this methodology across languages and 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 Comparing the Gender Bias of Defining Sets

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). As we have discussed, not all defining set words occur in the Wikipedia corpus for Wolof.

The defining set pairs were specifically chosen because we expect them to be highly gendered. So not surprisingly, in most cases, the defining set words indicated male or female bias as expected, but there were some exceptions. More surprisingly, a common exception was the word husband. Husband has a female bias in every language except Wolof where it did not occur in the corpora. 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 and it would be interesting to explore it further in future work. Husband is part of the set of 3 pairs (queen-king, wife-husband, and madam-sir) that were added in this study. In our ongoing work, we are repeating this analysis without the pair husband-wife in the defining set.

In Chinese, in particular, we saw more surprising results with words like father and son taking on a female bias. 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 2, 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.

Thus, for the Chinese Wikipedia corpus, even though the defining set was chosen to be highly gendered, when PCA is used to reduce the number of dimensions, there is not a clearly dominant gender direction. We believe this is the key reason that the gender bias of the defining set words is not as intuitive for Chinese as it is for other languages. If the primary PCA dimension does not capture gender it may suggest the need to add more defining set word pairs and it would be interesting to repeat this analysis with an expanded defining set in Chinese.

The word boy in German, Junge, is also an exception and 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 meanings than simply boy. We saw this with the defining set word "fille" in French which means both girl and daughter. Also, we have mentioned the issue of Word2Vec changing all words to lowercase and this also contributes to combining words in German that should be considered different parts of speech. Since all nouns in German are capitalized, maintaining capitalization would have provided some level of term separation. We suspect that this may contribute to Junge having a feminine gender bias. This problem of disambiguation is not unique to German and multiple meanings for words should be considered when selecting terms. For example, in English we included doctor as a profession, however had concerns of ambiguity with the verb to doctor. Such disambiguation of terms warrants further investigation in all languages.

4.2 Comparing the Gender Bias of Professions and Profession Sets

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. Figure 3 shows the gender bias scores for the 32 profession words for the English Wikipedia corpus. Not surprisingly, we see many of the same patterns as documented by Bolukbasi et al. In this figure, nurse is the profession with the largest absolute value of bias having a female gender bias of 0.32. Engineer is the profession with the largest male gender bias at -0.26.

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

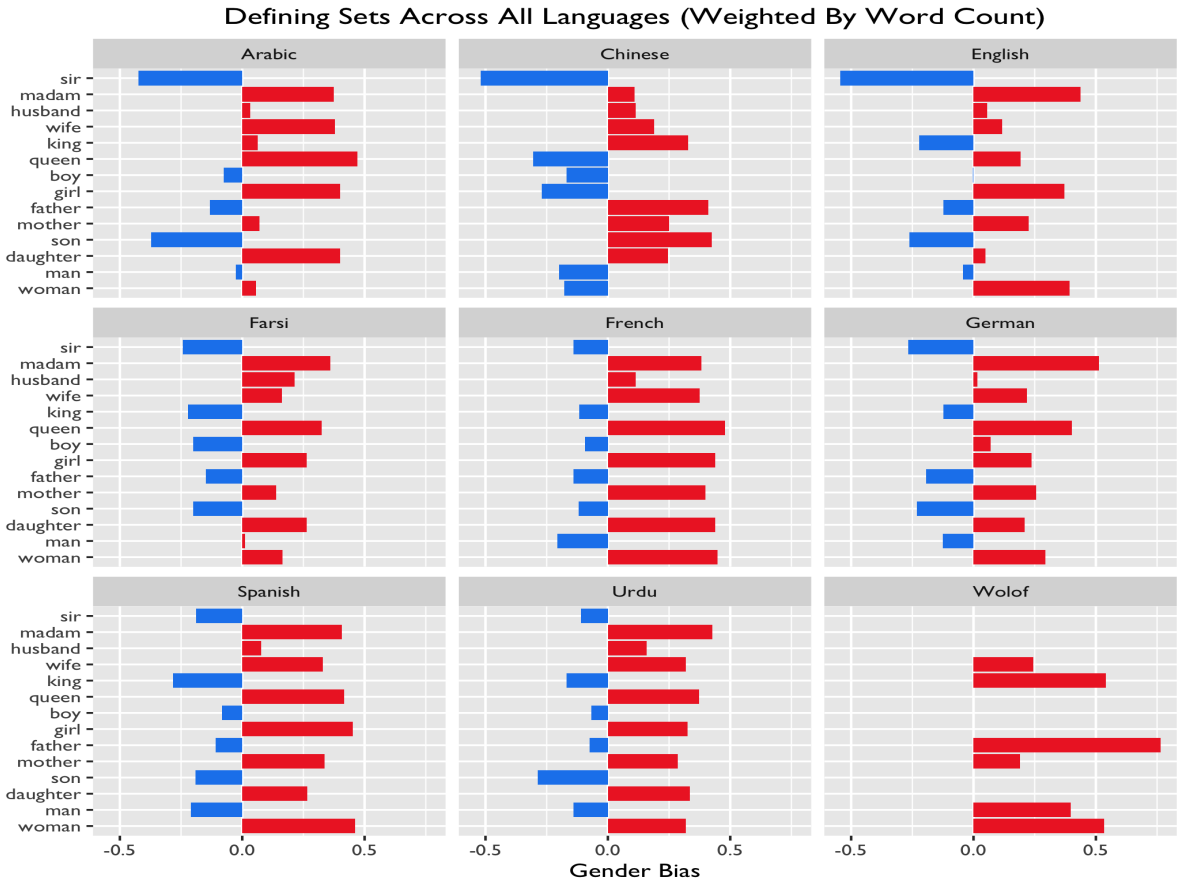

Figure 1: Defining Sets Across Languages The x-axis represents per-word gender bias scores as proposed by Bolukbasi et al. Female gender bias is represented as a positive number (red bar) and male gender bias is represented as a negative number (blue bar). Not all defining set words occur in the small 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.

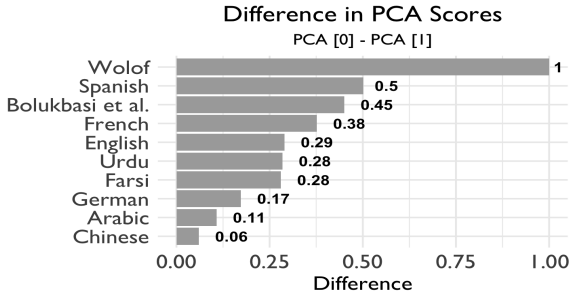

Figure 2: Difference in importance between first and second principal components by language. A larger difference increases the confidence we've isolated the gender direction.

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 also 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 the breakdown of the gender bias scores for the Spanish profession words. We show female only variants, male only variants and neutral only variants. Notice in Figure 4, that the gender bias for all female words is indeed female and that the gender bias for all male words is indeed male. This an intuitive

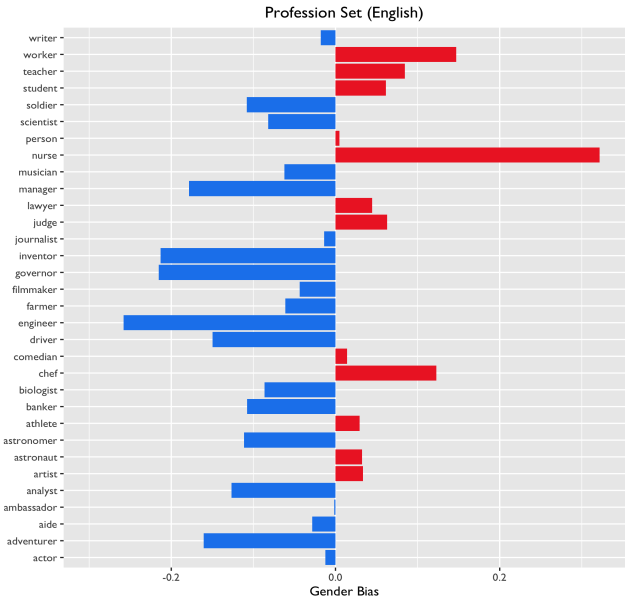

Figure 3: Per Profession Gender Bias for English Wikipedia.

and encouraging result that further supports the use of per-word gender bias calculations across languages. Neutral words show a mix of male and female bias. At http://tinyurl/clarksonnlpbias, we provide a technical report with a breakdown like this for all 5 of the gendered languages in our study.

In Figures 5 and 6, we compare these profession-level gender bias scores across languages. In Figure 5, we show results for the languages without grammatically gendered nouns. In Figure 6, we show results across all languages using the weighted average (weighted by word count). We have removed the y-axis labels in Figure 6 for readability, but the order is the same as in the previous figures and here our emphasis is on observing the patterns across languages rather than on a drill down into specific words. Earlier in the paper, we note some problems with the Chinese results (lack of a non-dominant PCA dimension) and Wolof (a very small corpus). It is interesting to note how similar English and French are. Notice also the similarities in patterns between Spanish, English, Arabic, German, French, Farsi and Urdu. When using an evenly weighted average, instead, the languages with grammatically gendered nouns were similar to each other, but not to English and Farsi. Based on these results, our recommendation is to use the weighted average as we have done in Figure 6.

4.3 Comparing the Gender Bias of Corpora

Having analyzed the gender bias of our defining set and profession words, in this section, we discuss how to combine them into an overall gender bias score for each corpus. Bolukbasi et al. did not use their gender bias metric to compare different corpora. In Babaeianjelodar et al., we used an evenly weighted average across 327 profession words in English. We observed that the corpus-level gender bias score for a hate speech corpora like RtGender was substantially higher than more "representative" corpora like the contents of the GLUE benchmarks. We were also able to use their calculated gender bias metric to diagnose a surprisingly low level of gender bias in a hate speech corpora like IdentityToxic and when we investigated this surprising result, we were indeed able to confirm that IdentityToxic contained mostly hate speech targeting race and sexual orientation, rather than gender. Our earlier work comparing corpora in English demonstrated the ability to meaningfully compare the gender bias across corpora in English, but the study described in this paper is the first to attempt to do such a comparison across languages.

Wikipedia offers an interesting basis for an initial cross-language comparison. As we have discussed the Wikipedia corpora in various languages vary substantially in size and quality, however, they do have the same goal of offering an open-collaborative online encyclopedia. They have similar patterns of authorship (i.e. maintained by a community of volunteer editors using a wiki-based editing system). The comparison is certainly not exactly apples-to-apples, but it is an interesting and meaningful one, especially given how often Wikipedia is used in NLP research.

In both Bolukbasi et al. and Babaeianjelodar et al. results across the profession set words were averaged evenly. Here, inspired by our findings with the weighted average in the last section, we compute a weighted average over entire documents or corpora as well. Instead of summing the gender bias scores for each profession word and dividing by the number of profession words, the weighted average adds the gender bias for each profession word each time it occurs and then divides by the total number of times a profession word occurs.

In Figure 7, we show both the average and weighted average. They are close, but different for all languages. Interestingly, the gender bias metric is negative or male for most of the languages. In Spanish the simple average is male biased and the weighted average is female biased. Besides Chinese and Wolof for which we have previously described some substantial concerns, only the weighted average in Spanish demonstrate overall female bias, but mildly so.

In future work, we would like to gain more experience using these metrics to compare more corpora for which we have more intuition of a ground truth- first within one language and then between languages. We expect there are important lessons to be learned from comparison within a language as we did in Babaeianjelodar et al. with English and then perhaps between gendered languages separately from non-gendered languages.

Even with this in mind, our methodology and the results presented here substantially improve on the state of the art in calculating and comparing gender bias across human languages. We would love to apply these metrics to the corpora we used in Babaeianjelodar et al. for which there is a ground truth understanding of their contents, but have not had an opportunity to do so. It is notable that those results were done with BERT rather than Word2Vec which although similar would introduce an additional level of variation in the results.

5 BEYOND WIKIPEDIA

Given how often Wikipedia is used in NLP research, documenting the difference in gender bias across Wikipedia corpora in different languages is a good first step. However, Wikipedia also has clear

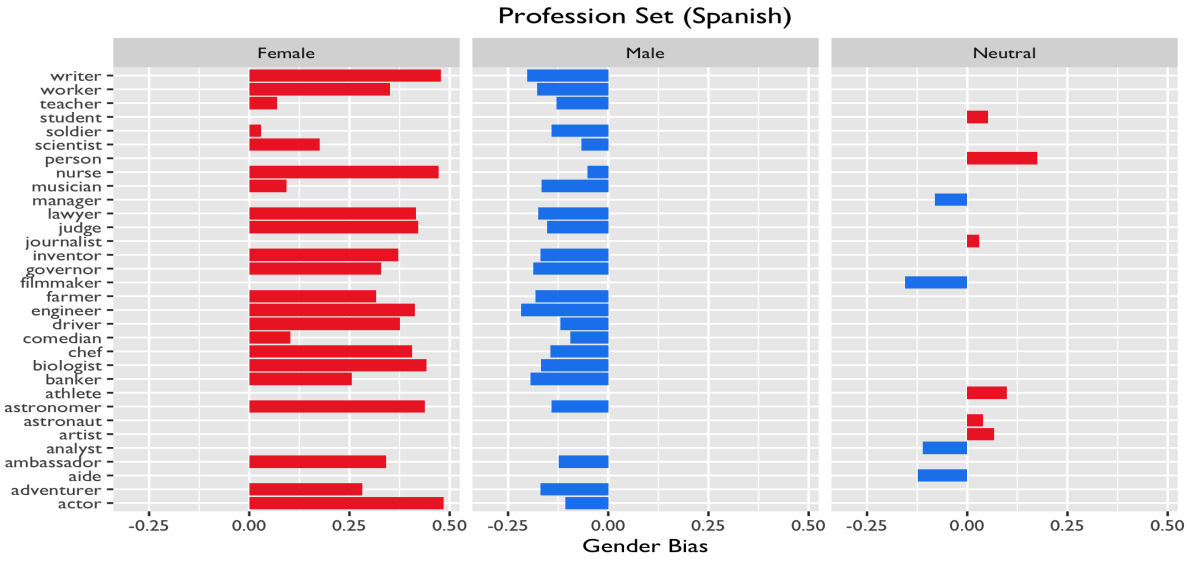

Figure 4: Per Profession Gender Bias for Spanish. Broken down into female only variants, male only variants and neutral variants.

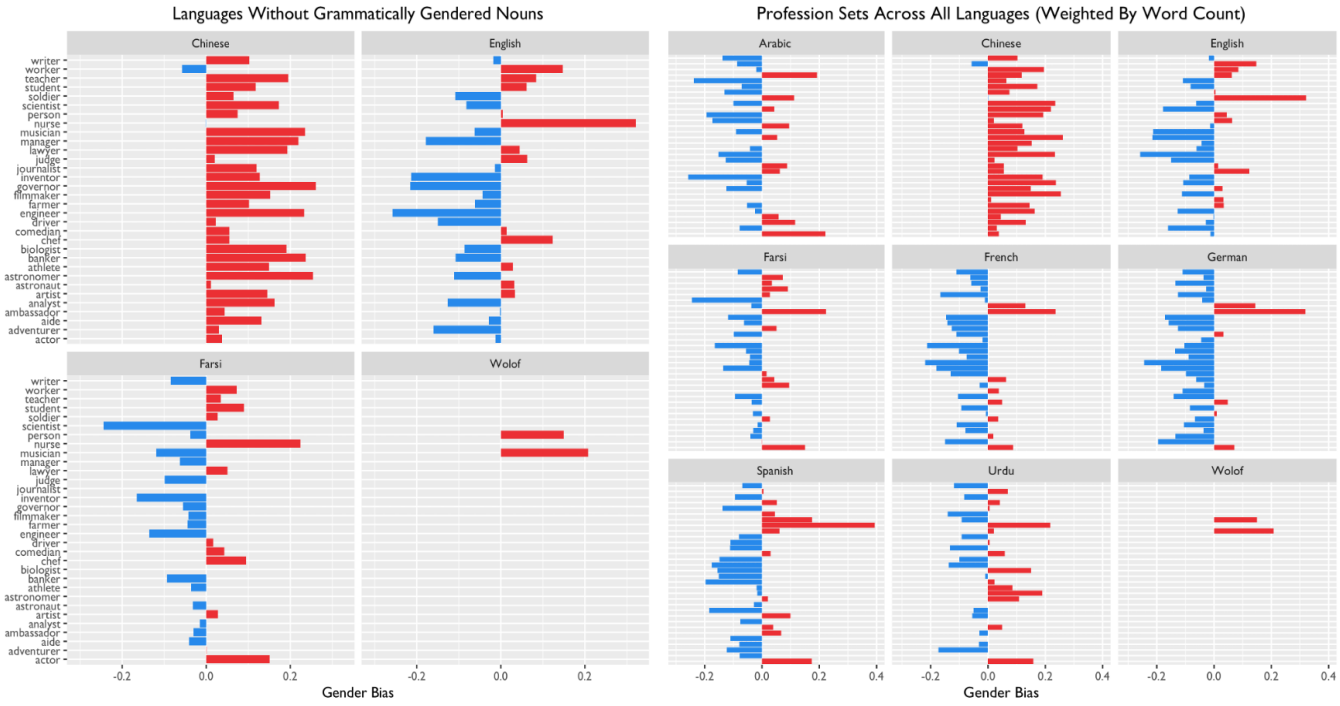

Figure 5: (LEFT)Per-Profession Gender Bias Metrics for Languages Without Grammatically Gendered Nouns
Figure 6: (RIGHT)Per-Profession Gender Bias Metrics for All Languages Weighting by Word Count

limitations and we would also like to experiment with additional corpora beyond Wikipedia.

A corpus of documents that are widely translated into many languages would be one interesting type of apples-to-apples comparison. According to Itc Translations, the Bible, the Universal Declaration of Human Rights and the Adventures of Pinocchio are among the most widely translated documents. However, clearly those documents would not be equally reflective of different cultures and in some languages may very well not even contain language generated by native speakers (e.g. they may be produced

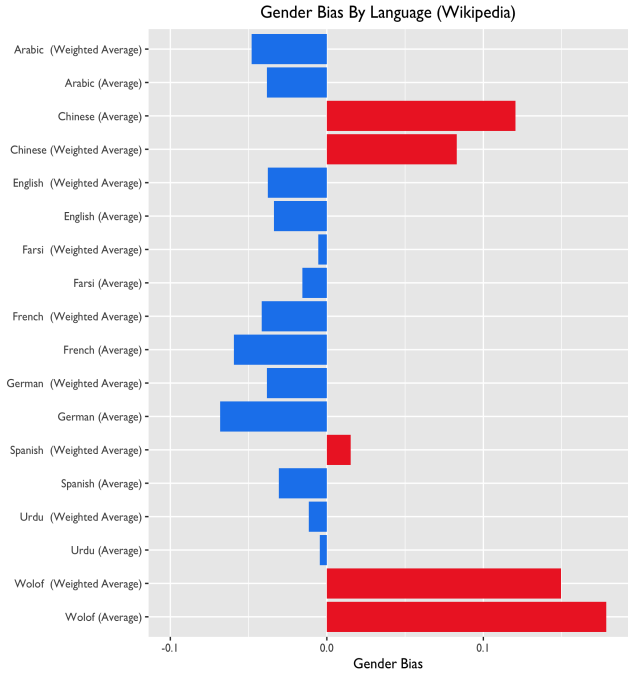

Figure 7: Per-Corpora Gender Bias Metrics: Overall gender bias across Languages for Wikipedia.

with translation software). We see similar problems with Wikipedia. Wikipedia prioritizes the voices of those creating content digitally and does not necessarily reflect the views of many speakers of those languages.

We would like to work with scholars in other disciplines such as sociology, linguistics and literature to compare gender bias across different corpora of culturally important texts written. by native speakers. Even within one language, it would be interesting to examine collections with different emphasis such as gender of author, different time periods, different genres of text, different country of origin,etc. We could also work to produce versions of our tools that scholars in these disciplines could use to analyze different corpora for themselves.

As an initial experiment in this space, we assembled a small collection of works that might be considered "classics" in English, Chinese and Spanish to provide an additional datapoint beyond just Wikipedia. Some example texts include *Pride and Prejudice* by Jane Austin and *The Great Gatsby* by F. Scott Fitzgerald in English, *Cien años de soledad* by Gabriel García Márquez and *Ficciones* by Jorge Luis Borges in Spanish and 司马迁 (Records of the Grand Historian) by Qian Sima and 萧红 (Tales of Hulan River) by Hong Xiao) in Chinese. Unlike with Wikipedia, these are pieces that have had a profound impact on the culture and were written by native speakers of the language. Where Wikipedia focuses on recently produced digital writing, many of these texts are older (e.g. as far back as 475 BC in the case of 论语 (The Analects) in Chinese).

The Chinese corpus we assembled in particular spans a wide range of years (475 BC-1992) and this allowed us to observe some important nuances that we did not see with Wikipedia. Classical Chinese (before 1900) uses a vocabulary and grammar that differs significantly from modern Chinese. It may not be surprising that there are changes in professions over that length of time, but we even found changes when it comes to some of the most fundamental defining set words. For example, the word woman can be translated in many ways, including "女子", "女人", and "妇女". "女子" was popularly used in ancient times, but its usage has decreased in modern writing. This shows us that as languages evolve over time, defining sets and profession sets may also have to evolve to measure gender bias.

6 CONCLUSION

In this paper, we extended an influential method for computing gender bias from Bolukbasi et al. It had only been applied in English and we made key modifications that allowed us to extend the methodology to 8 additional languages. Specifically, we modified and translated the original defining sets and profession sets. We extended the methodology to include languages with grammatically gendered nouns. With this, we quantified how gender bias varies across 9 languages within Wikipedia. We also assembled an initial classics corpora in 3 of the 9 languages and applied our methodology to it as well.

As speakers of 9 languages, we also used this process as an opportunity to shed light on the ways in which the modern NLP-pipeline does not reflect the voices of much of the world. For most languages, corpora are small and tool support is weak. Many published research methods, like Bolukbasi et al.'s gender bias metric calculations, are designed without consideration of the complexities of the multiple languages. This highlights the difficulties that speakers of many languages still face in having their thoughts and expressions fully included in the NLP-derived conclusions that are being used to direct the future. Despite substantial and admirable investments in multilingual support in projects like Wikipedia and Word2vec, we are still making NLP-guided decisions that systematically and dramatically under-represents many voices.

This work is an important 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. We focused on 9 languages from approximately 7,000 languages in the world. 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 and such that NLP can deliver more value to every part of the world.

For more information about our group and to access the extended tech report for this paper, please go to this link: http://tinyurl.com/clarksonnlpbias.

7 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

[1] Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. 2020. Quantifying Gender Bias in Different Corpora. In *Companion Proceedings of the Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 752–759. https://doi.org/10.1145/3366424.3383559

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

[3] BERT. 2020. BERT Pretrained models on Github. https://github.com/google-research/bert#bert.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 4356–4364.

[5] J; Buolamwini and T. Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (New York, USA) *(FAccT'18)*. 77–91.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

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

[8] D. Karani. 2018. Introduction to Word Embedding and Word2Vec. https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL]

[10] NLTK. 2005. Natural Language Toolkit. http://www.nltk.org/.

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

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

[13] 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. In *Participatory ML Workshop,Thirty-seventh International Conference on Machine Learning (ICML 2020), July 17 2020.*

[14] Wikipedia. 2020. German language. https://en.wikipedia.org/wiki/German_language.

[15] Wikipedia. 2020. List of languages by total number of speakers. https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers.

[16] Wikipedia. 2020. List of Wikipedias. https://en.wikipedia.org/wiki/List_of_Wikipedias.

[17] Wikipedia. 2020. Wolof Wikipedia. https://en.wikipedia.org/wiki/Wolof_Wikipedia.

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

[19] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. arXiv:1812.06280 [cs.CL]

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