UNIVERSITÉ DU QUÉBEC À MONTRÉAL

GENDER BIAS IN THE CONTEXT OF INDIGENOUS LANGUAGES AND NATURAL LANGUAGE PROCESSING

DISSERTATION

PRESENTED

AS PARTIAL REQUIREMENT

TO THE MASTERS IN COMPUTER SCIENCE

BY

OUSSAMA HANSAL

SEPTEMBER 2023

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

BIAIS DE GENRE DANS LE CONTEXTE DES LANGUES AUTOCHTONES ET DU TRAITEMENT DU LANGAGE NATUREL

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

OUSSAMA HANSAL

SEPTEMBRE 2023

ACKNOWLEDGEMENTS

Throughout the writing of this dissertation, I have received a great deal of support and assistance.

I would first like to express my sincere appreciation to my research director, Ms. Fatiha Sadat, Professor at the University of Quebec at Montreal (UQAM), for supervising my dissertation, for her valuable support, help, great advice, and encouragement during the entire period of this work. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

My sincere acknowledgements go to all the professors of the Computer Science department of UQAM, for the quality of their teaching.

A big thank you to the members of the jury for their interest in my work by agreeing to examine this dissertation and complement it with their valuable proposals.

I also thank all of my lab colleagues for their support and help.

Finally, I could not have completed this dissertation without the support of my wife, my parents, and my siblings. Thank you for your patience, moral support, and encouragement throughout my years at the university.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF NOTATIONS

AI	Artificial Intelligence
ML	Machine Learning
NLP	Natural Language Processing
INLP	Iterative NulL-space Projection
MT	Machine Translation
NMT	Neural Machine Translation
TGBI	Translation Gender Bias Index
GFST	Gender-Filtered Self-Training
WEAT	Word Embedding Association Test
IAT	Implicit Association Test
NN	Neural Network-based
BPE	Byte Pair Encoding
TER	Translation Error Rate

RÉSUMÉ

Le Traitement Automatique du Langage Naturel (TALN) est un domaine qui combine la linguistique, les sciences cognitives, l'informatique notamment l'Intelligence Artificielle (IA) pour explorer l'interaction des systèmes informatiques avec le langage humain. Au cours des dernières années, les modèles de TALN ont été largement utilisés dans diverses applications du monde réel, allant de la traduction automatique aux outils de développement des langues en danger et/ou autochtones. Grâce aux récents progrès de l'apprentissage profond, les modèles de TALN jouent un rôle important dans la promotion de la diversité linguistique pour les langues à faibles ressources. Malgré la popularité et l'influence croissantes de ces modèles, ils sont toujours soumis à des préjugés sexistes, sociaux et implicites inhérents aux données d'entraînement.

Ce travail de recherche explore l'inuktitut, une langue autochtone parlée par les communautés inuites du nord du Canada, caractérisée par sa grammaire, sa morphologie complexe et sa structure unique qui la rendent difficile à traiter informatiquement par rapport à l'anglais ou le français. Dans cette étude, nous nous concentrons sur la tâche de détection et d'atténuation des biais de genre en inuktitut, ce qui pose plusieurs difficultés et défis qui font partie intégrante des caractéristiques du traitement des données dans cette langue. Nous examinons les principaux défis linguistiques de l'inuktitut et nous nous concentrons sur la détection et l'atténuation du biais de genre dans cette langue.

Dans un premier temps, nous utilisons des méthodes de regroupement et de WEAT pour quantifier les biais de genre en inuktitut. Ensuite, nous proposons d'atténuer le biais de genre existant dans le corpus de l'inuktitut à l'aide de trois méthodes importantes : le "HARD Debias", le "SENT Debias" et la projection itérative dans l'espace vide (INLP). Enfin, nous évaluons les performances de ces modèles dans des tâches en aval et expliquons comment les approches de détection et de réduction des biais de genre dans les embeddings anglais peuvent être transposées aux embeddings inuktitut en tenant compte des caractéristiques particulières de la langue.

Nous avons mené différents types d'expériences pour optimiser et évaluer les approches proposées, car elles sont axées sur l'anglais ou d'autres langues européennes. Sur la base des évaluations et des résultats obtenus, nous constatons que le biais de genre est présent dans le corpus d'inuktitut que nous avons traité, et nous soulignons également que les techniques utilisées pour mesurer et réduire le biais de genre pour l'anglais peuvent être utilisées pour l'inuktitut. De plus, nous examinons l'impact des méthodes de débiaisage sur les tâches en aval, qui ne montrent pas d'effet significatif sur la performance. Enfin, nous présentons une liste des mots les plus biaisés détectés dans le corpus inuktitut.

Mots-clés : Traitement Automatique du Langage Naturel, TALN, Inuktitut, Biais de genre, Apprentissage profond.

ABSTRACT

Natural language processing (NLP) is a field that combines linguistics, cognitive science, computer science, more specifically Artificial Intelligence (AI) to explore the interaction between computers and human language. In recent years, NLP models have been widely used in various real-world applications, ranging from machine translation to the development of tools for endangered and Indigenous languages. With recent advancements in deep learning, NLP models play a crucial role in promoting language diversity and equity while exploring low-resource languages. Despite the increasing popularity and influence of these models, they are still susceptible to gender, social, and implicit biases inherent in the training data.

This research work delves into Inuktitut, an Indigenous language spoken by Inuit communities in Northern Canada. Inuktitut is characterized by its unique grammar, complex morphology and unique structure, which pose computational challenges compared to other languages such as English or French. This study focuses on the task of detecting and mitigating gender bias in Inuktitut, which presents several difficulties and challenges intrinsic to the data processing characteristics of this language. We explore the primary linguistic challenges of Inuktitut and concentrate on the detection and mitigation of gender bias within this language.

Initially, we employ clustering and WEAT methods to quantify gender bias in Inuktitut. Subsequently, we propose three significant methods, namely "Hard Debias," "SENT Debias" and "Iterative Null Space Projection (INLP)", to mitigate the existing gender bias in the Inuktitut corpus. Finally, we evaluate the performance of these models in downstream tasks and explain how approaches for detecting and reducing gender bias in English embeddings can be adapted to Inuktitut embeddings, taking into account the language's specific characteristics.

We conducted various experiments to optimize and assess the proposed approaches, focusing on English and other European languages.

Based on the evaluations and obtained results, we observe the presence of gender bias in the processed Inuktitut corpus. Additionally, we emphasize that the techniques used to measure and reduce gender bias in English can be applied to Inuktitut. Furthermore, we examine the impact of debiasing methods on downstream tasks, which show insignificant effects on performance. Lastly, we present a list of the most biased words detected in the Inuktitut corpus.

Keywords: Natural Language Processing, NLP, Inuktitut, Gender bias, Deep learning.

INTRODUCTION

Natural language processing (NLP) has tremendous potential to benefit society. By processing and analyzing language, machines can power a variety of useful applications, including personal assistants, medical record management systems, spam filters, and search engines that help people find information quickly and efficiently. However, despite its importance, bias in NLP systems often goes unnoticed and is often not detected until after the systems have been launched and used by consumers.

As AI adoption accelerates, it becomes increasingly important to minimize bias in AI models, and we all have a role to play in identifying and mitigating bias so that we can use AI reliably and positively. It is essential to acknowledge the limitations of our data, models, and technical solutions to bias, both to raise awareness and to allow for the consideration of human methods to limit bias in machine learning.

Like many other problems, bias in NLP can be addressed early or late in the process. There are numerous sources of unintended demographic bias in the NLP pipeline. Bias can occur at all stages of a machine learning pipeline, and implementing mitigation strategies at each stage is critical to eliminating bias. Biases can emerge and infiltrate the process at multiple stages, starting from data collection and persisting through various subsequent steps. These biases have been extensively studied and documented in recent research (Hovy & Prabhumoye, 2021a; Chang et al., 2019; Caliskan, 2021) shedding light on their pervasive nature and potential implications

Most of the models used to detect bias in NLP were developed around the detection of prejudice (Michael Kearns et al., 2022) . However, prejudice is different from bias. Prejudice is an attitude that involves hatred or intolerance towards a group of people based on their race, ethnicity, gender identity, sexuality, or other personal characteristics. Current NLP models have proven to be effective at detecting prejudices. However, unlike prejudice, biases are not always obvious. While some biases can be detected through context, others may not, making it difficult for automated systems to identify them. In fact, detecting and mitigating bias within automated systems proves to be more challenging than detecting it within human beings due to several important factors, such as dealing with imprecise sentiment analysis. Humans, on the other hand, can express nuanced sentiments when discussing bias.

As mentioned earlier, researchers have developed methods to detect bias, such as algorithms used for processing language data. However, the problem arises from the fact that these algorithms, like any other, are created by humans who inevitably bring their own biases into their work, whether implicit or cognitive. Therefore, even if we were to develop our own NLP algorithms from scratch, there would still be a risk of bias entering our output. This issue becomes even more problematic with public data services that allow anyone with an internet connection to upload data sets for processing. Once these data sets are submitted to these services, their fate lies in the hands of the developers. In an ideal world, we could design features to detect and mitigate bias, but currently, such features do not exist. This means we can only hope that these services will eventually incorporate ways to detect bias themselves. Perhaps in the future, new startups will emerge solely dedicated to detecting and potentially mitigating bias in NLP. Until then, it is of utmost importance for all of us to keep this in mind, considering that our society is becoming increasingly automated and heavily dependent on Artificial Intelligence (AI) solutions, including chatGPT, Google BARD, and other similar technologies.

Numerous AI systems are often portrayed as more advanced than they truly are. Terms like machine learning, deep learning, and neural networks are frequently associated with elevated levels of intelligence in these software programs or devices, although this is not always accurate. It is essential not to hesitate when it comes to inquiring about the inner workings of AI systems and comprehending how they arrive at their conclusions. These technologies should actively pursue greater transparency, with the potential biases they may exhibit being recognized. Moreover, it is of utmost importance to take tangible measures in addressing these biases by actively collaborating with diverse groups in the field of technolgy.

0.1 Problematic

Training a computer to understand our world is akin to training a child. We impart upon them what we consider important, our values, desired behavior, and emotions. However, if any of these aspects are flawed or biased, as they often are, their development will be hindered. Consequently, in order to address bias in machine learning, we must identify the biases present in our algorithms and examine their underlying causes to initiate their removal. Understanding the existence of biases allows us to refrain from reinforcing them, while the absence of this knowledge hampers our ability to comprehend why our models may be malfunctioning.

Machine learning systems possess the capability to perform a wide range of tasks, from autonomous driving to disease prediction and aiding accurate medical diagnoses. Despite the promising potential, recent research has revealed a high susceptibility to bias in machine learning algorithms across various real-world applications. Examples of such biases include facial recognition software's failure to identify darker-skinned faces, prediction models overlooking responses from individuals with lower incomes, and translation systems producing nonsensical output when confronted with foreign language sentences containing gender pronouns. These unfavorable outcomes stem from inadvertent flaws embedded within training datasets. Biased systems not only compromise the quality of the products but can also lead users or decision-makers to erroneous conclusions. To date, existing approaches to address machine learning bias have primarily focused on post-deployment detection through visual inspections. Detection is limited to specific scenarios where human intuition can identify potential issues, which is not always feasible given the complexity of the algorithms. Moreover, mitigating biases often requires extensive refactoring efforts, which may not always be practical.

Bias in the fields of NLP and machine learning (ML) is a growing concern. This is because these systems are often trained on data that reflects the biases of the real world. For example, if a dataset contains more text from male speakers than female speakers, then an NLP system trained on this data will be more likely to associate male pronouns with occupations and hobbies that are typically associated with men.

This can lead to a number of problems, including:

- Gender discrimination: NLP systems that are biased against women may be less likely to recommend women for jobs or promotions (Dastin, 2018).
- Sexism in online advertising: Online advertisers may use NLP systems to target women with ads for products or services that are typically marketed to women. This can reinforce gender stereotypes and make it difficult for women to find products or services that are tailored to their needs. Overcoming these issues necessitates a comprehensive understanding of the origins of these biases (McKenzie et al., 2018).

Overcoming these issues necessitates a comprehensive understanding of the origins of these biases. Such insights can inform future experiments and guide choices in data collection to ensure equal treatment in all applications of NLP.

0.2 Definition of Bias

Bias is a complex concept with multiple overlapping definitions (Campolo et al., 2017). It has long been recognized as an inherent aspect of human decision-making (Kahneman & Tversky, 1973). When we employ cognitive bias, we rely on preconceived notions that may or may not accurately reflect reality, using them to form judgments (Garrido-Muñoz et al., 2021). According to the Open Education Sociology Dictionary[1], bias refers to an unjust prejudice in favor of or against a person, group, or thing.

Machine learning bias can manifest in various forms, encompassing racial and gender discrimination, as well as age discrimination. Research studies have highlighted the presence of biases in machine learning algorithms, suggesting that these biases are inherited from the data they are trained on (Buolamwini & Gebru, 2018a; Obermeyer & Mullainathan, 2019). This poses a significant challenge in addressing and mitigating bias in machine learning systems. Despite efforts to mitigate bias, studies have shown that even state-of-the-art models can still exhibit biased behavior (Zhao et al., 2017; Sap et al., 2019). Consequently, machine learning systems can perpetuate and amplify existing societal biases (Barocas & Selbst, 2016).

[1] Open Education Sociology Dictionary: `https://sociologydictionary.org/bias/`

0.3 Promoting Language Diversity through NLP

NLP is a vital research field that focuses on enabling computers to understand and process human language. Its applications span across various industries, including healthcare, finance, and education. In recent years, there has been a growing recognition of the importance of NLP in promoting language diversity, particularly in preserving and revitalizing endangered Indigenous languages. Indigenous languages are spoken by native people of specific regions, serving as an integral part of their culture and identity.

The loss of linguistic diversity not only erodes cultural heritage but also impacts the knowledge systems and social structures of Indigenous communities. NLP can play a crucial role in preserving and promoting Indigenous languages. One avenue is through the development of machine learning models that accurately process and understand these languages. This can facilitate the creation of digital tools for language learning and documentation, such as automated speech recognition systems, language translation, and text-to-speech synthesis. For example, (Ranta & Goutte, 2021) developed a machine learning model for identifying the Corsican language, a lesser-studied language. This model could be used to improve the accuracy of language identification systems for other endangered languages.

Another aspect where NLP can contribute to language diversity is by developing language technologies that support the use of Indigenous languages in digital communication. For example, (Mager et al., 2018) discuss the challenges of developing language technologies for indigenous languages of the Americas. They discuss the lack of data, the complex grammatical structures, and the cultural factors that make it difficult to develop effective NLP systems for these languages. However, there are some initiatives underway to increase the visibility of these languages in the digital world. For example, The No Language Left Behind (NLLB) project which is a research effort by Meta AI to develop high-quality machine translation capabilities for most of the world's languages. The project has made significant progress, it has open-sourced models capable of delivering evaluated, high-quality translations directly between 200 languages. This includes low-resource languages like Asturian, Luganda, Urdu, and more (Costa-jussà et al., 2022).

The NLLB [2] project has the potential to have a significant impact on the lives of people around the world. By making it possible to translate between 200 languages, the project will help to break down language barriers and make information more accessible to people who speak different languages. This could have a positive impact on education, healthcare, and other important areas.

Additionally, NLP can be employed to build applications supporting the use of Indigenous languages in education, healthcare, and other sectors. (Cruz & Waring, 2019) discuss the use of NLP technologies to support the documentation, revitalization, and preservation of endangered languages. They provide a case study of the use of NLP to document the K'iche' language of Guatemala. (Zhang et al., 2022) present a case study of the use of NLP to revitalize the Cherokee language. They discuss the challenges of developing NLP systems for Cherokee, and they propose a roadmap for future research in this area.

Despite the initiatives to utilize NLP in promoting language diversity, several challenges remain. A significant obstacle is the lack of data available for Indigenous languages, which hampers the development of machine learning models. Furthermore, many Indigenous languages possess complex grammatical structures, posing challenges for accurate modeling using existing NLP techniques.

[2] The No Language Left Behind (NLLB) project `https://ai.meta.com/research/no-language-left-behind`

Cognitive biases pose significant challenges as they are deeply intertwined with historical factors like colonialism and language domination. The recognition and removal of these biases are essential for conducting accurate and respectful research. As highlighted by (Datta, 2018) the process of decolonizing research involves acknowledging and addressing the inherent biases that have emerged from colonial practices and power dynamics. By actively engaging in this decolonization process, researchers can strive to ensure that their methodologies and interpretations are more inclusive, just, and reflective of the unique perspectives of indigenous communities.

Furthermore, (Zavala, 2013) delves into the complexities of decolonizing research strategies within indigenous contexts. (Zavala, 2013) recognizes that the biases deeply rooted in dominant research paradigms can be particularly difficult to identify and eliminate when studying indigenous realities. The domination of languages rich in resources has also contributed to the marginalization of indigenous knowledge systems. Thus, embracing a decolonizing approach becomes crucial in reshaping research practices and fostering a more equitable collaboration between researchers and indigenous communities. By doing so, researchers can build meaningful connections that empower indigenous voices and facilitate the co-creation of knowledge, contributing to a more balanced representation of indigenous reality in academic discourse.

Despite these challenges, the importance of promoting language diversity through NLP cannot be overstated. With the continued disappearance of Indigenous languages, there is an urgent need for innovative solutions to preserve and revitalize these languages. NLP can play a significant role in this endeavour, contributing to the preservation of linguistic and cultural diversity. Efforts should be made to address the challenges, expand the availability of data, and advance NLP techniques to ensure the success of language revitalization and preservation initiatives.

0.4 Types of Bias

Understanding and addressing bias in machine learning models is essential due to the various forms it can take, including gender, racial, religious biases, as well as unfair recruiting practices and age discrimination. To tackle these issues effectively, it is important to comprehend the different types of bias inherent in machine learning algorithms.

Research (Shashkina, 2022) identifies several prevalent types of machine learning bias, which include reporting bias, selection bias, group attribution bias, and implicit bias. Reporting bias occurs when the frequency of occurrences in the training dataset does not accurately represent real-world proportions. Selection bias arises when the training data is not representative or randomly selected, leading to skewed outcomes. Group attribution bias occurs when machine learning systems generalize individual characteristics to entire groups, irrespective of the accuracy of the generalization. Implicit bias emerges when machine learning systems rely on data influenced by personal experiences, which may not be universally applicable.

One area where gender bias poses a significant challenge is in word embeddings. Biased word embeddings can have detrimental effects on models. For example, a resume filtering model based on biased word embeddings could discriminate against female candidates for positions like programmers, while excluding male candidates for roles such as hairdressers. Similarly, a question answering model that incorrectly assumes gender stereotypes, such as all doctors being male and all nurses being female, may provide erroneous answers when applied to medical reports. These instances underscore the negative impact of gender bias in machine learning systems (Bolukbasi et al., 2016).

0.5 Objectives

The main objectives of this dissertation are as follows:

- Explore and analyze gender bias in NLP.
- Investigate the unique characteristics of the Inuktitut language, an Indigenous and endangered language, including its structure, grammatical variations, and associated challenges.
- Review and evaluate the current state-of-the-art tools developed for detecting and mitigating biases in language processing.
- Apply two specific methods to quantify the extent of bias present in the Inuktitut language.
- Employ three promising methods to mitigate the bias identified within the Inuktitut corpus.
- Assess the performance and effectiveness of these mitigation methods in downstream tasks, such as machine translation.
- Compile and present a comprehensive list of the most biased words detected within the Inuktitut corpus.
- Contribute to the revitalization and preservation efforts of Inuktitut as an Indigenous language of Canada.

0.6 Contributions

Our contributions encompass three distinct aspects. Firstly, we concentrate on detecting and mitigating bias in the Inuktitut language, which is a critically endangered language spoken by Inuit people, mainly of Nunavut and other areas such as Nunavik, in Canada. The challenge lies in the scarcity of data available for training our models effectively. Furthermore, the unique complexity of Inuktitut's polysynthetic nature sets it apart from languages like English or French, presenting additional intricacies for NLP research.

The second aspect pertains to the applicability of existing bias quantification, detection, and mitigation methods. These approaches have primarily been developed and tailored for English or, to a lesser extent, European languages. It remains uncertain whether these methods can be effectively applied to other languages, including Inuktitut. This further adds to the complexity of our task.

Lastly, we focus on detecting and mitigating bias specifically in the context of Inuktitut. To the best of our knowledge, our study is the first of its kind to apply these methods to this particular language. By addressing bias in Inuktitut, we aim to contribute to a more inclusive and equitable representation of this language in NLP research.

0.7 Organization of the dissertation

The structure of the dissertation is as follows:

Chapter I: BIAS AND DISCRIMINATION IN NLP AND AI
This chapter provides an overview of bias and discrimination in NLP and AI. It highlights the challenges and implications of bias in NLP and AI systems.

Chapter II: OVERVIEW OF GENDER BIAS IN NLP
In this chapter, we present a comprehensive review of gender bias in NLP. We discuss previous studies that have examined gender bias in NLP models and explore the relationship between gender bias and Indigenous languages. This chapter provides a critical analysis of the current state of research in this area.

Chapter III: LINGUISTIC CHALLENGES IN INDIGENOUS LANGUAGES

This chapter focuses on the linguistic challenges associated with Indigenous languages, with a specific focus on Inuktitut. We delve into the unique characteristics of Inuktitut, its complex polysynthetic nature, and the implications of these linguistic challenges for NLP research.

Chapter IV: METHODOLOGY FOR DETECTING AND MITIGATING BIAS

In this chapter, we present our methodology for detecting and mitigating bias in NLP systems. We describe the specific techniques and approaches used in our study, emphasizing their applicability to the Inuktitut language. We also discuss the challenges and limitations encountered during the implementation of our methodology.

Chapter V: RESULTS

Chapter V is dedicated to presenting the results obtained from our study. We analyze and interpret the data collected during the detection and mitigation of bias in the Inuktitut language. The chapter provides insights into the effectiveness of our approaches and their impact on reducing bias in NLP systems.

Chapter VI: CONCLUSION AND FUTURE WORK

In the final chapter, we summarize our work and its contributions. We discuss the implications of our findings and reflect on the limitations and potential future directions of research in this area. Additionally, we present preliminary work and outline future research possibilities to further address bias in NLP and AI systems.

Finally, we include a showcase of our published work, highlighting any relevant publications or contributions stemming from this dissertation.

CHAPTER I

BIAS AND DISCRIMINATION IN NLP AND AI

NLP is a crucial technology with diverse applications across various industries. While NLP has revolutionized the way we interact with computers and analyze textual data, it is not immune to challenges, and one prominent issue it confronts is bias.

Bias in NLP arises from two main sources: data bias and algorithm bias. Data bias occurs when the training dataset used for the algorithm does not accurately represent the real-world population, resulting in skewed predictions. Algorithm bias, on the other hand, occurs when the algorithm makes assumptions that may not hold true in all contexts.

Since NLP algorithms learn from the data they are trained on, their accuracy is inherently tied to the quality and representativeness of that data. Numerous studies have identified biases, particularly related to race and gender, in NLP models. These biases often seep into the predictions and decision-making processes of black-box models widely used in NLP. The models tend to inherit and magnify the biases present in their training data, leading to unfair treatment and negative consequences in real-world applications.

To address these challenges, it is crucial to actively monitor and mitigate bias in NLP models. Techniques such as data augmentation, the creation of diverse datasets, and the use of fairness-aware learning algorithms can help alleviate bias and promote fairness in NLP. By taking these steps, we can ensure that NLP technology is used ethically and effectively across industries, resulting in positive societal impacts.

1.1 Background on NLP

NLP is a critical subfield of AI that focuses on the interaction between humans and computers using natural language. Over the years, NLP has made significant progress, resulting in the development of advanced models capable of understanding and generating human-like language. However, despite these advancements, NLP still grapples with issues related to bias and discrimination.

Bias and discrimination in NLP refer to systematic errors and unfair treatment of individuals or groups based on their gender, race, ethnicity, or other personal characteristics. These issues persist in NLP due to the reliance on human-generated data for training AI models.

The datasets used to train NLP models often reflect societal biases and discrimination, causing the models to inherit and amplify these biases. (Buolamwini & Gebru, 2018b) found that large language models, such as BERT and RoBERTa, are biased against women and people of color. For example, these models were more likely to associate words like "doctor" and "engineer" with men, and words like "nurse" and "teacher" with women. Consequently, such biases can lead to harmful consequences in real-world applications (Bryson, 2018).

One major source of bias in NLP is the training dataset itself. If the dataset lacks diversity or fails to represent various groups adequately, the resulting model can exhibit bias towards certain characteristics or groups. For example, if an NLP model is trained predominantly on data from a specific demographic, such as white males, it may struggle to perform accurately on data from other demographics. This bias can have serious implications in practical scenarios like hiring processes or legal decision-making.

Additionally, biases can also originate from the algorithms used in NLP (Michael Kearns et al., 2022). Algorithms often incorporate built-in assumptions and biases, which can result in unfair treatment of individuals or groups. For instance, an algorithm trained to identify names may exhibit bias towards certain names based on their frequency in the training data. Consequently, individuals may be incorrectly identified or excluded based on their name, leading to discriminatory outcomes.

The impact of bias and discrimination in NLP can be severe, perpetuating exclusion and harm to individuals and groups. Thus, it is crucial to address these issues to ensure the ethical and effective use of NLP technology. Researchers have devised various techniques to mitigate bias and discrimination in NLP models. These techniques include data augmentation, which aims to diversify the training data, diverse dataset creation to incorporate a broad range of perspectives, and fairness-aware learning algorithms that explicitly consider and mitigate biases during model training.

By employing these techniques, NLP models can achieve improved performance and fairness in real-world applications, ultimately benefiting society as a whole. However, it is essential to recognize that addressing bias and discrimination in NLP is an ongoing process that requires ongoing research, interdisciplinary collaboration, and a commitment to promoting fairness and equity in AI systems.

1.2 Problem of Gender Bias in NLP Models

Bias in NLP systems often goes unnoticed and is typically not detected until after the systems are launched and used by consumers. However, it can have adverse effects on our society, leading people to believe false information about society or themselves and potentially changing their behavior for better or worse (Stanczak & Augenstein, 2021).

This research specifically focuses on the study of gender bias in NLP systems, which can cause representational harm by treating individuals from different categories unfairly. For example, studies have shown that negative selections occur more frequently in male-dominated jobs compared to other types of jobs (Davison & Burke, 2000). Similar findings have been reported in competency assessment and performance evaluations, where women were rated less positively than men in male gender-typed line jobs, but not in staff jobs, as observed in a prominent financial services organization (Lyness & Heilman, 2006).

By examining common examples of bias in the workplace, we can begin to grasp how it can harm individuals when such biases are propagated through downstream NLP tasks. It not only undermines inclusivity but also diminishes productivity in the work environment. While biases are present in each one of us, it is crucial to recognize their impact on our lives and the lives of others.

Recent research in NLP has revealed that word embeddings, a popular technique, can incorporate social and implicit biases inherent in the training data (Swinger et al., 2019; Schlender & Spanakis, 2020; Caliskan, 2021). Although most models used to detect bias in NLP have focused on identifying prejudice, it is important to distinguish between prejudice and bias. Prejudice involves harboring hatred or intolerance towards a group of people based on their race, ethnicity, gender identity, sexuality, or other personal characteristics (Kroon et al., 2021). While current NLP models excel at detecting prejudice, biases are not always obvious. Some biases can be detected through context, while others may be more subtle, making it challenging for automated systems to identify them.

In fact, detecting and mitigating bias within automated systems is more complex than detecting bias in human beings. This is because automated systems do not have the same ability to express nuanced sentiments as humans do.

Our efforts are based on the assumption that observed gender bias in systems indicates a lack of sufficient interest in detecting and mitigating bias. We also believe that by separating genders and professions in word embeddings, systems can be better equipped to detect and mitigate gender bias rather than perpetuating it.

Gender bias represents a significant challenge in NLP models, impacting their accuracy and fairness and potentially leading to discriminatory outcomes. Researchers have identified two main types of gender bias prevalent in NLP models: direct and indirect bias.

Direct gender bias occurs when an NLP model explicitly associates certain genders with specific traits or actions. For instance, a model that consistently assigns masculine pronouns to high-status professions like "doctor" or "engineer" and feminine pronouns to low-status professions like "nurse" or "secretary" exhibits direct gender bias. This type of bias reinforces stereotypes and reinforces unequal gender representations.

Indirect gender bias arises when an NLP model is trained on data that reflects societal biases, resulting in biased predictions. For example, if historical hiring data shows a disproportionate representation of men in certain professions, the model may learn to favor male candidates when making predictions, even if qualifications are similar between male and female candidates. This indirect bias perpetuates existing gender disparities.

To mitigate gender bias in NLP models, various approaches have been proposed. One strategy is to train models on more diverse and balanced datasets that encompass a wide range of gender identities and roles. Another approach involves using gender-neutral language and avoiding gender-specific assumptions in model design and evaluation. Additionally, researchers have explored debiasing techniques that aim to explicitly remove or reduce bias from NLP models during training or post-training stages.

Multiple studies have shed light on the issue of gender bias in NLP models. For example, a study conducted by (Bolukbasi et al., 2016) revealed gender bias in word embeddings, where certain gender-associated words were found to be associated with different levels of status and occupation. Another study by (Caliskan et al., 2017) demonstrated that language models trained on large corpora can acquire and reproduce gender biases present in the training data, influencing text generation.

Addressing gender bias in NLP models requires collaborative efforts among researchers, practitioners, and policymakers. It necessitates ongoing research, the development of comprehensive evaluation metrics, and the integration of fairness considerations into the design and deployment of NLP systems. By mitigating gender bias in NLP models, we can strive for more inclusive, equitable, and ethical applications of NLP technology.

1.3 Conclusion: Purpose of Study

In conclusion, this study emphasizes the significance of NLP in promoting language diversity, specifically focusing on Indigenous languages. The ability to process, analyze, and generate text in multiple languages is essential for preserving and promoting linguistic diversity, particularly for endangered languages. NLP techniques, including machine translation, text-to-speech synthesis, and language modeling, hold the potential to facilitate communication, education, and cultural exchange across linguistic barriers.

Furthermore, the integration of NLP in language revitalization endeavors can greatly enhance the effectiveness of language preservation and revitalization programs. By providing tools for language documentation, analysis, and translation, NLP can contribute to the preservation of linguistic and cultural heritage among Indigenous communities.

However, there are existing challenges in the field of NLP for Indigenous languages, such as limited availability of data, lack of language resources, and the necessity for culturally sensitive approaches to language processing. Overcoming these challenges necessitates interdisciplinary research and collaboration. The importance of collaboration between diverse fields is necessary because researchers can create tailored solutions and ensures that the models and resources developed are respectful of cultural differences and contribute to the preservation and revitalization of these languages.

To conclude, this study underscores the potential of NLP in promoting linguistic diversity and facilitating cross-cultural communication, particularly in the context of Indigenous languages. By addressing the challenges associated with NLP for Indigenous languages, we can support language revitalization efforts and preserve the rich cultural heritage of Indigenous communities. Continued research and collaboration are crucial for advancing NLP techniques and promoting language diversity on a global scale.

CHAPTER II

RELATED WORK

2.1 Gender Bias in NLP and its Impact on Endangered Languages

Gender bias is a significant and pressing issue in NLP systems. Research has demonstrated that NLP models often reflect and perpetuate gender stereotypes, leading to biased outcomes in various applications. This bias becomes even more critical when considering endangered languages, which face unique challenges in computational processing and are spoken by marginalized communities subjected to discrimination and inequality.

The sources of gender bias in NLP are multifaceted. Firstly, biased training data contributes to the perpetuation of gender stereotypes in NLP models. For instance, certain professions tend to be associated with a specific gender in training data, leading to models that reinforce these associations, even when they lack substantial evidence. Additionally, word embeddings, widely used in NLP, can also reflect gender biases learned from biased training data. Occupations like "programmer" or "engineer" may be more closely linked to words like "man" or "he," while occupations like "nurse" or "teacher" may be more closely associated with words like "woman" or "she."

Addressing gender bias in NLP requires a multi-faceted approach. Debiasing techniques, such as reweighting training data, modifying word embeddings, and employing gender-neutral language, show promise in mitigating bias. Moreover, diversifying and expanding training data to be more representative and inclusive can help combat bias. Developing algorithms that are less likely to perpetuate gender stereotypes is another avenue of research.

While previous studies on gender bias in NLP have primarily focused on European languages, recent research has started to explore gender bias in endangered languages like Inuktitut, Maori, and Quechua. These studies have revealed that gender bias can manifest differently in these languages, with gendered pronouns, gender-specific vocabulary, and gendered verb forms playing a role. For example, in Inuktitut, spoken by Inuit communities in northern Canada, gender bias exists with certain words and phrases being strongly associated with one gender.

Mitigating gender bias in endangered languages necessitates tailored approaches. Techniques such as debiasing, dataset augmentation, and inclusive training data have shown potential in reducing gender bias in European languages, but their efficacy in endangered languages remains largely unexplored.

In conclusion, gender bias is a prominent issue in NLP, and its impact is amplified in endangered languages spoken by marginalized communities. Researchers have started investigating gender bias in these languages and proposed various approaches to mitigate it. However, further research is needed to fully comprehend the extent of gender bias in endangered languages and develop effective techniques to minimize its influence. By addressing gender bias in NLP, we can promote fairness, inclusivity, and respect for linguistic diversity in endangered language preservation and revitalization efforts.

2.2 Studies on Indigenous languages in different parts of the world

Studies on Indigenous languages have received attention in recent years, focusing on different regions and languages. Many studies have examined ways and tools for promoting indigenous languages, shedding light on the potential impacts on linguistic preservation and cultural identity.

This section aims to provide an overview of previous studies that have investigated Indigenous languages, with a specific focus on Australia and New Zealand, South America, and Africa. Each sub-section highlights key findings from relevant published papers to shed light on the complexities of these languages.

2.2.1 Australia and New Zealand

One significant study on Australian Indigenous languages is the work by (Hobson et al., 2018). The author of "Re-awakening Languages: Theory and Practice in the Revitalisation of Australia's Indigenous Languages" wrote a book that offers a comprehensive overview of the challenges and opportunities facing the revitalization of Indigenous languages in Australia. The book begins by providing an overview of the history of language loss in Australia, and then goes on to discuss the theoretical and practical challenges of language revitalization. The book also includes several case studies that highlight the successes and failures of language revitalization efforts in Australia.

Moreover, studies have explored the cultural significance of Australian Indigenous languages and their role in shaping traditional knowledge systems. (Simpson, 2019) discusses the state of Australia's Indigenous languages, which are facing a number of challenges, including declining numbers of speakers, a lack of government support, and the influence of English. The paper proposes several ways to help promote Australia’s Indigenous languages.

Additionally, (Bowern, 2014) conducted extensive fieldwork and analysis of various Indigenous languages, focusing on the Pama-Nyungan language family. Her findings contributed to a deeper understanding of the genetic relationships between languages within this family, helping to reconstruct their ancient phonological and morphological patterns.

In New Zealand, the Indigenous language of the Māori people, te reo Māori, has also been a subject of extensive study and revitalization efforts. The investigation of the Māori language has not only focused on its linguistic aspects but also on its revitalization in contemporary society.

One influential research project on te reo Māori was carried out by (Moorfield, 2005). They made significant contributions to the documentation and preservation of the Māori language. Their comprehensive dictionary and grammar guide provided learners and researchers with valuable linguistic resources, aiding in the maintenance and revitalization of te reo Māori.

2.2.2 South America

In South America, where a diverse array of indigenous languages have thrived for millennia, language revitalization efforts have become increasingly crucial in recent decades. One remarkable example of such efforts is seen in (Lagos et al., 2017). The authors explore the challenges and opportunities of language revitalization in the context of Pehuenche communities in Chile. They argue that language revitalization is not simply a matter of teaching people a language but rather a complex process that involves negotiating competing ideologies about language, culture, and identity. They draw on the concept of "language ideology" to analyze the different ways in which people in the Pehuenche communities think about and use the Mapuche language. The authors argue that these different language ideologies can create tensions and conflicts, but they can also be a source of strength and resilience in the struggle to revitalize the Mapuche language.

Additionally, (Moore & Galucio, 2016) discusses the current state of language documentation in Brazil and the challenges and opportunities that lie ahead. The authors argue that Brazil has a unique opportunity to document its rich linguistic diversity, but that this opportunity is under threat due to a number of factors, including the ongoing decline of indigenous languages, the lack of resources for language documentation, and the increasing influence of globalization. The authors conclude by calling for a renewed commitment to language documentation in Brazil, arguing that it is essential for the preservation of indigenous cultures and languages.

Another study is the one conducted by (Durazzo, 2022), which explores the Tuxá people's efforts to revitalize their ancestral language, Dzubukuá. The Tuxá are an Indigenous people who live in the Brazilian state of Bahia. They have a strong connection to their land, which they believe is inhabited by other-than-human entities known as encantados. The Tuxá believe that Dzubukuá is a sacred language that is connected to the encantados. As a result, they are committed to revitalizing the language to maintain their connection to their culture and traditions.

2.2.3 Other Regions

In other regions, such as Africa and Asia.

(ENNAJI, 2023) examines the efforts to revitalize the Amazigh language in North Africa. The author argues that the Amazigh language has been marginalized for centuries, and that the current revitalization efforts are a necessary step to preserve the language and culture of the Amazigh people.

(Egwuonwu-Chimezie, 2023) explores the role of indigenous languages in sustainable development in Nigeria. The paper argues that indigenous languages can contribute to sustainable development by promoting cultural identity, preserving traditional knowledge, and facilitating communication and social cohesion. The paper also discusses the challenges to indigenous language revitalization in Nigeria, such as the dominance of English as the language of education and government. The paper concludes by calling for greater investment in indigenous language revitalization in Nigeria.

(Abd Ghani, 2015) discusses the challenges and successes of revitalizing the Semai language, spoken by an indigenous people of Malaysia. The paper argues that the Semai language is facing many threats, including the dominance of the Malay language, the lack of educational resources in Semai, and the assimilation of the Semai people into mainstream Malaysian society. However, the paper also highlights some positive initiatives that have been taken to revitalize the Semai language, such as the development of a Semai-Malay dictionary and the introduction of Semai language classes in schools. The paper concludes by calling for more support for the revitalization of the Semai language, so that it can continue to be a vibrant part of Malaysian culture.

In summary, previous studies have shown that studying Indigenous languages is complex and multifaceted and differs from studying European languages. These findings highlight the importance of taking into account the particular characteristics of Indigenous languages when analyzing and mitigating gender bias in NLP models.

2.3 Previous Studies on Gender Bias in NLP Models

Interest in understanding, assessing, and reducing gender bias continues to grow in the NLP field, with recent studies showing how gender disparities affect 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 bias 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 nurse 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. The proposed methodology uses words that should be neutral to gender, such as "*engineer*" and "*nurse*", and thus reduces the bias by subtracting the vector components of gender directions from gender-neutral words. Gender directions are defined by the first principal component of a word vector of each word consisting of a gender definition word pairs, such as "*he*" and "*she*".

(Liang et al., 2020b) proposed a modified method that relies heavily on the sentences used to reduce bias. We hypothesize that because English uses the common pronouns he and she extensively, which are not used in Inuktitut, because it does not use pronouns; 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. In both scenarios it showed great results.

(Sun et al., 2019) provided a survey about how to recognize and mitigate gender bias in NLP with multiple gender debiasing methods.

(Savoldi et al., 2021) presented a study of gender bias in MT and proposed a unified framework about the concepts, sources, and effects of bias in MT. The authors integrated knowledge from related disciplines, which can be instrumental to guide future research and make it thrive. 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 and/or languages without pronouns.

(Vanmassenhove et al., 2018) showed the improvement of the translation quality by integrating gender information as an additional feature into NMT systems for multiple language pairs.

(Cho et al., 2019) proposed scheme for making up a test set, with 4,236 generated sentences, that evaluates the gender bias, the proposed measure which is named as translation gender bias index (TGBI), in a MT system, in Korean, a language with gender-neutral pronouns.

(Font & Costa-Jussa, 2019) applied word embedding techniques, such as hard-debias embeddings (Bolukbasi et al., 2016) and Gender Neutral Glove (Zhao et al., 2018b), to equalize gender bias in English-Spanish NMT. Their proposed system was evaluated on a test set of occupations in an English-Spanish Neural MT. They reported an increase of BLEU performance when using pretrained word embeddings and slightly better for the debiased model.

(Stanovsky et al., 2019) presented the first challenge set and evaluation protocol for gender bias analysis in MT. They used two coreference resolution datasets composed of English sentences which cast participants into non-stereotypical gender roles.

(Kocmi et al., 2020) presented the largest evidence for the phenomenon in more than 19 systems submitted in the WMT 2020 over four diverse target languages such as Czech, German, Polish and Russian. They used WinoMT to examine gender coreference and bias from English to languages with grammatical genders. They found that all systems consistently use spurious correlations in the data rather than meaningful contextual information.

(Prates et al., 2020) showed evidence that gender bias in Machine Translation tools, such as Google Translate, can be exhibited and a strong tendancy towards male defaults.

(Basta et al., 2020) used contextual information, such as the previous sentence and the speaker information, in order to mitigate gender bias in a decoder-based NMT model. They evaluated on WinoMT with +5% accuracy and in translation quality with +1 BLEU point.

(Choubey et al., 2021b) proposed gender-filtered self-training (GFST) to improve gender translation accuracy on unambiguously gendered inputs. Their approach uses a source monolingual corpus and an initial model to generate gender-specific pseudo-parallel corpora which are 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.

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 & 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.

(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.

(Ramesh et al., 2021) evaluated and quantified gender bias in Hindi-English Machine Translation. They used word embedding fairness evaluation (Badilla et al., 2020) which is a metric that measures the gender bias in word embeddings and the translation gender bias index (Cho et al., 2019) which measures the gender bias in the translations produced by an MT system to evaluate and quantify gender bias in MT systems. The results showed that there is a significant amount of gender bias in Hindi-English MT systems.

The bias study in machine learning is not only restricted to the computer science field. Interdisciplinary research can also help address this challenge across disciplines such as psychology, sociology, linguistics, cognitive science, and more (Datta, 2018).

(Reyhner, 1999) argues that indigenous language revitalization is a complex and challenging task, but it is one that is essential for preserving cultural heritage and maintaining linguistic diversity.

(Sun et al., 2019) provides a comprehensive overview of the state of the art in gender bias mitigation in NLP. The authors discuss four forms of representation bias that can lead to gender bias in NLP systems: distributional bias, stereotype bias, structural bias, and algorithm bias. They then review a variety of methods for mitigating gender bias in NLP.

(Chen et al., 2021) investigates the extent of gender bias and under-representation in NLP systems across 9 languages: Chinese, Spanish, English, Arabic, German, French, Farsi, Urdu, and Wolof. The authors find that gender bias is present in all of the languages they study, and that it is often more pronounced in languages with grammatically gendered nouns. They also find that women are underrepresented in the Wikipedia corpora for all of the languages they studied.

(Hassan, 2016) conducted a wide study on the influence that English has had on other language communities, such as Inuit communities. 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).

It is evident that the dominance of English has transferred more than just vocabulary to Indigenous communities. The overwhelming presence of English forces Indigenous speakers to abandon not only their language but also their way of thinking, and in particular their way of thinking about women. The imposition of the "*civilized*" European's values on "*the savage*" had a destructive influence on the positive and vibrant roles that Indigenous women played within their communities. Consequently, the roles of women in Indigenous communities were reduced to include only those that were subordinate to Indigenous men. Men were assigned the role of hunting, and as such, became the "*bread-winner*" 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).

These processes, as a result, not only changed the way Indigenous women perceive themselves, but also the way their communities perceive them. This is a negative perception of Indigenous women.

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*" is a foundational principle of Inuit society, and it is reflected in many aspects of their culture, including their gender roles. Inuit traditionally value the equality of gender roles, and they believe that both men and women have important contributions to make to the community.(Williamson, 2006)

The National Research Council of Canada[1] created language technologies to assist Indigenous communities in preserving their languages and extending their use. Their project aimed to work within the empowerment paradigm, where collaboration with communities and fulfillment of their goals is the heart of the project.

[1] Canadian Indigenous languages technology project: `https://nrc.canada.ca/en/research-development/`

Unfortunately, there was no research on gender biases in any of the studied Indigenous languages.

Understanding how human biases are incorporated into word embeddings can help us understand 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, the existent methods are insufficient and the bias remain hidden within the embeddings. The frequency of words is not taken into account, regardless of the gender distances, therefore biased terms can remain clustered together. Furthermore, according to (Mulsa & Spanakis, 2020) when bias approaches are applied to contextualized word embeddings, the approach needs to be changed because the embedding representation of each word varies based on the context.

(Zhu et al., 2019) explored several post-processing techniques for debiasing pre-trained language models. These techniques included adversarial training, fine-tuning, and data augmentation. Adversarial training involves training a discriminator to identify the gender of an input sentence, while the generator (i.e., the language model) is trained to produce sentences that are difficult for the discriminator to classify by hiding gender cues. Fine-tuning involves training the language model on a dataset that has been augmented with gender-balanced examples. Data augmentation involves generating new training examples by replacing certain words in existing sentences with their gender-neutral equivalents. (Zhu et al., 2019) found that all three techniques were effective at reducing gender bias in pre-trained language models, but that adversarial training was the most effective.

In 2020, (Costa-jussà & de Jorge, 2020) showed that fine-tuning an NMT model on a gender-balanced dataset can improve the accuracy of the model's translations and reduce the gender bias in the translations. The study also showed that fine-tuning on a gender-balanced dataset can improve the performance of the model on downstream tasks, such as question answering.

(Zhang et al., 2018) introduced a novel debiasing technique called Adversarial learning which involves training two models simultaneously: a predictor and an adversary. The predictor is responsible for making predictions, while the adversary is responsible for trying to predict a protected variable, such as gender or race. The goal of the predictor is to make accurate predictions while minimizing the amount of information that it reveals about the protected variable to the adversary. Adversarial learning has been shown to be effective in mitigating unwanted biases in a variety of machine learning tasks. However, it is important to note that adversarial learning is not a silver bullet. It can be difficult to implement effectively, and it can sometimes lead to a decrease in accuracy.

Other studies have also investigated the impact of various factors on the effectiveness of debiasing techniques, such as the size and quality of the training data. For example, a study by (Zhao et al., 2018a) found that debiasing techniques were more effective when trained on larger and more diverse datasets. They also found that debiasing techniques that focused on specific components of the language model, such as the word embeddings, were more effective than those that applied to the entire model.

Another study by (Schick et al., 2021) surveyed a number of debiasing techniques for language models, including counterfactual data augmentation, dropout, iterative nullspace projection, self-debias, and sentence debias. They evaluated the effectiveness of these techniques on a variety of intrinsic bias benchmarks, as well as on downstream tasks. The study found that self-debias was the most effective debiasing technique, followed by counterfactual data augmentation and iterative nullspace projection. However, they also found that debiasing techniques can sometimes have a negative impact on the language modeling ability of the model. The study concluded that more research is needed to develop more effective debiasing techniques that do not negatively impact language modeling ability.

These studies suggest that the effectiveness of debiasing techniques depends on a variety of factors, including the size and quality of the training data, the specific components of the language model that are targeted, and the downstream task. More research is needed to better understand the factors that influence the effectiveness of debiasing techniques and to develop more effective debiasing methods.

Overall, these studies demonstrate that there are several promising techniques for mitigating and debiasing gender bias in NLP models. However, further research is needed to evaluate the effectiveness of these techniques across different languages and domains, and to develop new methods that can address the unique challenges posed by non-Western languages and dialects.

These studies have contributed to our understanding of gender bias in NLP models and provided insights into how to mitigate and debias these models. However, much work remains to be done, particularly in the context of endangered languages and the unique challenges they pose for NLP processing.

2.4 Conclusion: Limitations and Gaps in Existing Literature

While previous studies have contributed significantly to our understanding of gender bias in Indigenous languages and NLP models, several limitations and gaps in the existing literature remain (Kaneko et al., 2022). In this section, we discuss some of these limitations and identify areas for future research.

One significant limitation is the scarcity of data for many Indigenous languages. This lack of data poses a challenge for accurately detecting and mitigating gender bias in these languages. Additionally, this limitation makes it difficult to develop and test NLP models for these languages, hindering progress in the field.

Another limitation is the limited scope of research on gender bias in Indigenous languages. Most studies have focused on a few languages, mainly those spoken in Europe, and North America. Thus, there is a need for research that covers a broader range of Indigenous languages and geographic regions.

Many existing techniques for detecting and mitigating gender bias in NLP models are developed for English-based languages. However, these methods may not be directly applicable to Indigenous languages with different linguistic structures and characteristics. Thus, there is a need for research that specifically addresses the unique linguistic challenges of Indigenous languages.

Finally, the existing literature on gender bias in Indigenous languages and NLP models has primarily focused on binary gender categories, overlooking other forms of intersectionality such as ethnicity, race, and sexual orientation. Future research should address these gaps by exploring the intersectionality of gender bias in Indigenous languages.

In summary, while significant progress has been made in addressing gender bias in Indigenous languages and NLP models, several limitations and gaps in the existing literature remain. Future research should aim to address these limitations by exploring new data sources, covering a broader range of Indigenous languages and regions, considering the unique linguistic characteristics of these languages, and examining the intersectionality of gender bias.

CHAPTER III

LINGUISTIC CHALLENGES IN INDIGENOUS LANGUAGES

Indigenous languages pose unique challenges that make them an ideal case study for NLP, offering insights into the workings of NLP systems. Two primary challenges associated with Indigenous languages are the scarcity of data and the existence of numerous dialects. Fortunately, recent advancements in neural networks have facilitated the processing of these languages, providing opportunities to unravel the mysteries of language.

(Le & Sadat, 2020a) investigated the use of neural machine translation to address the challenges of translating Inuktitut. They found that NMT can be effective for Inuktitut translation, but that it is important to consider the specific challenges of polysynthetic languages when training and evaluating NMT models. For example, they found that using additional features extracted from the source-target alignment information and bilingual lexicons can improve the performance of NMT models for Inuktitut

In this chapter, we examine the main linguistic challenges presented by Inuktitut, an Indigenous language spoken in Eastern Canada and the official language of the government of Nunavut. By delving into the structure of Inuktitut words, the levels of grammatical variations, dialectal differences in spelling, and gender animacy, we can gain a deeper understanding of the NLP challenges specific to

Inuktitut.

3.1 Inuktitut in the Context of NLP

Inuktitut is spoken primarily by Inuit communities in Canada. It belongs to the polysynthetic language family, wherein words are formed by combining multiple morphemes (meaningful units) into a single word. As a result, Inuktitut exhibits a highly agglutinative nature, with words that can become lengthy and intricate. Additionally, the language features a complex system of inflectional and derivational morphology, further contributing to its intricacy.

3.1.1 History and Demographics

Inuktitut, the language of the Inuit people, is one of Canada's Indigenous languages and is spoken by approximately 39,770 individuals according to the Canadian Encyclopedia [1]. It belongs to the Eskimo-Aleut language family, which is a small group of languages (Swift, 2004). The term "Inuit" [2] translates to "people" or "the people" and refers to the Indigenous populations living in Northern regions. Inuktitut has a rich oral tradition, with stories, songs, and legends passed down through generations. According to The Library and Archives Canada Blog [3] The language was first written down using the Latin script in the 19th century.

[1] The Canadian Encyclopedia: inuktitut `https://www.thecanadianencyclopedia.ca/en/article/inuktitut`

[2] The Canadian Encyclopedia: inuit `https://www.thecanadianencyclopedia.ca/en/article/inuit`

[3] The Library and Archives Canada Blog: `https://thediscoverblog.com/2015/06/11/aboriginal-syllabic-scripts/`

In the field of NLP, Inuktitut presents several challenges. One major challenge stems from the limited availability of linguistic resources such as annotated corpora, lexicons, and grammars. This scarcity of resources can be attributed, in part, to the minority status of Inuktitut, resulting in less support compared to widely spoken languages. Additionally, the complex morphology of the language poses difficulties in developing accurate language models.

Despite these challenges, recent efforts have been made to develop NLP tools for Inuktitut. The Pirurvik Centre in Iqaluit[4], Nunavut, has created freely available online spell-checker and grammar-checker tools for Inuktitut [5]. Other initiatives, such as the Nunavik Inuit Language Corpus and the Tusaalanga Interactive Language Learning Platform, aim to build linguistic resources specifically for Inuktitut.

In the following section, we will explore the specific linguistic features and also challenges faced by Inuktitut in the realm of NLP and examine the ongoing research endeavors to tackle them.

3.2 Linguistic Features of Inuktitut

3.2.1 Morphological Complexity and Morphophonemic Challenges

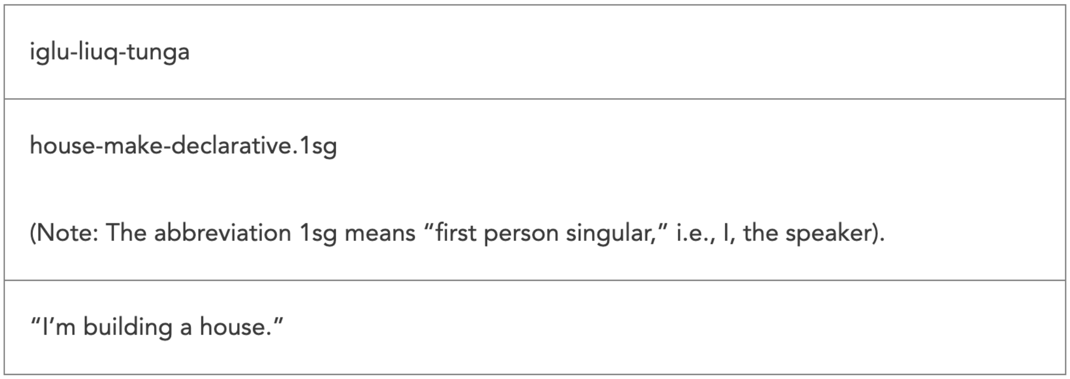
Inuktitut, like many other Indigenous languages, exhibits a high degree of morphological complexity, which presents challenges in the realm of NLP (Gasser, 2011; Littell et al., 2018). The morphology of Inuktitut is characterized by the formation of words through the combination of a word base with multiple suffixing

[4] The Pirurvik Centre in Iqaluit: `https://www.pirurvik.ca/`

[5] The Web applications for processing Inuktut text: `https://iutools.org/`

morphemes, resulting in long and intricate sentence-like words (Mithun, 2015).

From a linguistic perspective, a single word in Inuktitut can encompass what would typically constitute a full clause or sentence in other languages. This polysynthetic nature of Inuktitut allows for the expression of complex ideas within a single word (Kudlak & Compton, 2018). For instance, the word "igluliuqtunga" (as shown in Figure 3.2) carries the meaning of "I am going to the igloo to sleep" (Kudlak & Compton, 2018).



iglu-liuq-tunga
house-make-declarative.1sg (Note: The abbreviation 1sg means “first person singular,” i.e., I, the speaker).
“I’m building a house.”

Figure 3.1 An example sentence in Nunavut Inuktitut, one of the Canadian Indigenous languages, written in the Latin script with its translation (Kudlak & Compton, 2018)

The phenomenon of extensive suffixation in Inuktitut gives rise to lengthy words that can be highly unique, with each word carrying a wealth of grammatical and semantic information (Mithun, 2015). This poses significant challenges for NLP systems, as capturing the intricate morphological structure of Inuktitut words and accurately processing their meaning require specialized approaches and resources.

In the subsequent sections, we will delve into the specific challenges posed by the morphological and morphophonemic characteristics of Inuktitut and examine the ongoing research efforts aimed at addressing these challenges in the context of NLP.

Moreover, the composition of verbs and the incorporation of nouns in Indigenous languages, including Inuktitut, are significantly more complex than in European languages like English or French. Indigenous languages are highly inflected and exhibit unique structural characteristics (Dorais, 2010).

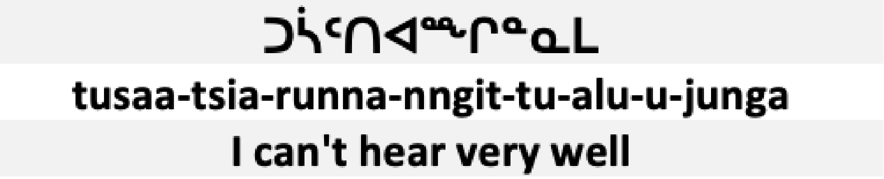
Our current research focuses on Inuktitut, a polysynthetic language belonging to the Inuit language family and spoken in Northern Canada. However, Inuktitut is considered endangered as it is primarily used for "lower" tasks, such as private conversations and non-specialized jobs, while English or French dominate in higher education and high-income professions (Dorais, 2010).

When comparing word composition in English, the structure of words in Inuit languages, including Inuktitut, can vary significantly in their surface form. Words can be relatively short, consisting of a word base, lexical suffixes, and grammatical ending suffixes, or they can be much longer, incorporating up to ten or even fifteen formative morphemes, depending on the regional dialect (Lowe, 1985; Kudlak & Compton, 2018; Le & Sadat, 2020b; Le & Sadat, 2022).

The composition of an Inuit word can be represented as: **Word base** + Lexical suffixes + *Grammatical ending suffixes.*

For example, in Nunavut Inuktitut (Figure 3.2), a sentence is formed by the word **tusaa**-tsia-runna-nngit-tu-alu-u-*junga*, which translates to "I can't hear very well" in English (Dorais, 1990; Micher, 2018).

In this example, the lengthy word is composed of a root word morpheme **tusaa** (meaning: "to hear"), followed by six lexical suffixes (**tsia, runna, nngit, tu, alu, u**) and one grammatical ending suffix (**junga**), which are separated by hy-



ᑐᓵᑦᑎᐊᖖᒋᓐᓇᒪ

tusaa-tsia-runna-nngit-tu-alu-u-junga

I can't hear very well

Figure 3.2 An example of a sentence in Nunavut Inuktitut, one of the Canadian Indigenous languages, in the Inuktitut script, its romanization, and its translation

phens. All these segmented morphemes are synthetically combined into a single unit.

Notably, in this single Inuktitut word, the information expressed is equivalent to that of two complete clauses in English (Dorais, 1990; Micher, 2018). These characteristics of Inuktitut's word composition pose significant challenges for NLP systems, requiring specialized approaches and resources to accurately process and analyze the rich morphological structure and semantics of Inuktitut words.

In Inuit family languages, a single word can encompass the meaning of an entire sentence in English. The word composition typically involves augmenting lexical constituents with multiple formative suffixing morphemes attached to a word base. It is common for full sentences to be represented by a single word (Lowe, 1985). This complex morphological structure is characterized by the use of a variety of lexical suffixes and grammatical ending suffixes, making the task of morphological segmentation for polysynthetic languages particularly challenging. The analysis of Indigenous language morphology poses several difficulties in the field of NLP, primarily due to their morphological complexity (Arppe et al., 2017).

One specific challenge is determining the basic unit of a word and identifying the subword units within it (Arppe et al., 2017). The morphophonemics of Inuktitut further contribute to its complexity, as Inuktitut roots can exhibit various morphological suffixes (Mithun, 2015). In Inuktitut, each morpheme specifies sound variations that can occur to its left and/or itself. These modifications are phonologically conditioned by the individual morphemes, rather than their contexts. This not only exacerbates the issue of data sparsity but also presents challenges in morphological analysis that will be addressed in the research topics of this project. The underlying representations of these morphemes are referred to as 'deep' morphemes in this study, contrasting with 'surface' morphemes (Micher, 2018). Understanding and effectively analyzing these complex morphophonemic patterns are crucial for developing accurate NLP models for Inuktitut and other Indigenous languages.

3.2.2 Dialectal Variations

Inuktitut exhibits considerable dialectal variations, with differences in phonology, morphology, and vocabulary across its various dialects. For instance, the Inuktitut spoken in Nunavut displays significant distinctions from other regional variants. This poses a challenge in developing language models that are accurate and effective across different dialects, necessitating the creation of dialect-specific linguistic resources.

The presence of abundant spelling variations in Inuktitut further complicates its computer processing. Within the Inuktitut language family, numerous dialects exist, including Uummarmiutun, Siglitun, Inuinnaqtun, Natsilik, Kivallirmiutun, Aivilik, North Baffin, South Baffin, Arctic Quebec, and Labrador (Dorais, 1990). These dialects primarily differ in their phonology, which is reflected in their respective spellings. The lack of standardization and the prevalence of dialectal changes contribute to spelling variance, significantly impacting the overall sparsity of data in accessible corpora for experimentation (Micher, 2018). Consequently, the polysynthetic nature of Inuktitut, combined with its morphophonemic characteristics and spelling variations, renders it a particularly challenging language for NLP (Micher, 2018).

3.2.3 Gender Animacy and Abundance of Grammatical Suffixes

Inuktitut features a complex system of noun classes that are determined by both gender and animacy. Each noun is assigned to a specific class, which has implications for plural formation, pronoun selection, and other grammatical aspects. This poses a challenge in developing language models that accurately account for these noun classes.

It is worth noting that Inuktitut does not have gender markings for nouns and verbs. Instead, it distinguishes words based on animacy, which refers to whether a noun is considered living or not. This cultural understanding categorizes nouns as animate or inanimate, without specific gender distinctions. The absence of gender markings adds further challenges when designing NLP applications and tools, particularly when translating to languages with gendered nouns, as it may result in gender stereotyping and biases (Choubey et al., 2021a).

The distinction between animate and inanimate nouns is common in many Amerindian language families, including Inuktitut. Animate nouns typically include humans and animals, while inanimate nouns encompass non-living objects such as stones or tables. This animacy distinction reflects the social activity and perceived rationality of the noun (Micher, 2018).

Inuktitut employs a large number of grammatical suffixes to indicate various meanings, including tense, mood, aspect, and negation. These suffixes can be combined in different ways to form complex words and can be attached to both verbs and nouns. The abundance of suffixes presents a challenge in developing accurate language models, particularly for tasks that require disambiguating multiple suffixes.

According to (Micher, 2018), Inuktitut exhibits inflectional morphology to express a wide range of grammatical features. This includes nine verbal moods, two sets of subject and subject-object markers per mood, four person distinctions, three numbers (singular, dual, and plural), eight noun cases, and noun possessors. Demonstratives in Inuktitut also have multiple dimensions, including location, directionality, specificity, and previous mention. These grammatical features are expressed through the use of grammatical inflection in the language (Micher, 2018).

3.2.4 Linguistic uniqueness

The linguistic characteristics of Inuktitut, with its highly agglutinative morphology and complex noun phrase system, present significant challenges for applying NLP techniques developed for more widely spoken languages. As a result, researchers are working to develop innovative methodologies specifically tailored to the unique linguistic properties of Inuktitut.

Inuktitut, predominantly spoken in northern Canada, is characterized by its extensive use of affixation and agglutinative nature. This means that complex word forms are created by adding affixes that convey additional grammatical and semantic information. However, this linguistic feature differs greatly from the morphology of more common languages, necessitating a deep understanding of Inuktitut's distinctive linguistic structure.

In addition, Inuktitut exhibits a sophisticated noun phrase system with noun classes or classifiers. Nouns are categorized into different groups based on attributes such as shape, size, and animacy. This complexity adds an extra layer of challenge in NLP tasks, as models need to account for the interaction between noun classes and other grammatical components.

To address these challenges, researchers are actively working on developing specialized approaches and tools specifically designed for Inuktitut. This involves extensive exploration of the language's morphology, syntax, and semantic nuances to construct robust computational models capable of effectively processing and analyzing Inuktitut text.

Building annotated corpora that capture the intricacies of Inuktitut's morphological and syntactic structures is a crucial part of these efforts. These corpora serve as valuable resources for training machine learning algorithms and developing language models specific to Inuktitut. Additionally, collaboration with native speakers and linguistic experts is essential to ensure the accuracy and cultural appropriateness of the developed tools.

Adapting NLP techniques to Inuktitut is an ongoing process that requires continuous refinement and innovation. By fostering interdisciplinary collaboration between computational linguists, native speakers, and experts in Inuktitut linguistics, significant progress is being made to overcome the challenges posed by the unique linguistic features of Inuktitut.

Ultimately, the application of NLP techniques in the context of Inuktitut has the potential to facilitate language processing tasks, improve communication, and preserve the rich linguistic heritage of this remarkable Arctic language. With dedicated research and development efforts, the future of NLP in Inuktitut looks promising.

3.3 Data Sparsity

Inuktitut, being a minority language, encounters significant challenges due to the scarcity of available linguistic data when compared to more widely spoken languages. This poses difficulties in developing accurate language models, especially for tasks that require a considerable amount of annotated data, such as machine translation or speech recognition.

In a notable research study conducted by (Micher, 2018), Inuktitut's polysynthetic nature and capacity to incorporate numerous morphemes within single words, coupled with unpredictable morphophonological processes and spelling variations, contribute to the sparsity of Inuktitut data. This scarcity is even more pronounced compared to other languages characterized by morphological complexity that have been studied in the field of NLP.

The lack of extensive and diverse annotated datasets impedes the training of accurate and robust models that can effectively capture the intricacies of Inuktitut's grammar, syntax, and semantics. Consequently, the performance of NLP applications in Inuktitut, such as machine translation systems or speech recognition algorithms, may be compromised due to the limited availability of training resources.

To address the data sparsity challenge in Inuktitut, collaborative endeavors involving native speakers, linguists, and computational linguists are essential for collecting, annotating, and curating more extensive and more representative corpora in Inuktitut. These efforts enable the development of language models and data-driven approaches that can overcome the limitations posed by data sparsity, thereby enhancing the accuracy and performance of NLP techniques applied to Inuktitut.

(Micher, 2018) suggest that recognizing and addressing the unique challenges associated with data sparsity in Inuktitut can contribute to the preservation and revitalization of the language, enabling its continued use in digital communication and ensuring its cultural and linguistic richness thrives in the digital age.

3.4 Conclusion

In conclusion, the linguistic characteristics of Inuktitut present unique challenges when applying NLP techniques. The highly agglutinative morphology, complex noun class system, and scarcity of linguistic data pose hurdles that require specialized methodologies and tools.

The morphology of Inuktitut, with its extensive use of affixation and intricate word forms, diverges significantly from widely spoken languages. This necessitates a deep understanding of Inuktitut's linguistic framework to develop accurate language models.

The noun class system adds another layer of complexity, categorizing nouns based on various attributes. NLP models must account for the interplay between noun classes and other grammatical components.

The scarcity of available linguistic data in Inuktitut hampers the development of accurate models. The limited availability of annotated datasets affects the training of robust models, hindering the performance of NLP applications.

Addressing these challenges requires interdisciplinary collaboration and efforts to collect, annotate, and curate larger and more representative corpora. Collaborative endeavors involving native speakers, linguists, and computational linguists are crucial for developing language models specific to Inuktitut.

By recognizing and addressing the unique challenges of Inuktitut, the field of

NLP can contribute to the preservation and revitalization of the language. This enables its continued use in digital communication and ensures the preservation of its cultural and linguistic richness.

Moving forward, continuous research and development efforts are necessary to refine and innovate NLP techniques for Inuktitut. With dedicated efforts, NLP can facilitate language processing tasks, enhance communication, and contribute to the preservation of Inuktitut's linguistic heritage.

CHAPTER IV

METHODOLOGY

Although machine learning models have achieved remarkable results on various tasks, they often fall short in addressing biases. Recent studies have highlighted the impact of bias on NLP technologies, leading to a growing interest in identifying, analyzing, and mitigating bias within the NLP community. This problem is not new, as it is well-known that NLP systems inherently contain and reflect algorithmic bias, raising significant concerns about their social impact. Given the widespread use of NLP systems and tools in everyday life, it is essential to acknowledge that our models have real-life consequences, which may not align with our intended objectives (Ehni, 2008).

Within this broader research area, it is crucial to consider the specific challenges related to bias in Indigenous languages. Indigenous languages possess a wealth of secondary data about individuals, their identities, and their demographic groups, which are often leveraged to develop NLP systems. However, the focus on creating these systems has sometimes shifted the emphasis away from creating models as tools of understanding and towards achieving optimal results, making it more challenging to comprehend the underlying processes (Hovy & Prabhumoye, 2021b).

By recognizing and addressing bias in NLP systems, particularly in the context of Indigenous languages, we can strive for more equitable and fair technologies. This requires concerted efforts from the NLP community to develop methods and tools that promote transparency, fairness, and inclusivity. By embracing a responsible and ethical approach, we can ensure that NLP systems have a positive impact on society while mitigating the potential biases that may arise.

4.1 Clustering and WEAT methods to quantify gender bias in Inuktitut

The primary objective of this research is to detect and measure potential gender bias in Inuktitut language data. To accomplish this, two main techniques will be employed: clustering and the Word Embedding Association Test (WEAT).

Clustering is a method used to group similar words or phrases based on their semantic or contextual properties. In the context of gender bias, clustering allows us to identify clusters of words strongly associated with men or women and quantify the level of association within each cluster. By applying clustering to Inuktitut data, we can identify word groups commonly linked to men or women, enabling the analysis of their distribution across various texts or contexts for potential patterns of bias.

The Word Embedding Association Test (WEAT) is a widely used evaluation metric that measures the association between different word categories within word embeddings. This method involves calculating the cosine similarity between the average vector of words in one category and the average vector of words in another category. By comparing the resulting similarity scores for different pairs of categories, we can identify words strongly associated with either men or women in Inuktitut and quantify the extent of that association. Through the comparison of association scores among different word sets, any instances of bias within the language data can be identified.

By employing these evaluation metrics in combination, a comprehensive understanding of the gender bias in Inuktitut language data can be achieved. Additionally, the effectiveness of various debiasing methods in mitigating this bias can be assessed. This approach enables researchers to gain insights into the presence of gender bias and develop strategies to address and rectify it within the Inuktitut language.

4.2 Mitigation Techniques: Hard Debias, SENT Debias, INLP

Once instances of gender bias in the Inuktitut language data have been identified, our research will explore various techniques for mitigating or eliminating this bias. Three primary techniques that we will investigate are Hard Debias, SENT Debias, and Iterative NullSpace Projection (INLP).

4.2.1 Hard Debias (Bolukbasi et al., 2016)

One of the initial strategies used to detect and minimize bias in word embeddings was *Hard Debias*. Hard Debias involves directly modifying the language data to remove any gender associations that may be present. This can be achieved by replacing gendered words with neutral alternatives or modifying sentence structures to eliminate implicit gender associations. For example, using "they" instead of "he" or "she" can contribute to gender neutrality.

(Bolukbasi et al., 2016) proposed a method for Hard Debias of word embeddings, wherein they first identify the gender subspace of the embedding space using a set of gender-defining words. They then project all words onto the orthogonal complement of the gender subspace, effectively eliminating gender associations. This method has been shown to be effective at reducing gender bias in word embeddings, although it can result in a loss of important linguistic features and context (Bolukbasi et al., 2016). Algorithm 1 outlines the procedures involved in the Hard Debias algorithm.

Algorithm 1 Hard debias (Bolukbasi et al., 2016)

```
Require: Word sets $W$, defining sets $D$, $D_1$, $D_2$, ..., $D_k$, embedding matrix $R$,
    integer parameter $k \geq 1$.
        Step 1: Identify genger subspace
Ensure: Bias subspace $B$.
Let $\mu_i$ be the mean of defining set $D_i$ for $i = 1, 2, ..., k$.
Let $C = R^T R$.
Let $B$ be the first $k$ rows of SVD($C$).
        Step 2: Neutralize and Equalize
Require: Words to neutralize $NW$, family of equality sets $E_1$, $E_2$, ..., $E_m$, where
    each $w \in NW$.
Ensure: New embedding matrix $R'$.
for each word $w \in NW$ do
    Let $a$ be the $w$th row of $R$.
    Set $a$ to the mean of defining set $D_i$ where $w \in E_i$.
end for
for each set $E_i$ do
    Let $\mu_i$ be the mean of defining set $E_i$.
    for each word $w \in E_i$ do
        Set the $w$th row of $R'$ to $\mu_i$.
    end for
end for
```

The first step in this algorithm is to identify the gender subspace. This is done by taking the word sets W, defining sets D, and the embedding matrix C. The defining sets are sets of words that are known to be associated with either the male or female gender. For example, the defining set for the male gender might be the set man, boy, he, his, him. The embedding matrix C is a matrix that represents the embeddings of all of the words in the vocabulary.

Once the word sets and the embedding matrix are defined, the gender subspace is identified by finding the first k rows of the SVD of C, where k is a parameter that is typically set to 2 or 3. The SVD of a matrix is a factorization of the matrix into three matrices: U, S, and V. The U matrix contains the left singular vectors of the matrix, the S matrix contains the singular values of the matrix, and the V matrix contains the right singular vectors of the matrix.

The gender subspace is the first k rows of the U matrix. This is because the left singular vectors of a matrix represent the directions in which the matrix has the most variance. In the case of the word embedding matrix, the first k rows of the U matrix will represent the directions in which the word embeddings have the most variance with respect to gender.

The second step is to hard debias the embeddings. This is done by subtracting the mean of the defining sets from the embeddings of the words in the defining sets. For example, if the mean of the defining set for the male gender is 0.5, then the embeddings of the words in the defining set will be subtracted by 0.5. This step effectively shifts the embeddings of the words in the defining sets to be more neutral. This means that the embeddings of the words in the defining sets will no longer be biased towards either the male or female gender.

The Hard Debias Algorithm has been evaluated on a number of datasets, and it has been shown to be effective in reducing gender bias in word embeddings. For example, in the study by (Bolukbasi et al., 2016), the Hard Debias Algorithm was able to reduce the gender bias in word embeddings by up to 80

However, the Hard Debias Algorithm is not perfect. It is still possible for the algorithm to introduce some error into the embeddings. For example, the algorithm might shift the embeddings of some words too far in the neutral direction, which could result in a loss of meaning.

Overall, the Hard Debias Algorithm is a promising approach for debiasing word embeddings. It is effective in reducing gender bias, and it is relatively simple to implement.

4.2.2 SENT Debias (Liang et al., 2020a)

(Liang et al., 2020a) addresses the problem of bias in sentence representations. The authors propose a method called Sent-Debias, which can be used to reduce the bias in sentence representations while still preserving their performance on downstream tasks.

Sent-Debias works by first identifying a set of words that are associated with certain biases. These words are then used to create bias attribute sentences, which are then used to train a linear regression model. The linear regression model is used to estimate the projection of a sentence representation onto the bias subspace. This projection is then removed from the sentence representation, which reduces the bias in the representation (Liang et al., 2020a).

(Liang et al., 2020a) evaluated Sent-Debias on a number of downstream tasks, including sentiment analysis, linguistic acceptability, and natural language understanding. They found that Sent-Debias was able to reduce the bias in sentence representations without significantly impacting the performance on these tasks.

Algorithm 2 provides an overview of the processes involved in the SENT Debias algorithm.

Algorithm 2 SENT-DEBIAS (Liang et al., 2020a)

```
1: Initialize (usually pretrained) sentence encoder $M_o$.
2: Define bias attributes (e.g. binary gender $g$ and $m$).
3: Obtain words $D = \{w_1, w_2, ..., w_n\}$ indicative of bias attributes.
4: For each word $w_i \in D$, contextualize it to obtain a sentence representation $s_i = M_o(w_i)$.
5: Compute the bias subspace $V$ by PCA on the set of sentence representations $\{s_1, s_2, ..., s_n\}$.
6: For each new sentence representation $h$, project it onto the bias subspace to obtain the projection $h' = V \cdot h$.
7: Subtract the projection $h'$ from the original sentence representation to obtain the debiased sentence representation $h - h'$.
```

The SENT Debias algorithm works by first identifying a set of words that are indicative of bias attributes. For example, the words "he" and "she" could be indicative of the bias attribute "gender". The algorithm then constructs a sentence representation for each word in the set of indicative words. This is done by using a pretrained sentence encoder.

The next step is to compute the bias subspace. This is done by taking the average of the sentence representations for the indicative words. The bias subspace is a vector space that captures the bias of the training data.

Finally, the algorithm projects the sentence representations of new sentences onto the bias subspace. This is done by finding the projection of each sentence representation onto the bias subspace. The projection of a sentence representation is a vector that represents the bias of the sentence. The algorithm then subtracts the projection from the sentence representation. This removes the bias from the sentence representation.

The algorithm has been evaluated on a number of datasets. It has been shown to be effective at removing bias from NLP models(Liang et al., 2020b).

The algorithm has also been analyzed to understand its error sources. One of the main error sources is the choice of indicative words. If the set of indicative words does not capture all of the bias in the training data, then the algorithm will not be able to remove all of the bias from the NLP model. Another error source is the choice of the pretrained sentence encoder. If the pretrained sentence encoder is not robust to bias, then the algorithm will not be able to remove all of the bias from the NLP model.

Overall, SENT-DEBIAS is a promising debiasing algorithm for NLP models. It has been shown to be effective at removing bias from NLP models. However, the algorithm has some limitations, such as the choice of indicative words and the choice of the pretrained sentence encoder.

SENT-DEBIAS is a powerful tool for removing bias from NLP models. However, it is important to be aware of the limitations of the algorithm. The choice of indicative words and the choice of the pretrained sentence encoder can have a significant impact on the effectiveness of the algorithm.

4.2.3 Iterative NullSpace Projection (INLP) (Ravfogel et al., 2020)

INLP, which stands for Iterative NullSpace Projection, is a method for eliminating gender bias from word embeddings (Algorithm 3). This technique involves iteratively projecting the embedding space onto the null space of the gender subspace. By doing so, INLP minimizes gender associations in the word embeddings while preserving important linguistic features and context.

(Ravfogel et al., 2020) proposed a method for INLP that iteratively projects the embedding space onto the null space of the gender subspace. They also minimize the reconstruction error of the projected space. The evaluation of their method on several benchmark datasets demonstrated its effectiveness in reducing gender bias while preserving linguistic features (Ravfogel et al., 2020). In a study by (Ravfogel et al., 2020), INLP was able to reduce gender bias by up to 90 on several benchmark datasets.

Algorithm 3 Iterative null-space projection (Ravfogel et al., 2020)

```
Require: $(X, Z)$: a training set of vectors and protected attributes
Require: $n$: Number of rounds
Ensure: A projection matrix $P$
$Xprojected \leftarrow X$
$P \leftarrow I$
for $i = 1$ to $n$ do
    $W \leftarrow TrainClassifier(Xprojected, Z)$
    $B \leftarrow GetNullSpaceBasis(W)$
    $PN(W) \leftarrow B \cdot B^T$
    $P \leftarrow P + PN(W)$
    $Xprojected \leftarrow Xprojected - PN(W) \cdot Xprojected$
end for
return $P$
```

Algorithm 3 summarizes the INLP algorithm which takes as input the embedding matrix and the gender subspace. It then projects the embedding matrix onto the null space of the gender subspace and uses the projected embedding matrix to train a new model. In the first step, the gender subspace is computed. This can be done by finding the principal components of the matrix that represents the gender associations between words. Then, the embedding space is projected onto the null space of the gender subspace. This is done by multiplying the embedding matrix by the projection matrix, which is the matrix that spans the null space of the gender subspace. The projected embedding space is then used to train a new model. This model is less likely to be biased than the original model, because it is not trained on the gender subspace.

4.3 Overcoming sparsity due to complexity

As discussed in Chapter III, working with Inuktitut language data presents a challenge due to its relatively sparse nature compared to more widely spoken languages. This scarcity of data poses difficulties in developing accurate language models, especially for tasks that rely on large amounts of annotated data.

To overcome the challenge of data sparsity in Inuktitut, our research will explore various techniques and approaches. Firstly, we will investigate the development of more efficient algorithms specifically designed to handle sparse data. These algorithms can effectively process and analyze the available data, even in cases where the data points are limited.

Additionally, we will explore the integration of additional sources of linguistic information to supplement the existing data. This can involve leveraging resources such as lexicons, dictionaries, or other language references to enrich the dataset and provide additional context. By incorporating this supplementary information, we can enhance the coverage and accuracy of language models trained on sparse data.

Another technique we will investigate is transfer learning. Transfer learning involves utilizing pre-trained language models from other languages and adapting them to the Inuktitut language. By leveraging the knowledge and representations learned from larger datasets in other languages, we can enhance the performance of language models in Inuktitut, even with limited annotated data.

By exploring these techniques and approaches, we aim to address the challenge of data sparsity in Inuktitut and develop more accurate and robust language models. These efforts will contribute to improving the effectiveness of NLP tasks in Inuktitut, such as machine translation or speech recognition, and facilitate the broader adoption and application of language technologies in this unique linguistic context.

4.4 Morphological Analyzer and Word Embeddings

In our study of Inuktitut, we will employ various tools and techniques to gain deeper insights into the language's linguistic features and effectively analyze the available data. One of those techniques is the one investigated by (Le & Sadat, 2020b). The authors developed a morphological segmenter for Inuktitut, which helped to improve the accuracy of NMT models. They then trained an NMT model on a parallel corpus of Inuktitut and English sentences. The model was able to achieve a significant improvement over previous results.

We believe that these tools will help us to better understand the linguistic features of the language, and process Inuktitut text more accurately.

4.4.1 Morphological Analyzer

To facilitate the analysis of Inuktitut, we will utilize a morphological analyzer proposed by (Le & Sadat, 2020c), which is a tool specifically designed to deconstruct words into their constituent morphemes. In the case of Inuktitut, a language known for its complex system of affixes, a morphological analyzer proves particularly valuable. By breaking down words into their meaningful units, we can identify and examine the different morphological forms they possess. This analysis enables us to gain a comprehensive understanding of the language's morphology, including the intricate relationships between affixes and the underlying meaning they convey.

(Le & Sadat, 2020c) proposed an Inuktitut Neural Network-based(NN) word segmenter. This method uses a set of rich features, including part-of-speech tags, morphological tags, and dependency relations, to segment Inuktitut words into their constituent morphemes.

The morphological analyzer will assist us in exploring the rich morphological structure of Inuktitut words, uncovering patterns, and identifying variations that may exist within the language. This in-depth analysis of the language's morphology serves as a foundation for subsequent investigations into gender bias and other linguistic phenomena.

4.4.2 Word Embeddings

Word embeddings are powerful representations that encode words as high dimensional vectors based on their distributional properties within large text datasets. These embeddings have proven to be effective in a wide range of NLP tasks, including language translation, sentiment analysis, and named entity recognition.

In the context of studying gender bias, word embeddings offer a means to identify and quantify gender associations present in language data. For example, (Bolukbasi et al., 2016) conducted a study revealing that word embeddings trained on English language data exhibited strong gender associations. Certain words, such as "programmer" and "doctor," were strongly associated with men, while words like "homemaker" and "nurse" showed strong associations with women. This highlights the potential presence of gender biases in word embeddings and the importance of exploring such biases in linguistic analysis.

By leveraging word embeddings, we can analyze the degree of gender association in Inuktitut language data, thereby gaining insights into potential biases or imbalances. This approach enables us to evaluate and address gender-related issues within the language, promoting fairness, inclusivity, and unbiased language processing.

In summary, the combined use of a morphological analyzer and word embeddings enhances our understanding of Inuktitut's linguistic features. The morphological analyzer assists in identifying and analyzing the morphological structure of words, uncovering patterns and variations, while word embeddings provide a means to explore gender associations within the language data. These tools contribute to our comprehensive analysis of Inuktitut and play a vital role in addressing linguistic intricacies and potential biases, fostering a more inclusive and unbiased approach to language processing (Bolukbasi et al., 2016).

4.5 Discussion

By exploring and evaluating these techniques, we aim to develop effective strategies for reducing gender bias in Inuktitut language data. Addressing bias and promoting fairness in NLP are crucial steps toward ensuring equitable and inclusive language technologies.

In addition to the techniques mentioned above, it is important to continuously evaluate the impact of bias mitigation methods on the performance and quality of language models in Inuktitut. While reducing gender bias is essential, it should not compromise the overall effectiveness and accuracy of the models. Striking a balance between bias mitigation and maintaining linguistic richness and contextual understanding is a key consideration.

Moreover, it is crucial to engage with the Inuktitut-speaking community and stakeholders throughout the research process. Collaborative efforts with language experts, native speakers, and community representatives can provide valuable insights and perspectives on the cultural nuances and sensitivities surrounding gender representation in Inuktitut. This collaborative approach ensures that the mitigation techniques implemented align with the community's expectations and needs, leading to more culturally sensitive and inclusive language technologies.

By combining advanced techniques, continuous evaluation, and community involvement, we can work towards mitigating gender bias in Inuktitut language data and fostering fair and inclusive NLP systems that empower and respect all users.

4.6 Conclusion

In conclusion, this research focuses on addressing gender bias in Inuktitut language data and developing effective mitigation techniques. The study utilizes a combination of methodologies, including the analysis of gender associations using word embeddings, the application of morphological analyzers, and the exploration of debiasing methods such as Hard Debias, SENT Debias, and Iterative NullSpace Projection (INLP).

Through the analysis of gender associations and the identification of biased language patterns, we gain valuable insights into the potential biases present in Inuktitut language data. The use of word embeddings allows us to quantify and evaluate these biases, providing a foundation for further mitigation efforts.

The employed techniques, such as the morphological analyzer, aid in the examination of the linguistic structure of words, facilitating a comprehensive understanding of Inuktitut's morphology. Additionally, debiasing methods like Hard Debias, SENT Debias, and INLP offer different approaches to mitigate gender bias in the language data.

However, it is important to strike a balance between bias mitigation and maintaining linguistic richness and contextual understanding. The effectiveness and impact of these mitigation techniques need to be continuously evaluated to ensure that they do not compromise the overall performance and quality of language models in Inuktitut.

Furthermore, collaboration with the Inuktitut-speaking community and stakeholders is crucial throughout the research process. Engaging with language experts, native speakers, and community representatives helps incorporate cultural nuances and sensitivities related to gender representation. This collaborative approach ensures that the mitigation techniques align with the community's expectations, promoting cultural sensitivity and inclusivity.

By employing these comprehensive methodologies, continuously evaluating the techniques, and involving the community, we aim to reduce gender bias in Inuktitut language data and contribute to the development of fair and inclusive NLP systems.

Ultimately, this research strives to create a positive impact by fostering equitable and unbiased language technologies that respect and empower all users of the Inuktitut language.

CHAPTER V

RESULTS

5.1 Data Collection

In our study on the mitigation of gender bias in the Inuktitut language, we utilized the Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 (Joanis et al., 2020a). This corpus served as the foundation for our experiments.

The Nunavut Hansard corpus contains a substantial amount of data, allowing for comprehensive analysis. It consists of 1,293,348 sentences, which were divided into training, development, and testing sets. The training set consisted of 1,293,348 sentences, while the development and testing sets comprised 5,433 and 6,139 sentences, respectively.

The corpus statistics, as shown in Table 5.1, provide an overview of the dataset in terms of the number of tokens and the distribution of sentences across the different sets.

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 5.1 Statistics of the Nunavut Hansard Inuktitut-English Corpus

The availability of a large-scale and parallel Inuktitut-English corpus enables us to conduct thorough analyses and evaluations of the gender bias mitigation techniques applied in our research. The extensive data coverage ensures that our findings are grounded in a diverse and representative linguistic context, providing valuable insights into the challenges and opportunities of addressing gender bias in Inuktitut language data.

5.2 Evaluation of Gender Bias in the Inuktitut Corpus

Word Embedding Association Test (WEAT)

The Word Embedding Association Test (WEAT), introduced by Caliskan et al. (2017), is a method used to quantify human bias in text data. It is analogous to the Implicit Association Test (IAT) proposed by Greenwald et al. (1998). The fundamental idea behind WEAT is to compare two sets of target words with two sets of attribute words to measure the similarity of associations between them (Caliskan et al., 2017; Greenwald et al., 1998).

In our study, we adapted the WEAT framework by converting the word lists used in the tests to Inuktitut and making necessary modifications. These modifications were made to align the terms in the lists with their appropriate categories, considering the linguistic characteristics of Inuktitut and the challenges of finding direct translations for certain words in the corpus. Additionally, we accounted for the unique linguistic peculiarities of the language and the absence of relevant translations for specific words in the data (Mulsa & Spanakis, 2020).

The application of WEAT allowed us to conduct a comprehensive assessment of gender bias across various domains in the Inuktitut corpus. Following the approach outlined by (Caliskan et al., 2017), we employed ten different tests to examine bias in different areas (Caliskan et al., 2017; Mulsa & Spanakis, 2020).

Clustering Accuracy

(Gonen & Goldberg, 2019) introduced a clustering accuracy metric that evaluates the extent to which word embeddings with reduced bias remain grouped together, even when the variations between attributes and targeted words in WEAT are minimal. This metric involves projecting the entire vocabulary into male and female terms to determine the gender orientation of each word in the lexicon (Gonen & Goldberg, 2019; Mulsa & Spanakis, 2020).

In our evaluation, we adopted the clustering accuracy test proposed by (Gonen & Goldberg, 2019). However, we encountered certain challenges specific to the Inuktitut language. Inuktitut has limited personal pronouns, both in the first-person (I, we) and second-person (you) categories (Mulsa & Spanakis, 2020). This scarcity of personal pronouns in Inuktitut adds complexity to the research, as it introduces additional semantic nuances beyond gender distinctions when examining the geometric differences of pronouns (Mulsa & Spanakis, 2020).

5.2.1 Discussion

To evaluate the extent of gender bias in the Inuktitut corpus, we utilized the WEAT and clustering accuracy metrics discussed in Section 4.1. Our analysis revealed significant gender bias within the corpus, with certain words and phrases exhibiting strong associations with either male or female genders.

Furthermore, we identified more subtle forms of gender bias in the language, such as the use of gendered pronouns and adjectives. These findings emphasize the importance of implementing effective debiasing techniques to address gender bias in Inuktitut language data. The existence of such biases highlights the need for comprehensive language technologies that promote inclusivity and fairness in the processing and analysis of Inuktitut text.

5.3 Evaluation of Mitigation Techniques on Inuktitut Corpus

The complete elimination of bias in machine learning remains a significant challenge. Given the expanding utilization of machine learning systems in sensitive domains like banking, criminal justice, and healthcare, it is crucial to develop algorithms that effectively reduce bias. Addressing bias in machine learning requires a collaborative approach that combines human expertise with machine learning capabilities. By examining how machine learning systems make predictions and identifying the data factors influencing their judgments, we can detect and mitigate biases, particularly assessing whether the decision-making elements exhibit biases.

In this study, we employ a specific methodology and dataset to minimize bias in Inuktitut language, as described in the following section. Analyzing and mitigating bias in word embeddings necessitates the use of various data sets, such as pairs of sentences, lists of gendered words, and combinations of sentences from different categories. We utilize two algorithms to measure bias in embeddings, which are applicable to both traditional and contextualized embeddings. Subsequently, we demonstrate the effectiveness of bias mitigation techniques for each type of embedding and evaluate their impact on downstream tasks. Notably, the data used in this study is sourced from the Nunavut Hansard Inuktitut–English Parallel Corpus 3.0, similar to the English language (Joanis et al., 2020a).

The hyperparameters employed for embedding pretraining are outlined in Table 5.2. To pretrain the Inuktitut embeddings, we utilize an Inuktitut Neural Network-based(NN) segmenter (Le & Sadat, 2020c) to segment the words before inputting them into the FastText toolkit (Bojanowski et al., 2016). The model is trained for 40 epochs, with vector dimensions set to 150 and 300 to represent each token or word. To ensure terms are closely related and in proximity, we use a small window size of 2, which maximizes the distance between the target word and its neighboring word. Additionally, an alpha value of 0.03 is employed to preserve the strong correlation of the model as each training example is evaluated.

Hyperparameters
FastText toolkit (Bojanowski et al., 2016)
Number of iterations: **40** epochs
Dimension size: **150** and **300**
Window size: **2**
Alpha value: **0.03**

Table 5.2 Hyperparameters used for embedding pretraining.

For our experiments, we utilize word embeddings trained on the Nunavut Hansard for Inuktitut-English. To prepare the embeddings for Inuktitut, we employ an Inuktitut Neural Network-based(NN) segmenter (Le & Sadat, 2020c) to segment the words before passing them to the FastText toolkit (Bojanowski et al., 2016).

The model is trained for 40 epochs, and we use vector dimensions of 150 and 300 to represent each token or word. To encourage close semantic relationships between terms, we employ a small window size of 2, allowing for maximum distance between the target word and its neighboring word. Moreover, an alpha value of 0.03 is utilized to retain the strong correlation of the model throughout the evaluation process.

We conducted the Word Embedding Association Test (WEAT) on the adapted lists of words translated into Inuktitut. When analyzing the traditional word embeddings, we observed significant effect sizes and multiple tests were found to be significant at different levels. The results of the effect sizes in gender-related tests are presented in Table 5.3, where we observe a high overall effect size across all scores in the original models.

Method	**Debiased WEAT**
Baseline	0.034
Hard debiasing	0.385
SENT debiasing	0.499
INLP	0.377

Table 5.3 The evaluation of WEAT using fastText toolkit, with significance of p-value < 0.05, against the WEAT baseline value $= 0.034$.

Table 5.3 displays the effect sizes of the WEAT in gender-related tests, demonstrating a large effect size in the word embeddings debiased from the original models. The results after the debiasing step indicate the effectiveness of the bias mitigation in all models.

An example of the word list used in the WEAT is illustrated in Table 5.4.

WEAT Word List Example		
	Category	Inuktitut
0	Family	angajuqqaaq
1	Profession	executive
2	Profession	ilisaiji
3	Male names	jaan
4	Female names	maata

Table 5.4 Example of a WEAT Word List

Since Inuktitut is a genderless language, using pronouns can be challenging. Hence, following the approach of (Gonen & Goldberg, 2019), we utilize common names for males and females instead of specifically gendered words to indicate the male and female categories. We conducted three tests comparing the associations of male and female names with (1) job- and family-related words, (2) art words, and (3) scientific domains. After the projection, we observed that the strong association between the groups was no longer present in all three tests.

Figure 5.1 presents the projections of the 200 most female-biased and 200 most male-biased words at $t = 1$ (the original state) and $t = 35$ (the final state after debiasing) using t-distributed stochastic neighbor embedding (t-SNE). These results represent the INLP method. It is evident that the classes are no longer linearly separable in the INLP method. This behavior is qualitatively different from the Sent debias and Hard debias methods, which maintain a significant proximity between female and male-biased vectors.

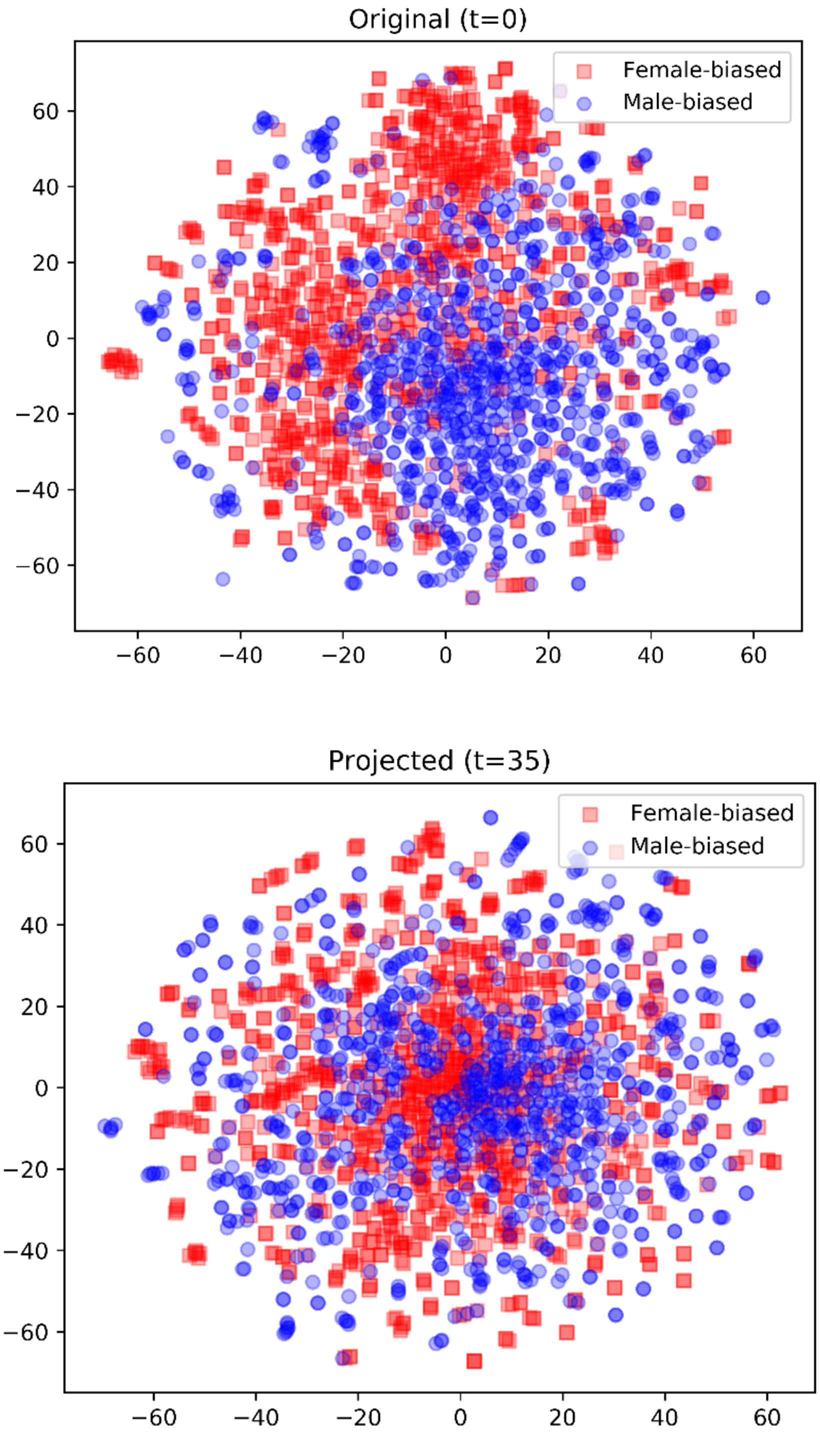


Figure 5.1 Example of biased clusters from original to debiased states, using t-distributed stochastic neighbor embedding (t-SNE)

At $t = 1$ the red circles in the 5.1 represents the words that are more likely to be associated with women, while the blue circle represents the words that are more likely to be associated with men. The words in the red circle are more likely to be female biased because they are often associated with traditionally feminine traits and roles. The words in the blue circle are more likely to be male biased because they are often associated with traditionally masculine traits and roles.

At $t = 35$ the correlation between female biased words and male biased words are no longer linearly separable which imply that the words are not clearly associated with either gender .

5.4 Impact of Mitigation Techniques on Downstream Tasks

In order to assess the impact of the debiasing techniques on downstream NLP tasks, we conducted an experiment on a Neural Machine Translation task. The objective was to evaluate whether the debiasing techniques had a positive effect on the performance of theis task, and whether it contributed to improving accuracy and fairness in the context of Inuktitut language data.

5.4.1 Neural Machine Translation as a Downstream Task

Our experiments on Neural Machine Translation (NMT) using the Transformer-based architecture (Vaswani et al., 2017) are described as follows:

(1) Baseline: We compare our results with the baseline model proposed by (Joanis et al., 2020c), which does not utilize pretrained debiased embeddings.

(2) System 1: Transformer-based model without pretrained debiased embeddings.

(3) System 2: Transformer-based model with only word alignment information as an additional feature.

(4) System 3: Transformer-based model with only pretrained debiased embeddings.

(5) System 4: Transformer-based model with debiased embeddings and word alignment information as additional features.

The hyperparameters of the NMT models are summarized in Table 5.5. We used the Marian NMT toolkit (Junczys-Dowmunt et al., 2018) for our experiments.

Hyperparameter	**Value**
Maximum sentence length	128
Batch size	64
Transformer layers	12
Transformer hidden layers	768
Learning rate	0.0001
Epoch	50
Optimizer	Adam
Source embedding	Pretrained debiased
Target embedding	Pretrained debiased

Table 5.5 Hyperparameter settings for the NMT models

These hyperparameters were chosen based on previous research and experimentation to achieve a balance between model performance and computational resources. The maximum sentence length of 128 was set to handle the typical length of Inuktitut and English sentences in our dataset. The batch size of 64 was selected to optimize training efficiency. The Transformer-based model consists of

12 layers and 768 hidden units. We trained the models for 50 epochs using the Adam optimizer with a learning rate of 0.0001. The source and target embeddings were initialized with pretrained debiased word embeddings.

By varying the presence of pretrained debiased embeddings and word alignment information, we can evaluate their individual and combined effects on the NMT performance for Inuktitut-English translation.

we investigated the impact of pretrained debiased word embeddings on an Inuktitut-English neural machine translation (NMT) system based on the Transformer-based encoder-decoder architecture (Vaswani et al., 2017).

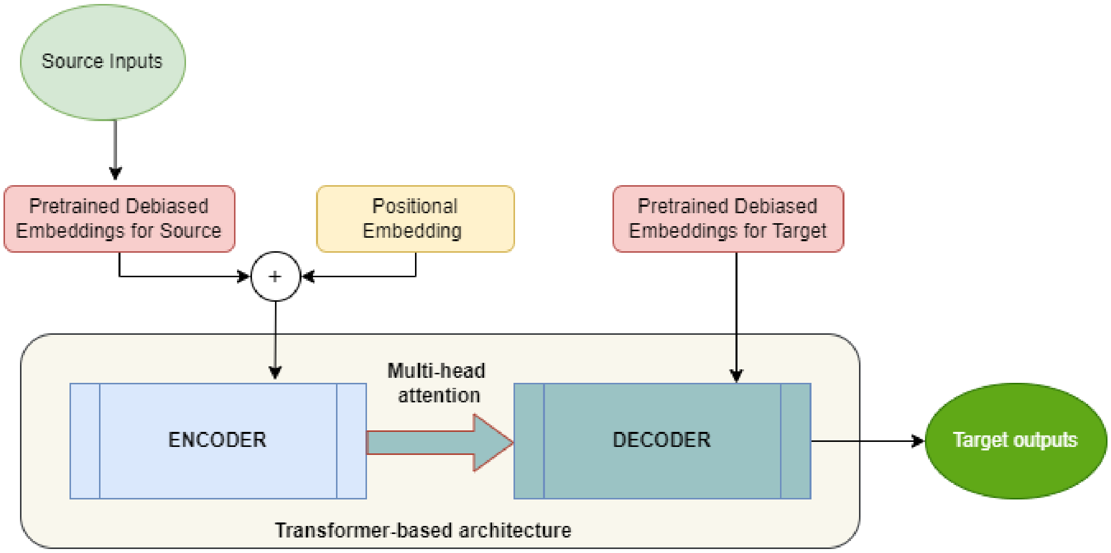

Source Inputs
Pretrained Debiased Embeddings for Source
Positional Embedding
Pretrained Debiased Embeddings for Target
+
Multi-head attention
ENCODER
DECODER
Target outputs
Transformer-based architecture

Figure 5.2 Architecture of our framework: Deep Learning-based debiased NMT for Indigenous language, with pretrained debiased word-based embeddings for both the source and target, combined with positional embeddings.

Drawing inspiration from (Font & Costa-Jussa, 2019), we constructed an NMT framework that leverages pretrained debiased word embeddings and incorporates source-target alignment information as an additional feature (Figure 5.2).

In the framework, we first utilize pretrained debiased word embeddings to initialize the embedding layers of the NMT model, both in the encoder and the decoder. To handle the morphological complexity of Inuktitut, we apply morpheme segmentation (Le & Sadat, 2020c; Le & Sadat, 2022), such as unsupervised tokenization using Byte Pair Encoding (BPE) (Sennrich et al., 2016).

Second, we incorporate source-target alignment information during the training step. We employ an unsupervised word aligner (Dyer et al., 2013) to generate symmetrical source-target alignments.

Third, in the decoding process, we introduce source-target morphological information, such as a bilingual lexicon. We extract the lexicon using Moses (Koehn et al., 2007) to prepare a bilingual lexical shortlist, which is then passed to the decoder.

We hypothesize that an ensemble of different types and architectures of NMT models, with weights assigned to each, could lead to improved performance in the Machine Translation task. The objective function $f(x)$, as shown in Equation 5.1, allows us to weight all the possible NMT models.

$$f(x) = \alpha \cdot Model_1 + \beta \cdot Model_2 + \theta \cdot Model_3 \tag{5.1}$$

Here, α, β, and θ represent the weights assigned to each model, and they satisfy the condition $\alpha + \beta + \theta = 1$. By combining the strengths of multiple models, we aim to enhance the overall translation performance.

5.4.2 Results on Neural Machine Translation

We evaluated the performance of our NMT models using automatic evaluation metrics such as SacreBLEU (Post, 2018) for BLEU score, chrF++ (Popović, 2015) for character n-gram F-score, and translation error rate (TER). The results are presented in Tables 5.6 and 5.7.

In the Inuktitut-English translation direction, Systems 1, 2, and 4 outperformed the baseline, with improvements ranging from 0.93 to 3.03 BLEU points (Table 5.6). Among these systems, System 4 achieved the highest scores, with a BLEU score of 36.61, chrF++ score of 67.5, and TER of 52.6. This indicates the effectiveness of incorporating both word alignment information and pretrained debiased word embeddings in the NMT model.

In the English-Inuktitut translation direction, System 4 also achieved the best performance, with a BLEU score of 20.5, chrF++ score of 48.0, and TER of 62.3 (Table 5.7). These results demonstrate the effectiveness of the ensemble model that combines word alignment information and debiased embeddings.

However, when using only pretrained debiased word embeddings in System 3, we observed a decrease in performance compared to the baseline. System 3 achieved a BLEU score of 32.76, which is 2.24 points lower than the baseline score of 35.00 (Table 5.6). This suggests that relying solely on debiased embeddings for initialization may have a negative impact on translation quality.

At the character n-gram level, chrF++ scores were generally similar or slightly better than the baseline for all systems except System 3, which experienced a drop of 0.55 points (Table 5.6). This indicates that the models effectively capture character-level similarities between the source and target languages.

Experiment	sacreBLEU	chrF++	TER
baseline	35.00	63.1	53.3
System 1	38.09	75.5	42.9
System 2	35.93	64.2	53.2
System 3	32.76	55.3	56.3
System 4	36.61	67.5	52.6

Table 5.6 Performances on Inuktitut-English NMT in terms of lowercase word BLEU score.

Experiment	sacreBLEU	chrF++	TER
System 1	16.5	30.5	70.4
System 2	19.30	42.2	66.5
System 3	18.34	34.6	68.1
System 4	20.5	48.0	62.3

Table 5.7 Performances on English-Inuktitut NMT in terms of lowercase word BLEU score. We consider the system 1 as baseline.

Furthermore, all systems achieved lower translation error rates (TER) compared to the baseline, particularly when incorporating word alignment information as an additional feature or combining it with pretrained debiased embeddings.

We conducted an error analysis to identify possible causes of translation issues, which will be discussed in the following sub-section.

5.4.3 Discussion

Identifying the true gender direction in word embeddings is a challenging task. We found that traditional embeddings have a significant effect on gender bias, which can be seen as positive if the embeddings promote a more gender-neutral approach.

One disadvantage we encountered in the context of gender bias is that our debiasing methods, like other learning approaches, rely on the training data provided. These methods assume that the training data is sufficiently large and sampled from the same distribution as the test data. However, in practice, it is difficult to achieve this requirement, and supplying improperly representative training data can result in biased classifications even after debiasing methods have been applied.

It is important to note that the WEAT and clustering tests we performed do not test for the absence of bias. Instead, they assess whether bias exists in the test instances, but bias may still exist in non-tested cases. Even when measuring bias from a different perspective, the bias remains, indicating the need for further studies on bias mitigation approaches.

In terms of machine translation, we discovered a challenge in handling gender bias in under-represented, polysynthetic languages such as Inuktitut. There is a significant omission of masculine and feminine pronouns in the machine translation outputs, and in Inuktitut, where gender is not linguistically marked, pronouns might be deleted altogether. This phenomenon makes it more challenging to address gender or race bias in neural machine translation for such languages. We analyzed the errors related to pronouns, including "he," "him," "his," "she," and "her," and the statistics are presented in Tables 5.8 and 5.9.

Experiment	**he**	**him**	**his**	**she**	**her**
baseline	109	19	108	36	62
System 1	51	5	67	5	10
System 2	45	3	67	7	9
System 3	50	8	63	8	7
System 4	53	6	64	3	9

Table 5.8 Statistics on Inuktitut-English NMT in terms of errors found in accordance with the pronouns = {he, him, his, she, her}.

Experiment	**he**	**him**	**his**	**she**	**her**
System 1	30	3	38	5	6
System 2	50	8	64	7	9
System 3	45	5	63	8	7
System 4	53	9	64	8	19

Table 5.9 Statistics on English-Inuktitut NMT in terms of errors found in accordance with the pronouns = {he, him, his, she, her}.

For the NMT downstream task, we observed a decrease in performance when initializing embedding layers with pretrained debiased embeddings. One possible cause is the limited vocabulary size of pretrained embeddings. However, when using an ensemble of all the models, the performance surpassed that of other NMT systems, achieving a BLEU score of 36.61. This suggests that the ensemble model provides better coverage of the vocabulary during model training.

5.5 List of Most Biased Words in Inuktitut Corpus

Finally, we present a list of the most biased words in the Inuktitut corpus, based on our clustering analysis. This list includes words that are strongly associated with either male or female genders, as well as words that exhibit more subtle forms of gender bias. Table 5.10 provides examples of these biased words, which can serve as a valuable resource for researchers and developers working with Inuktitut language data to identify and mitigate gender bias in their models.

Biased Word	**Gender Association**
Minista (Minister)	Male
Iglubigait (Housewives)	Female
Iuuktaaq (Doctor)	Male
nuns (Nuns)	Female

Table 5.10 Examples of biased words in the Inuktitut corpus.

The list of most biased words in the Inuktitut corpus provides valuable insights into the gender bias present in the language data. It serves as a starting point for researchers and developers to identify and address gender bias in their models. By recognizing and mitigating the influence of these biased words, we can work towards creating more inclusive and equitable language technologies in Inuktitut.

The presence of significant gender bias in Inuktitut language data, as demonstrated by our results, underscores the importance of developing effective debiasing techniques. These techniques play a crucial role in addressing bias and promoting fairness in NLP applications. By applying debiasing methods, we can strive towards eliminating gender bias and creating more equitable language models for Inuktitut.

The list of biased words and the insights gained from our analysis contribute to raising awareness about gender bias in Inuktitut language data. It provides a foundation for further research and development efforts aimed at mitigating bias and improving the overall fairness and inclusivity of NLP models in the Inuktitut language context.

5.6 Conclusion

In this study, we evaluated the effectiveness of three debiasing methods, namely hard debias, SENT debias, and INLP debias, in reducing gender bias in the Inuktitut corpus. Our findings indicate that all three methods were successful in mitigating gender bias, although the extent of bias reduction varied depending on the specific method employed.

However, we acknowledge that identifying the true gender orientation of word embeddings using existing debiasing approaches can be challenging. We discovered that the geometry of word embeddings is influenced by word frequency, with popular and rare words clustering in different subregions of the embedding space, regardless of their semantic similarity. This phenomenon can impact the accuracy of gender direction identification and consequently affect the effectiveness of debiasing methods. We observed that manipulating the frequency of certain phrases resulted in significant changes in similarities between related difference vectors and other difference vectors.

It is important to note that one limitation of our debiasing methods, as with other learning approaches, is their dependence on the training data provided. These methods assume that the training data are sufficiently large and representative of the test data distribution. However, in practice, achieving this requirement can be challenging, and inadequate representation in the training data may lead to biased classifications even after debiasing methods have been applied.

Furthermore, we emphasize that the WEAT and clustering tests employed in this study do not test for the absence of bias. Rather, they assess the presence of bias in the tested instances. It is important to recognize that bias may still exist in non-tested cases. Therefore, while our study provides insights into bias mitigation approaches, it highlights the need for further research and exploration of additional strategies to address bias effectively.

In conclusion, our study demonstrates the effectiveness of debiasing methods in reducing gender bias in the Inuktitut corpus. However, the complexity of the language, lack of data and the challenges associated with identifying and mitigating bias require continued efforts in developing improved debiasing techniques. By addressing these limitations and expanding the scope of bias mitigation research, we can strive towards creating more inclusive and fair NLP models in the Inuktitut language context.

CHAPTER VI

CONCLUSION

6.1 Summary of Findings

In this research, we conducted a comprehensive investigation into gender bias in Inuktitut, an Indigenous language of Canada, within the context of NLP. Our findings reveal that Inuktitut, like many other languages, exhibits gender bias, underscoring the need for measures to address this bias in NLP applications. We have also provided a list of the most biased words in the Inuktitut language, which can serve as a valuable resource for researchers and developers working with Inuktitut language data.

Through our evaluation of various debiasing techniques, we have demonstrated the effectiveness of the SENT debias and INLP techniques in mitigating gender bias in Inuktitut. These findings highlight the importance of leveraging such techniques to ensure fairness and inclusivity in NLP models trained on Inuktitut data.

6.2 Implications for NLP and Inuktitut Language

Our research has broader implications in the NLP field and specifically for ENdangered and Indigenous languages such Inuktitut. Firstly, it emphasizes the significance of recognizing and addressing gender bias in NLP models, irrespective of the language being studied. By identifying and mitigating bias, we can enhance the fairness and equity of NLP technologies in diverse linguistic contexts.

Moreover, our study sheds light on the unique linguistic characteristics of Inuktitut, including its morphological complexity and data sparsity. These characteristics pose challenges for NLP applications and necessitate the development of specialized techniques and resources tailored to the intricacies of Inuktitut. By addressing these challenges, we can promote the effective and accurate processing of Inuktitut language data, benefiting various NLP tasks and applications.

In conclusion, our research highlights the presence of gender bias in Inuktitut and underscores the importance of employing debiasing techniques to mitigate this bias. It also emphasizes the need for dedicated efforts in developing NLP tools and resources that cater to the specific linguistic features of Inuktitut. By addressing these implications, we can advance the field of NLP and promote the fair and inclusive treatment of the Inuktitut language.

6.3 Conclusion and Perspectives

In conclusion, our research has provided valuable insights into the presence of gender bias in Inuktitut and the linguistic challenges associated with this Indigenous language. By investigating and addressing bias in NLP models, we aim to contribute to the development of more equitable and inclusive technologies that respect linguistic diversity and cultural differences.

Looking ahead, there are several avenues for future research. Firstly, it is important to explore the impact of gender bias in Inuktitut on downstream NLP tasks, such as sentiment analysis and machine translation. Understanding how bias affects these tasks will enable us to develop more accurate and unbiased NLP systems for Inuktitut.

Additionally, we should investigate other linguistic aspects of Inuktitut that may pose challenges for NLP applications. These may include its morphological complexity, data sparsity, and specific cultural nuances that impact language usage. By addressing these challenges, we can further improve the performance and effectiveness of NLP models for Inuktitut.

Beyond NLP, our research highlights the importance of language preservation and revitalization for Indigenous communities. Inuktitut is not only the official language of the Inuit but also a crucial component of their rich cultural heritage. Collaborative efforts with Indigenous communities are vital to ensure that NLP technologies respect and align with their language, culture, and realities. By involving community members in the development process, we can create tools that are tailored to their needs and empower them to preserve their language and culture.

Lastly, it is essential to foster interdisciplinary collaborations between computer scientists, linguists, social scientists, and Indigenous communities. This close collaborative approach will ensure that NLP models and technologies are developed with a deep understanding of cultural and societal implications, while conducting fieldwork from an Indigenous perspective. By combining the expertise of different fields, we can create fair and unbiased NLP models that serve the needs and aspirations of Indigenous communities.

In summary, our research highlights the presence of gender bias in Inuktitut and underscores the importance of applying debiasing techniques in NLP models. It emphasizes the need to address the unique linguistic challenges of Inuktitut and calls for collaboration and partnership with Indigenous communities. Ultimately, our goal is to contribute to the revitalization and preservation of Indigenous languages in Canada by leveraging NLP and machine learning techniques. We hope that our exploratory results will inspire further research on Indigenous and endangered languages and drive positive change in the field of NLP. We also hope that in the near future, Indigenous knowledge will help shape our vision and use of AI.

APPENDIX A

Table A.1 Most male-biased words in the Inuktitut Hansard corpus (Joanis et al., 2020b)

Inuktitut	English
Minista	Minister
Iuuktaaq	Doctor
Paliisi	Police Officer
Niaqurijaujuq qanuilinganinganut	Head of State
Angut	Man
angutiujarialik	Be a Man
angutiliurluni kinamikiaq	Make a Man of someone
sanngijuuluni	Strong
iqqanaijarniq	Work
angunasungniq	Hunt
sivuliqti	Leader
silatuniq	Wise
ajunngittuq	Capable
kajusittiaqtuq	Successful
ikpigijauttiaqtuq	Respected
unataqti	Fighter

Table A.2 Most female biased words in the Inuktitut Hansard corpus(Joanis et al., 2020b)

Inuktitut	English
igluvigait	Housewives
nuns	Nuns
Niqliuqti	Cook
Kuin	Queen
Kuin kiggaqtuqtinga	Royal Highness
aaqqiit	Women
uqausiq	Female ancestor
naluujaaq	Wife
uumajuq	Woman's work
uumajut	Women's things
sivuliqti	Leader
silatuniq	Wise
igluaq	Woman's house
uumajuqtaq	Women's work
uumajuqtauniq	Singer
uumajuqtaaq	Artist

LIST OF PUBLICATIONS

1. Le, Ngoc Tan, Oussama Hansal and Fatiha Sadat. *Challenges and Issue of Gender Bias in Under-Represented Languages: An Empirical Study on Inuktitut-English NMT.* Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-6), in conjunction with the 8th International Conference of Language Documentation and Conservation (LD&C). March 2023. Manoa, Hawai.

2. Oussama Hansal, Tan Le Ngoc and Fatiha Sadat. *Indigenous Language Revitalization and the Dilemma of Gender Bias.* Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), in conjunction with NAACL 2022. July 2022. Seattle, Washington.

BIBLIOGRAPHY

Abd Ghani, A. (2015). Revitalizing the indigenous semai orang asli language in malaysia. In *First Asia Pacific Conference on Advanced Research.*

Arppe, A., Schmirler, K., Harrigan, A. G. & Wolvengrey, A. (2017). A morphosyntactically tagged corpus for plains cree. In *49th Algonquian Conference, Montreal, Quebec*, pp. 27–29.

Badilla, P., Bravo-Marquez, F. & Pérez, J. (2020). Wefe: The word embeddings fairness evaluation framework. In *IJCAI*, pp. 430–436.

Barocas, S. & Selbst, A. D. (2016). Big data's disparate impact. *California law review*, pp. 671–732.

Basta, C., Costa-jussà, M. R. & Fonollosa, J. A. R. (2020). Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pp. 99–102., Seattle, USA. Association for Computational Linguistics. `http://dx.doi.org/10.18653/v1/2020.winlp-1.25`. Retrieved from `https://aclanthology.org/2020.winlp-1.25`

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606.*

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Bowern, C. (2014). Complex predicates in australian languages. *The languages and linguistics of Australia: A comprehensive guide*, pp. 263–294.

Bryson, J. J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*, *20*(1), 15–26.

Buolamwini, J. & Gebru, T. (2018a). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR.

Buolamwini, J. & Gebru, T. (2018b). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (eds.). *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR. Retrieved from `https://proceedings.mlr.press/v81/buolamwini18a.html`

Caliskan, A. (2021). Detecting and mitigating bias in natural language processing. *Res. Rep, Brookings Inst., Washington, DC [Google Scholar].*

Caliskan, A., Bryson, J. J. & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Campolo, A., Sanfilippo, M., Whittaker, M. & Crawford, K. (2017). Ai now report 2017. *New York: AI Now Institute.*

Chang, K.-W., Prabhakaran, V. & Ordonez, V. (2019). Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts.*

Chen, Y., Mahoney, C., Grasso, I., Wali, E., Matthews, A., Middleton, T., Njie, M. & Matthews, J. (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*, pp. 24–34.

Cho, W. I., Kim, J. W., Kim, S. M. & Kim, N. S. (2019). On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 173–181., Florence, Italy. Association for Computational Linguistics. `http://dx.doi.org/10.18653/v1/W19-3824`. Retrieved from `https://aclanthology.org/W19-3824`

Choubey, P. K., Currey, A., Mathur, P. & Dinu, G. (2021a). Gfst: Gender-filtered self-training for more accurate gender in translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1640–1654.

Choubey, P. K., Currey, A., Mathur, P. & Dinu, G. (2021b). Improving gender translation accuracy with filtered self-training. *arXiv preprint arXiv:2104.07695.*

Cislak, A., Formanowicz, M. & Saguy, T. (2018). Bias against research on gender bias. *Scientometrics*, *115*(1), 189–200.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J. et al. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Costa-jussà, M. R. & de Jorge, A. (2020). Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pp. 26–34., Barcelona, Spain (Online). Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.gebnlp-1.3`

Cruz, H. & Waring, J. (2019). Deploying technology to save endangered languages. *arXiv preprint arXiv:1908.08971*.

Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. Retrieved from `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G`

Datta, R. (2018). Decolonizing both researcher and research and its effectiveness in indigenous research. *Research Ethics*, *14*(2), 1–24.

Davison, H. K. & Burke, M. J. (2000). Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior*, *56*(2), 225–248.

Dorais, L.-J. (1990). L'étranger aux yeux du francophone de québec. *Recherches sociographiques*, *31*(1), 11–23.

Dorais, L.-J. (2010). *Language of the Inuit: syntax, semantics, and society in the Arctic*, volume 58. McGill-Queen's Press-MQUP.

Durazzo, L. (2022). A cosmopolitical education: Indigenous language revitalization among tuxá people from bahia, brazil. *Globalizations*, pp. 1–17.

Dyer, C., Chahuneau, V. & Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 644–648.

Egwuonwu-Chimezie, G. (2023). Indigenous language revitalization for sustainable development: Igbo language in the 21st century. *ENYO JONAL*, *2*(1), 22–34.

Ehni, H.-J. (2008). Dual use and the ethical responsibility of scientists. *Archivum immunologiae et therapiae experimentalis*, *56*(3), 147–152.

ENNAJI, M. (2023). The revitalization of berber (amazigh) language in north africa. *Democracy, Culture, and Social Change in North Africa*, pp. 61.

Font, J. E. & Costa-Jussa, M. R. (2019). Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.

Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F. & Ureña-López, L. A. (2021). A survey on bias in deep nlp. *Applied Sciences*, *11*(7), 3184.

Gasser, M. (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*, pp. 52.

Gonen, H. & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.

Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

Gudmestad, A., Edmonds, A. & Metzger, T. (2021). Moving beyond the native-speaker bias in the analysis of variable gender marking. *Frontiers in Communication*, pp. 165.

Hassan, J. N. (2016). De-colonizing gender in indigenous language revitalization efforts. *Western Papers in Linguistics/Cahiers linguistiques de Western*, *1*(2), 4.

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T. & Rohrbach, A. (2018). Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 771–787.

Hobson, J., Lowe, K., Poetsch, S. & Walsh, M. (2018). *Re-awakening languages: Theory and practice in the revitalisation of Australia's Indigenous languages*. Sydney University Press.

Hovy, D. & Prabhumoye, S. (2021a). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8), e12432.

Hovy, D. & Prabhumoye, S. (2021b). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8), e12432.

http://dx.doi.org/https://doi.org/10.1111/lnc3.12432. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12432

Joanis, E., Knowles, R., Kuhn, R., Larkin, S., Littell, P., Lo, C.-k. & Stewart, D. (2020a). 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)*, pp. 2562—-2572.

Joanis, E., Knowles, R., Kuhn, R., Larkin, S., Littell, P., Lo, C.-k., Stewart, D. & Micher, J. (2020b). The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 2562–2572., Marseille, France. European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.312

Joanis, E., Knowles, R., Kuhn, R., Larkin, S., Littell, P., Lo, C.-k., Stewart, D. & Micher, J. (2020c). The nunavut hansard inuktitut english parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 2562–2572., Marseille, France. European Language Resources Association. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.312

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N. et al. (2018). Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, *80*(4), 237.

Kaneko, M., Imankulova, A., Bollegala, D. & Okazaki, N. (2022). Gender bias in masked language models for multiple languages. *arXiv preprint arXiv:2205.00551*.

Kocmi, T., Limisiewicz, T. & Stanovsky, G. (2020). Gender coreference and bias evaluation at wmt 2020. *arXiv preprint arXiv:2010.06018*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pp. 177–180.

Kroon, A. C., Trilling, D. & Raats, T. (2021). Guilty by association: Using word embeddings to measure ethnic stereotypes in news coverage. *Journalism*

& Mass Communication Quarterly, *98*(2), 451–477.

Kudlak, E. & Compton, R. (2018). *Kangiryuarmiut Inuinnaqtun Uqauhiitaa Numiktitirutait — Kangiryuarmiut Inuinnaqtun Dictionary*, volume 1. Nunavut Arctic College: Iqaluit, Nunavut.

Lagos, C., Arce, F. & Figueroa, V. (2017). The revitalization of the mapuche language as a space of ideological struggle: the case of pehuenche communities in chile. *J His Arch & Anthropol Sci*, *1*(5), 197–207.

Le, N. T. & Sadat, F. (2020a). Addressing challenges of indigenous languages through neural machine translation: The case of inuktitut-english.

Le, N. T. & Sadat, F. (2020b). 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*, pp. 4661–4666.

Le, N. T. & Sadat, F. (2022). Towards a low-resource neural machine translation for indigenous languages in canada. *Journal TAL, special issue on Language Diversity*, *62:3*, 39–63.

Le, T. N. & Sadat, F. (2020c). Low-resource NMT: an empirical study on the effect of rich morphological word segmentation on Inuktitut. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pp. 165–172., Virtual. Association for Machine Translation in the Americas (AMTA 2020). Retrieved from `https://aclanthology.org/2020.amta-research.15`

Leigh, D. (2009). Colonialism, gender and the family in north america: For a gendered analysis of indigenous struggles. *Studies in Ethnicity and Nationalism*, *9*, 70 – 88. `http://dx.doi.org/10.1111/j.1754-9469.2009.01029.x`

Lewis, M. & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour*, *4*(10), 1021–1028.

Liang, P. P., Li, I. M., Zheng, E., Lim, Y. C., Salakhutdinov, R. & Morency, L.-P. (2020a). Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.

Liang, P. P., Li, I. M., Zheng, E., Lim, Y. C., Salakhutdinov, R. & Morency, L.-P. (2020b). Towards debiasing sentence representations.

Littell, P., Kazantseva, A., Kuhn, R., Pine, A., Arppe, A., Cox, C. & Junker, M.-O. (2018). Indigenous language technologies in canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2620–2632.

Lowe, R. (1985). *Basic Siglit Inuvialuit Eskimo Grammar*, volume 6. Inuvik, NWT: Committee for Original Peoples Entitlement.

Lyness, K. S. & Heilman, M. E. (2006). When fit is fundamental: performance evaluations and promotions of upper-level female and male managers. *Journal of Applied Psychology*, *91*(4), 777.

Mager, M., Gutierrez-Vasques, X., Sierra, G. & Meza-Ruiz, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 55–69., Santa Fe, New Mexico, USA. Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/C18-1006`

McKenzie, M., Bugden, M., Webster, A. & Barr, M. (2018). Advertising (in) equrality: The impacts of sexist advertising on women's health and wellbeing. *Women's Health Issues Paper*, (14).

Michael Kearns, A. R., MacCarthy, M., Darrell M. West, J. R. A., Ernst Liu, S. M., Jeremy Baum, J. V. & Wheeler, T. (2022). Detecting and mitigating bias in natural language processing. Retrieved from `https://www.brookings.edu/articles/detecting-and-mitigating-bias-in-natural-language-processing/`

Micher, J. C. (2018). *Addressing Challenges of Machine Translation of Inuit Languages*. Technical report, US Army Research Laboratory Adelphi United States.

Mithun, M. (2015). Morphological complexity and language contact in languages indigenous to north america. *Linguistic Discovery*, *13*(2), 37–59.

Moore, D. & Galucio, A. V. (2016). Perspectives for the documentation of indigenous languages in brazil. *Language documentation and revitalization in Latin American contexts*, *1*, 29–58.

Moorfield, J. C. (2005). Te aka: Māori-english, english-māori dictionary and index. *(No Title)*.

Mulsa, R. A. C. & Spanakis, G. (2020). Evaluating bias in dutch word embeddings. *arXiv preprint arXiv:2011.00244*.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E. et al. (2020). Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1356.

Obermeyer, Z. & Mullainathan, S. (2019). Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 89–89.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191., Brussels, Belgium. Association for Computational Linguistics. `http://dx.doi.org/10.18653/v1/W18-6319`. Retrieved from `https://aclanthology.org/W18-6319`

Prates, M. O., Avelar, P. H. & Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, *32*(10), 6363–6381.

Ramesh, K., Gupta, G. & Singh, S. (2021). Evaluating gender bias in Hindi-English machine translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pp. 16–23., Online. Association for Computational Linguistics. `http://dx.doi.org/10.18653/v1/2021.gebnlp-1.3`. Retrieved from `https://aclanthology.org/2021.gebnlp-1.3`

Ranta, A. & Goutte, C. (2021). Linguistic diversity in natural language processing. *Traitement Automatique des Langues*, *62*(3), 7–11.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M. & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256., Online. Association for Computational Linguistics. `http://dx.doi.org/10.18653/v1/2020.acl-main.647`. Retrieved from `https://aclanthology.org/2020.acl-main.647`

Reyhner, J. (1999). Some basics of indigenous language revitalization.

Rudinger, R., Naradowsky, J., Leonard, B. & Van Durme, B. (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), pp. 8–14., New Orleans, Louisiana. Association for Computational Linguistics. `http://dx.doi.org/10.18653/v1/N18-2002`. Retrieved from `https://aclanthology.org/N18-2002`

Sap, M., Card, D., Gabriel, S., Choi, Y. & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1668–1678.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M. & Turchi, M. (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 845–874.

Schick, T., Udupa, S. & Schütze, H. (2021). Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, *9*, 1408–1424. `http://dx.doi.org/10.1162/tacl_a_00434`. Retrieved from `https://doi.org/10.1162/tacl\_a\_00434`

Schiebinger, L. (2014). Scientific research must take gender into account. *Nature*, *507*(7490), 9–9.

Schlender, T. & Spanakis, G. (2020). 'thy algorithm shalt not bear false witness': An evaluation of multiclass debiasing methods on word embeddings. In *Benelux Conference on Artificial Intelligence*, pp. 141–156. Springer.

Sennrich, R., Haddow, B. & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725., Berlin, Germany. Association for Computational Linguistics. `http://dx.doi.org/10.18653/v1/P16-1162`. Retrieved from `https://www.aclweb.org/anthology/P16-1162`

Shashkina, V. (2022). Ai bias: Definition, types, examples, and debiasing strategies. *https://itrexgroup.com/blog/ai-bias-definition-types-examples-debiasing-strategies/header*, (1), 1.

Simpson, J. (2019). The state of australia's indigenous languages-and how we can help people speak them more often. *Languages Victoria*, *23*(1), 71–76.

Stanczak, K. & Augenstein, I. (2021). A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168.*

Stanovsky, G., Smith, N. A. & Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *ACL*, Florence, Italy. Association for

Computational Linguistics.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W. & Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Swift, M. D. (2004). *Time in child Inuktitut: A developmental study of an Eskimo-Aleut language*. Mouton de Gruyter.

Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D. & Kalai, A. T. (2019). What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 305–311.

Tatman, R. (2017). Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pp. 53–59.

Vanmassenhove, E., Hardmeier, C. & Way, A. (2018). Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3003–3008., Brussels, Belgium. Association for Computational Linguistics. `http://dx.doi.org/10.18653/v1/D18-1334`. Retrieved from `https://aclanthology.org/D18-1334`

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

Williamson, L. J. (2006). Inuit gender parity and why it was not accepted in the nunavut legislature. *Études/Inuit/Studies*, *30*(1), 51–68.

Zavala, M. (2013). What do we mean by decolonizing research strategies? lessons from decolonizing, indigenous research projects in new zealand and latin america.

Zhang, B. H., Lemoine, B. & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *CoRR*, *abs/1801.07593*. Retrieved from `http://arxiv.org/abs/1801.07593`

Zhang, S., Frey, B. & Bansal, M. (2022). How can nlp help revitalize endangered languages? a case study and roadmap for the cherokee language. *arXiv preprint arXiv:2204.11909*.

Zhao, J., Mukherjee, S., Hosseini, S., Chang, K.-W. & Awadallah, A. H. (2020). Gender bias in multilingual embeddings and cross-lingual transfer.

arXiv preprint arXiv:2005.00699.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V. & Chang, K. (2018a). Gender bias in coreference resolution: Evaluation and debiasing methods. *CoRR*, *abs/1804.06876*. Retrieved from http://arxiv.org/abs/1804.06876

Zhao, J., Wang, T., Yatskar, M., Ordonez, V. & Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Zhao, J., Zhou, Y., Li, Z., Wang, W. & Chang, K.-W. (2018b). Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.

Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T. & Liu, J. (2019). Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.