Indigenous Language Revitalization and the Dilemma of Gender Bias

Oussama Hansal Université du Québec à Montréal Oussama.Hansal@courrier.uqam.ca

Ngoc Tan Le Université du Québec à Montréal le.ngoc_tan@uqam.ca Fatiha Sadat Université du Québec à Montréal sadat.fatiha@uqam.ca

Abstract

Natural Language Processing (NLP), through its several applications, has been considered as one of the most valuable field in interdisciplinary researches, as well as in computer science. However, it is not without its flaws. One of the most common flaws is bias.

This paper examines the main linguistic challenges of Inuktitut, an indigenous language of Canada, and focuses on gender bias identification and mitigation. We explore the unique characteristics of this language to help us understand the right techniques that can be used to identify and mitigate implicit biases. We use some methods to quantify the gender bias existing in Inuktitut word embeddings; then we proceed to mitigate the bias and evaluate the performance of the debiased embeddings. Next, we explain how approaches for detecting and reducing bias in English embeddings may be transferred to Inuktitut embeddings by properly taking into account the language's particular characteristics. We compare the effect of the debiasing techniques on Inuktitut and English. Finally, we highlight some future research directions which will further help to push the boundaries.

1 Introduction

Despite the complexity of low resource and endangered languages, the study of these languages has pulled many researchers in recent years, while this can be an encouraging factor for the development of language technologies, the complex morphology of some languages and the lack of resources have been considered as barriers. Moreover, as many NLP tasks are trained on human language data, it is expected for these applications to exhibit biases in different forms. Hovy and Prabhumoy (2021) described five sources where bias can occur in NLP systems: (1) the data, (2) the annotation process, (3) the input representations, (4) the models, and finally (5) the research design.

Gender bias can be defined as prejudice toward one gender over the other. Though usually tacit, bias range from the use of gender defaults to associating between occupation and gender. As language technologies become widespread and deployed on a large scale, their social impact raises concerns both internally and externally (Hovy and Spruit, 2016; Dastin, 2018). To capture the situation, Sun et al. (2019) reviewed NLP studies on this topic. However, their investigation is based on monolingual applications where the underlying assumptions and solutions may not directly apply to languages other than English. Thus, depending on the language involved and the factors taken into account, gender stereotypes have been conceptualized differently from study to study. To date, gender stereotypes have been addressed through a narrow problem-solving approach. While technical countermeasures are necessary, the failure to take a broader look at and engage with relevant literature outside of NLP could be detrimental to the growth of the field.

For example, when translating from English to French this following sentence, by Google Translate¹:

(en) *The engineer has asked the nurse to help her get up from the bed.*

(fr) L'ingénieur a demandé à l'infirmière de l'aider à se lever du lit.

We can see that it identified the engineer as a male and the nurse as a female, even though we used "her" to indicate that we are referring to a female. Such inadequacies not only jeopardize the development of endangered languages applications, but also perpetuate and amplify existent biases.

Understanding how human biases are incorporated into word embeddings can help us understand

¹https://translate.google.ca/, consulted at April 14th, 2022

bias in NLP models, given that word embeddings are commonly used in NLP. While some significant work has been done toward minimizing the bias in the embeddings, it has been proved that some methods are insufficient and that the bias can remain hidden within the embeddings. The words frequency is not taken into account, regardless of the gender distances, therefore biased terms can remain clustered together. Furthermore, when applied to contextualized word embeddings, these bias approaches must be changed because the embedding representation of each word varies based on the context.

This research intends to shed light on this issue by evaluating recent efforts to identify and mitigate bias within the indigenous languages revitalization and preservation context. We focus on Inuktitut, one of the main Inuit language of Eastern Canada and the official language of the government of Nunavut.

Thus, this paper is structured as follows: Section 2 presents the state-of-the-art. Section 3 presents the bias statement. Section 4 discusses the linguistic challenges of indigenous languages, with a focus on Inuktitut. Sections 5 highlights gender bias detection and mitigation. Section 7 presents the evaluations and the experimental results; while comparing with other existing approaches. Section 8 discusses the necessity of a human in the loop paradigm. Finally, Section 9 concludes this paper and presents potential future work.

2 Related Work

Interest in understanding, assessing, and reducing gender bias continues to grow in the NLP field, with recent studies showing how gender disparities affect the language technologies. Sometimes, for example, when visual recognition tasks fail to recognize female doctors (Zhao et al., 2017; Rudinger et al., 2018), image caption models do not detect women sitting next to the machine (Hendricks et al., 2018); and automatic speech recognition works best with male voices (Tatman, 2017). Although previously unconcerned with these phenomena in research programs (Cislak et al., 2018); it is now widely recognized that NLP tools encode and reflect asymmetries controversial society for many seemingly neutral tasks, including machine translation (MT). Admittedly, this problem is not new.

A few years ago, Schiebinger (2014) criticized the phenomenon of "missing men" in machine translation after conducting one of his interviews through a commercial translation system. Although there are some feminine mentions in the text, the female pronoun "*she*" is mentioned several times by the masculine pronoun. Users of online machine translation tools have also expressed concern about gender, having noticed how commercial systems manipulate society's expectations of gender, for example by projecting the translation of engineer into masculinity and that of medical science into femininity.

Bolukbasi et al. (2016) proved the existence of gender bias in English word embeddings, and proposed a method called Hard Debias to mitigate the gender bias. Liang et al. (2020) proposed a modified method that relies heavily on the sentences used to reduce biases.

We hypothesize that because English uses the common pronouns *he* and *she* extensively, which are not used in Inuktitut, as much as in English, for different reasons ²; the mitigation step encompasses a smaller gender subspace in comparison to English, and thus the bias is reduced.

Another method is the Iterative Null space Projection (INLP), which is a post-hoc method that can work on pre-trained representations (Ravfogel et al., 2020). The INLP's concept aims to identify task direction by training linear classifiers and removing direction from representation. INLP is effective in reducing gender bias. It was tested and showed great results in both word embeddings and contextualized word embeddings.

Most of the solutions were mainly proposed to reduce gender bias in English, and may not work as well when it comes to morphologically complex or polysynthetic languages. Nevertheless, there have been recent studies that explored the gender bias problem in languages other than English. Zhao et al. (2020) studied gender bias which is exhibited by multilingual embeddings in four languages (English, German, French, and Spanish) and demonstrated that such biases can impact cross-lingual transfer learning tasks.

Lewis and Lupyan (2020) examined whether gender stereotypes are reflected in the large-scale distributional structure of natural language semantics and measured gender associations embedded in the statistics of 25 languages and related them to data on an international dataset of psychological gender associations.

²https://uqausiit.ca/

Choubey et al. (2021) proposed gender-filtered self-training to improve gender translation accuracy on unambiguously gendered inputs. Their approach used a source monolingual corpus and an initial model to generate gender-specific pseudoparallel corpora, which were then filtered and added to the training data. They evaluated their method from English to five languages, which showed an improvement in gender accuracy without damaging gender equality.

Ntoutsi et al. (2020) presented a wide multidisciplinary overview of bias in AI systems, with an emphasis on technological difficulties and solutions, as well as new research directions toward approaches that are well-grounded in a legal framework.

The bias study in machine learning is not only restricted to the computer science field. Interdisciplinary research can help address this challenge across disciplines such as psychology, sociology, linguistics, cognitive science, and more (Datta, 2018). Hassan (2016) conducted a wide study on the influence that English has had on other language communities such as Inuit community. It can be seen in the way that it has affected gender relations specifically, by disempowering women in indigenous communities, the same as described in (Gudmestad et al., 2021). Men were assigned the role of hunting, and as such, became the "breadwinner" of the family. Women, on the other hand, were relegated to take care of the house and children, leaving them with no economic power and a perceived subordinate role within the family (Leigh, 2009).

According to Williamson (2006), the Inuits use a concept that encapsulates history, philosophy and observations of the world surrounding them. They call it "*Qaujimajatuqangit*" which is translated as "*traditional knowledge*". For Inuit people, "*Qaujimajatuqangit*" establishes gender equality in several fundamental ways. It respects the balance between the gender roles, the importance of family, and the fluidity of both gender and sexuality.

3 Bias Statement

Bias in NLP systems often goes without notice, it's often not even detected until after the systems are launched and used by consumers, which can have adverse effects on our society, such as when it shows false information to people which leads them to believe untrue things about society or themselves; thereby changing their behavior for better or worse (Stanczak and Augenstein, 2021). The harm of bias in NLP has been understated by some people and overstated by others, who dismiss its relevance or refuse to engage with it altogether. In this paper, we focus on the study of gender bias. If a system associates certain professions with a specific gender, this creates a representational harm. Representational harm is when an individual who falls into one of those categories is treated less fairly than someone outside of that category because of their belonging to it. For example, negative selection have been reported to occur more frequently in male dominated jobs than in other types of jobs (Davison and Burke, 2000). Similar conclusions have been made in the areas of competency assessments and performance evaluations, women were rated less positively than men in line jobs (which tend to be male gender-typed), but not in staff jobs, according to a prominent financial services organization (Lyness and Heilman, 2006). By looking at common examples of bias in the workplace, we can begin to understand how it can harm people in the office When such representations are being used in downstream NLP tasks. It can make the work environment feel less inclusive and less productive. Every single one of us has biases, but it's important to acknowledge when and how they impact our lives and the lives of others. According to recent research in NLP, word embeddings can incorporate social and implicit biases inherent in the training data (Swinger et al., 2019; Schlender and Spanakis, 2020; Caliskan, 2021). Current NLP models have proven to be good at detecting prejudices (Ahmed et al., 2022). However, unlike with prejudice, biases are not always obvious. While some biases are detectable via context, others might not be-which makes it difficult for automated systems to detect them. In fact, detecting and mitigating bias within automated systems prove to be more challenging than detecting it within human beings due to several important factors as dealing with imprecise sentiment analysis; as opposed to humans who can express nuanced sentiments when discussing bias. Our effort is predicated on the assumption that observed gender bias in systems are an indication of an insufficient interest into detecting and mitigating bias, we also believe that separating genders and professions in word embeddings would allow systems to detect and mitigate gender rather than promote it.

4 Linguistic Challenges in Indigenous Languages

In this section, we present the main linguistic challenges of Canada's indigenous languages, especially Inuktitut, an Inuit language of Eastern Canada and official language of the government of Nunavut. Thus, to better understand the challenges of NLP in Inuktitut, we explore the structure of Inuktitut words, the levels of grammatical variations, the dialectal variations in spelling, and gender animacy.

4.1 Morphological complexity

Most of the indigenous languages, particularly in the Americas, belong to either the polysynthetic language group or the agglutinative language group. They have a complex, rich morphology that plays an important role in human learning versus machine learning (Gasser, 2011; Littell et al., 2018). Much of the research on their morphological analysis has focused only on linguistic aspects.

Comparing word composition in English, the word structure in Inuit languages is variable in its surface form. Words can be very short, composed of three formative features such as word base, lexical suffixes, and grammatical ending suffixes. Or they can be very long up to ten or even fifteen formative morphemes as features depending on the regional dialect (Lowe, 1985; Kudlak and Compton, 2018; Le and Sadat, 2020, 2022).

4.2 Morphophonemics

The morphophonemics of Inuktitut are highly complex, in addition to the variety of morphological suffixes that Inuktitut roots can take on (Mithun, 2015). In Inuktitut, each morpheme specifies the sound variations that can occur to its left and/or to itself. These modifications are phonologically conditioned by the individual morphemes themselves, rather than their contexts. This not only aggravates the data sparsity issue, but it also poses morphological analysis issues, which we shall address in the research topics of this project.

4.3 Dialectal variations

The third aspect of Inuktitut which contributes to the challenge of processing it with a computer is the abundance of spelling variation seen in the electronically available texts. Inuktitut, like all languages, can be divided into a number of different dialects, such as Uummarmiutun, Siglitun, Inuinnaqtun, Natsilik, Kivallirmiutun, Aivilik, North Baffin, South Baffin, Arctic Quebec, and Laborador (Dorais, 1990). The primary distinction between these dialects is phonological, which is reflected in spelling. As a result, spelling variance, either due to a lack of standardisation or due to numerous dialect changes, contributes significantly to the overall sparsity of the data in the corpora accessible for experimentation (Micher, 2018).

4.4 Gender animacy

Inuit languages are known to have some particular linguistics challenges. There is no gender marking in nouns, like you'll find in French and Spanish (*male | female*) nouns. Instead, Inuktitut distinguishes words along a dimension called *animacy*, because of the cultural understanding as to whether a noun is known to be alive or not. The singular and plural suffixes that are used in nouns, depend on whether is is animate or inanimate.

The animacy is described as a distinction between human and non-human, rational and irrational, socially active and socially passive³. For example, animate nouns are related to humans and animals most obviously, but other objects that are not considered alive, like stone, table, are considered as inanimate. Animate and inanimate gender is common in many Amerindian families such as Cree, Inuktitut, Quechuan, Aymara, Mapudungun, Iroquoian, and Siouan⁴.

5 Bias detection and mitigation

Although existing machine learning models achieve great results on many tasks, they generally fail in avoiding biases. Recent studies illustrate how bias affect NLP technologies, which has created a growing interest in identifying, analysing and mitigating bias within the NLP community. The problem is not new, it is well-known that NLP systems contain and reflect algorithmic bias in them, this controversial imbalances has developed a large scale of concerns about its social impact. NLP systems and tools are used in everyday life, The time of academic naivety is finished, therefore we must acknowledge that our models have an impact on people's lives, but not necessarily in the way we intend (Ehni, 2008).

³https://en.wikipedia.org/wiki/List_of_languages_by _type_of_grammatical_genders

⁴https://linguisticmaps.tumblr.com/post/169273617313/ grammatical-gender-or-noun-class-categories-new

To contextualize the plan within this larger research area, we will focus on indigenous languages that proves no exception to the existent problem of bias in NLP systems. Indigenous languages contain a wealth of secondary data about individuals, their identity and their demographic group, which are exploited to fulfil the objective of creating NLP systems. The focus on creating these systems has drifted us away from creating models as tools of understanding towards other tools that produce great results but are far more difficult to understand (Hovy and Prabhumoye, 2021).

Many questions may arise, such as: Is it possible that NLP models are biased by definition? What could be the source of this bias? Can we figure out what it is? Is there anything we can do about it?

5.1 Definition of Bias

Bias is a complex concept with overlapping definitions (Campolo et al., 2017). It has been considered as a fundamental human decision-making process since the beginning of time (Kahneman and Tversky, 1973). When we apply a cognitive bias, we are assuming that reality will behave in accordance with prior cognitive convictions that may or may not be accurate, with which we can make a judgement (Garrido-Muñoz et al., 2021). According to the Sociology dictionary⁵, bias is a term used to describe an unjust prejudice in favour of or against a person, group, or thing.

Machine learning bias can happen in a variety of ways, ranging from racial and gender discrimination to age discrimination. It also exists in machine learning algorithms throughout their development, which is the root problem of machine learning bias. Therefore, human biases are adopted and scaled by machine learning systems.

5.2 Types of Bias

Machine learning models incorporate bias in many shapes, including gender, racial and religious biases extending to unfair recruiting and age discrimination. But what are really the machine learning types of bias?

According to (Shashkina, 2022), the most common types of machine learning bias found in algorithms are listed below:

• Reporting bias: It happens when the frequency of occurrences in the training dataset does not

precisely reflect reality.

- Selection bias: This sort of bias happens when training data is either unrepresentative or not randomly selected.
- Group attribution bias: It happens when machine learning systems generalize what is true of individuals to entire groups that the individual is or is not a part of.
- Implicit bias: It happens when machine learning systems are based on data that is created on personal experience which does not necessarily apply broadly.

5.3 Mitigating Bias

We still have a long way to go before machine learning bias is completely eliminated. With the increased usage of machine learning systems in sensitive domains such as banking, criminal justice, and healthcare, we should aim to create algorithms that reduce bias in machine learning systems. Collaboration between human skills and machine learning is required to solve the problem of bias in machine learning. It will help us in the detection and mitigation of biases by figuring out how machine learning systems make predictions and what data aspects it uses to make judgments. This will help us understand whether the elements influencing the choice are biased.

6 Bias Mitigation for Inuktitut

In this study, we use a methodology and data for bias mitigation in Inuktitut, as described in the following section. To analyse and mitigate bias in word embeddings, multiple sets of data (e.g. pairs of sentences, lists of gendered words, and combinations of sentences from different categories) are required. Two algorithms are used to measure bias in embeddings, which are applicable to traditional embeddings. Then we demonstrate how we mitigate bias in either type of embedding and examine how well the bias mitigation works on downstream tasks. Furthermore, because this study is based on Inuktitut embeddings, the data used is from the Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 (Joanis et al., 2020) as for English.

6.1 Bias Measuring Methods

Word Embedding Association Test (WEAT)

This method, proposed by Caliskan et al. (2017), helps to measure human bias in data presented as

⁵Open Education Sociology Dictionary: https:// sociologydictionary.org/bias/

texts. It is similar to the Implicit Association Test (IAT) proposed by (Greenwald et al., 1998). The similarity of IAT and WEAT consists of using two lists of target words and two lists of attribute words. The first pair of lists represents the terms we want to compare and the second pair of lists represents the categories in which we suspect bias could exist (Mulsa and Spanakis, 2020). By using WEAT, Caliskan et al. (2017) defined ten tests to assess the bias in several areas (Mulsa and Spanakis, 2020).

In our study we converted the WEAT lists of words used in the tests to Inuktitut and modified them such that terms in these lists are only related with the appropriate category. Some of the modifications correspond to the different linguistic characteristics of the language and the lack of meaningful translations of certain words in the data. Some other changes are due to the language's various linguistic peculiarities and the lack of relevant translations for particular words in the data.

Clustering accuracy

Gonen and Goldberg (2019) provided a new metric that shows that word embeddings with reduced bias can stay grouped together even when the range across attributes and targeted words (in WEAT) is minimal. To determine the gender orientation of each word in the lexicon, the clustering accuracy test necessitates projecting the entire vocabulary into male and female terms (Mulsa and Spanakis, 2020).

The pronouns *he* and *she* were used by Gonen and Goldberg (2019), because they are commonly used and the only variation between them is in the gender subdomain.

Inuktitut has few personal pronouns, either in first person (I, we) or second person (you) ⁶; which represents a problem in this research by adding extra meaning besides gender to the geometrical difference of the pronouns (Mulsa and Spanakis, 2020).

6.2 Debiasing Methods

In this section, we present the debiaising methods used in this research with an application on the Inuktitut language.

Hard debias (Bolukbasi et al., 2016)

One of the earliest strategies used to detect and minimise bias in word embeddings was *Hard Debias*. Through post-processing, it removes gender bias by subtracting the component linked with gender from all embeddings. It takes a set of gender-specific word pairs and computes the gender direction in the embedding space as the first principal component of difference vectors of these pairs. Furthermore, it removes gender bias by projecting biased word embeddings onto a subspace orthogonal to the assumed gender direction (Bolukbasi et al., 2016). The gender orientation is skewed by the frequency of words.

SENT debias (Liang et al., 2020)

SENT-Debias is divided into four steps: 1) identifying words with bias attributes; 2) contextualising these words into bias attribute sentences and, as a result, their sentence representations; 3) estimating the sentence representation bias subspace; and 4) debiasing general sentences by eliminating the projection onto this bias subspace. These processes are summarized in Figure 1.

SENT-DEBIAS:	
 Initialize (usually pretrained) sentence encoder M_θ. 	
 Define bias attributes (e.g. binary gender g_m and g_f). 	
 Obtain words D = { (w₁⁽ⁱ⁾,,w_d⁽ⁱ⁾) }^m_{i=1} indicative of bias attribute. 	
4: $S = \bigcup_{i=1}^{m} \text{CONTEXTUALIZE}(w_1^{(i)},, w_d^{(i)}) = \{(s_1^{(i)},, s_d^{(i)})\}_{i=1}^{n}$	// words into sentences
5: for $j \in [d]$ do	
6: $R_j = \{M_\theta(s_i^{(i)})\}_{i=1}^n$	// get sentence representations
7: end for	
8: $\mathbf{V} = \mathbf{PCA}_k \left(\bigcup_{i=1}^d \bigcup_{\mathbf{w} \in \mathcal{R}_i} (\mathbf{w} - \boldsymbol{\mu}_i) \right)$	// compute bias subspace
9: for each new sentence representation h do	
10: $\mathbf{h}_{\mathbf{V}} = \sum_{j=1}^{k} \langle \mathbf{h}, \mathbf{v}_{j} \rangle \mathbf{v}_{j}$	// project onto bias subspace
11: $\hat{\mathbf{h}} = \mathbf{h} - \mathbf{h}_{\mathbf{V}}$	// subtract projection
12: end for	

Figure 1: SENT Debias Algorithm (Liang et al., 2020).

Iterative NullSpace Projection (Ravfogel et al., 2020)

INLP stands for Iterative Nullspace Projection, which is a method for eliminating data from neuronal representations (Figure 2). This algorithm is built on repeatedly training linear classifiers that predict a specific property that we want to eliminate; then projecting the representations onto their null-space. As a result, the classifiers loose sight of the target property, making it difficult to linearly divide the data based on it. While this method is relevant to a variety of applications, it was tested on bias and fairness use-cases and demonstrated that it can mitigate bias in word embeddings.

7 Data and Evaluations

We conducted some experiments on gender bias mitigation in Inuktitut language. We used the Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 (Joanis et al., 2020). The statistics of the training corpus are described in Table 1.

⁶https://uqausiit.ca/grammar-book

Input : (X, Z) : a training set of vectors and pro-
tected attributes
n: Number of rounds
Result: A projection matrix P
Function GetProjectionMatrix (X, Z) :
$X_{projected} \leftarrow X$
$P \leftarrow I$
for $i \leftarrow 1$ to n do
$W_i \leftarrow \text{TrainClassifier}(X_{projected}, Z)$
$B_i \leftarrow \text{GetNullSpaceBasis}(W_i)$
$P_{N(W_i)} \leftarrow B_i B i^T$
$P \leftarrow P_{N(W_i)}P$
$X_{projected} \leftarrow P_{N(W_i)} X_{projected}$
end return P

Figure 2: INLP Algorithm (Ravfogel et al., 2020).

Dataset	#tokens	#train	#dev	#test
Inuktitut	20,657,477	1,293,348	5,433	6,139
English	10,962,904	1,293,348	5,433	6,139

 Table 1: Statistics of Nunavut Hansard for Inuktitut-English

We performed our experiment using word embeddings, trained on the Nunavut Hansard for Inuktitut-English. In order to pre-train the embeddings for Inuktitut, we used an Inuktitut segmenter to segmentate the words before passing it to the FastText toolkit (Bojanowski et al., 2016). The model was trained for 40 epochs and we used 150 and 300 as the size of the dense vector to represent each token or word. In order to get terms that are more related and close to each other we used a small window of 2 which give us the maximum distance between the target word and its neighboring word. We also used an alpha value of 0.03 to preserve the strong correlation of the model after each training example is evaluated.

We performed the WEAT test on the adapted lists of words translated to Inuktitut. Among all the traditional word embeddings, we see high effect sizes and multiple tests are significant at different levels. The results of the WEAT effect sizes on gendered related tests are shown in Table 2 where we see an overall high effect size across all the scores on the original models.

The results of the WEAT effect sizes on gendered related tests are shown in Table 2 where we see a high effect size on the word embeddings debiased from the original models. The results after the debiasing step shows that the bias mitigation is

WEAT			
Methods	Original	Debiased	
SENT debias	0.0338	0.499	
INLP	0.0338	0.377	
Hard Debias	0.0338	0.385	

Table 2: Fasttext WEAT results, with significance of p-value, for three methods such as Sent debias, INLP, and Hard debias. Bold values are better.

effective in every model. An example of the list of words used is illustrated below in Table 3.

WEAT words list example				
	Category	Inuktitut		
0	family	angajuqqaaq		
1	prof	executive		
2	prof	ilisaiji		
3	male names	jaan		
4	female	maata		
	names			

Table 3: Example of WEAT words list

Because Inuktitut is a genderless language, it can be difficult to use pronouns. Therefore following (Gonen and Goldberg, 2019), we used common names for males and females instead of specifically gendered words to indicate the male and female categories (e.g. pronouns). Three tests compare the associations of male and female names to (1)job and family-related words, (2) art words, and (3) scientific domains. We observe that, following the projection, the substantial relationship between the groups is no longer there in the three tests. Figure 3 shows projections of the 200 most female-biased and 200 male-biased words projected at t = 1, which is basically the original state, and t = 35which is the final state after debiasing. These results represent the INLP method. The results clearly demonstrate that the classes are no longer linearly separable in the INLP method. This behavior is qualitatively different from the Sent debias and the Hard debias methods; which are shown to maintain much of the proximity between female and male-biased vectors.

7.1 Discussion

We hypothesize, in this paper, that identifying the true gender orientation of word embeddings using these existing Debias approaches could be challenging. We show that the geometry of word em-



Figure 3: Example of biased clusters from original to debiased states, using t-distributed stochastic neighbor embedding (t-SNE)

beddings is influenced by word frequency. Popular and rare words, for example, cluster in various subregions of the embedding space, regardless of the fact that the words in these clusters are semantically unrelated. This may have a negative impact on the process of determining gender direction and, as a result, the efficacy of debiasing methods to debias the gender. We saw that changing the frequency of certain phrases causes large changes in the similarities between the related difference vector and other difference vectors.

We noticed, in the context of gender bias, one disadvantage that we found out, is that all of our 3 debiasing methods, like other learning approaches, are dependent on the data that is supplied to it; and assumes that the training data is suitably large and sampled from the same distribution as the test data. In practice, this requirement is difficult to achieve, and failing to supply properly representative training data may result in biased classifications even after it has been applied.

We further emphasize that the WEAT and clustering tests do not test for the absence of bias; rather, they test if bias exists in the test instances, but bias may also exist in non-tested cases. Even if we measure bias from a different perspective, the bias remains, indicating that more studies on bias mitigation approaches are needed.

8 Human-in-the-Loop Paradigm

For indigenous peoples in general, the language is directly connected to their culture and identity. Thus, it is very important for indigenous peoples of Canada, to both, speak their language and practice their culture. Inuktitut not only represents the official language of Inuits but also represents the rich culture of this community. With recent advances, NLP models represent a big opportunity for the development of tools that will further help in preserving the language with respect for the culture and realities of the indigenous people where the language takes a big part of it.

Most communities in Nunavut offer Inuktitut or Inuinnaqtun for the first few years of education, and the government has vowed to develop completely bilingual students across the territory ⁷. As a result, the problem remains unsolved. As a nonindigenous people with a strong academic interests in social science, linguistics and NLP, Dorais (2010) cites that gaining a better grasp of the general sociolinguistic situation in Northern Canada is the first step toward a true solution to the Inuit culture and language difficulties. It is insufficient to describe how Inuit people communicate (which is the task of linguists). We must also attempt to comprehend what they are saying and what language means to them (Dorais, 2010). Revitalizing indigenous language should be done for, by and with indigenous communities. With the emergence of AI, especially deep learning, there is a large interest for the revitalization of indigenous languages. However, there is little interest in the field of computer science, and there are also very few or no researchers from Canada's Indigenous communities in the field of NLP.

⁷Source: https://www. thecanadianencyclopedia.ca/en/article/ inuktitut

It's evident that human skills like insight and creativity be easily computerized, therefore collaborating human skills with machine learning technologies is a great approach to keep human in the loop for developing technologies for us. Before building machine learning algorithms, it's a good idea to consult with humanists and social scientists to verify that the models we create don't inherit any of the biases that people have.

Machine learning models can assist us in revealing flaws in human decision-making. So, if these models trained on current human decisions reveal bias, it will be important to have a second look from human to keep this models fair. In the case of developing machine learning technologies for indigenous communities, it is important to keep the collaboration and partnership with them; before, while and after developing tools for them. Engaging communities to develop machine learning tools is very important, not only it will make the tool more suitable and tailored to their needs but it will also give the ownership to these communities.

9 Conclusion

This paper demonstrates that gender bias exists in Inuktitut, among other biases (as probably in other languages as well). Then, by appropriately translating the data and taking into account the language's specific characteristics, we illustrated how approaches used to measure and reduce biases in English embeddings can be applied to Inuktitut embeddings. Furthermore, we investigated the influence of mitigating approaches on downstream tasks, finding a major effect in traditional embeddings, which could be regarded as favourable if the embeddings utilised guarantee a more genderneutral approach. As a future work, we plan to investigate other types of biases in Inuktitut and collaborate with the Indigenous community. Our main objective remain the revitalization and preservation of Indigenous languages of Canada, using NLP and machie learning techniques. We hope that these exploratory results will encourage researches on Indigenous and Endangered languages.

References

Zo Ahmed, Bertie Vidgen, and Scott A Hale. 2022. Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning. *EPJ Data Science*, 11(1):8.

- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
- 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.
- Aylin Caliskan. 2021. Detecting and mitigating bias in natural language processing. *Res. Rep, Brookings Inst., Washington, DC [Google Scholar].*
- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. Ai now report 2017. *New York: AI Now Institute*.
- Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. Improving gender translation accuracy with filtered self-training. arXiv preprint arXiv:2104.07695.
- Aleksandra Cislak, Magdalena Formanowicz, and Tamar Saguy. 2018. Bias against research on gender bias. *Scientometrics*, 115(1):189–200.
- Jeffrey Dastin. 2018. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of Data and Analytics*, pages 296–299. Auerbach Publications.
- Ranjan Datta. 2018. Decolonizing both researcher and research and its effectiveness in indigenous research. *Research Ethics*, 14(2):1–24.
- Heather K Davison and Michael J Burke. 2000. Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior*, 56(2):225–248.
- Louis-Jacques Dorais. 1990. L'étranger aux yeux du francophone de québec. *Recherches sociographiques*, 31(1):11–23.
- Louis-Jacques Dorais. 2010. Language of the Inuit: syntax, semantics, and society in the Arctic, volume 58. McGill-Queen's Press-MQUP.
- Hans-Jörg Ehni. 2008. Dual use and the ethical responsibility of scientists. *Archivum immunologiae et therapiae experimentalis*, 56(3):147–152.
- Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.
- Michael Gasser. 2011. Computational morphology and the teaching of indigenous languages. In *Indigenous Languages of Latin America Actas del Primer Simposio sobre Enseñanza de Lenguas Indígenas de América Latina*, page 52.

- Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.
- Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.
- Aarnes Gudmestad, Amanda Edmonds, and Thomas Metzger. 2021. Moving beyond the native-speaker bias in the analysis of variable gender marking. *Frontiers in Communication*, page 165.
- Jenna N Hassan. 2016. De-colonizing gender in indigenous language revitalization efforts. *Western Papers in Linguistics/Cahiers linguistiques de Western*, 1(2):4.
- Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787.
- Dirk Hovy and Shrimai Prabhumoy. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass.*
- Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.
- Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings* of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 591–598.
- Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, and Darlene Stewart. 2020. The nunavut hansard inuktitut–english parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 2562—2572.
- Daniel Kahneman and Amos Tversky. 1973. On the psychology of prediction. *Psychological review*, 80(4):237.
- Emily Kudlak and Richard Compton. 2018. Kangiryuarmiut Inuinnaqtun Uqauhiitaa Numiktitirutait — Kangiryuarmiut Inuinnaqtun Dictionary, volume 1. Nunavut Arctic College: Iqaluit, Nunavut.
- Ngoc Tan Le and Fatiha Sadat. 2020. Revitalization of indigenous languages through pre-processing and neural machine translation: The case of inuktitut. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4661–4666.
- Ngoc Tan Le and Fatiha Sadat. 2022. Towards a lowresource neural machine translation for indigenous languages in canada. *Journal TAL, special issue on Language Diversity*, 62:3:39–63.

- Darcy Leigh. 2009. Colonialism, gender and the family in north america: For a gendered analysis of indigenous struggles. *Studies in Ethnicity and Nationalism*, 9:70 – 88.
- Molly Lewis and Gary Lupyan. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour*, 4(10):1021– 1028.
- Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.
- Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in canada: Assessment, challenges, and successes. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2620–2632.
- Ronald Lowe. 1985. Basic Siglit Inuvialuit Eskimo Grammar, volume 6. Inuvik, NWT: Committee for Original Peoples Entitlement.
- Karen S Lyness and Madeline E Heilman. 2006. When fit is fundamental: performance evaluations and promotions of upper-level female and male managers. *Journal of Applied Psychology*, 91(4):777.
- Jeffrey C Micher. 2018. Addressing challenges of machine translation of inuit languages. Technical report, US Army Research Laboratory Adelphi United States.
- Marianne Mithun. 2015. Morphological complexity and language contact in languages indigenous to north america. *Linguistic Discovery*, 13(2):37–59.
- Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating bias in dutch word embeddings. *arXiv preprint arXiv:2011.00244*.
- Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human

Language Technologies, Volume 2 (Short Papers), pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

- Londa Schiebinger. 2014. Scientific research must take gender into account. *Nature*, 507(7490):9–9.
- Thalea Schlender and Gerasimos Spanakis. 2020. 'thy algorithm shalt not bear false witness': An evaluation of multiclass debiasing methods on word embeddings. In *Benelux Conference on Artificial Intelligence*, pages 141–156. Springer.
- Victoria Shashkina. 2022. Ai bias: Definition, types, examples, and debiasing strategies. *https://itrexgroup.com/blog/ai-bias-definitiontypes-examples-debiasing-strategies/header*, (1):1.
- Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.
- Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.
- Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305– 311.
- Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59.
- Laakkuluk J. Williamson. 2006. Inuit gender parity and why it was not accepted in the nunavut legislature. *Études/Inuit/Studies*, 30(1):51–68.
- Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and crosslingual transfer. *arXiv preprint arXiv:2005.00699*.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.