Evaluating Bias Transfer Across Culture Representations Using RoBERTa: A Study of Indigenous and Canadian Political Discourse

Yanis Bencheikh, Jessica Ojo, Issa Saad
McGill University
{yanis.bencheikh, jessica.ojo, abdullah.saad}@mcgill.ca

Abstract

Transformer-based architectures, the backbone of large language models (LLMs), have driven the global adoption of natural language processing (NLP) to create conversational tools seemingly generalizable to diverse sociocultural linguistic contexts. When faced with the task of generating useful outputs in languages underrepresented in their training datasets, these models leverage knowledge transfer from overrepresented languages. However, this transfer can propagate biases proper to the group that dominates the training knowledge-learning space, leading to probable culturally insensitive misrepresentations of underrepresented cultures. Using the Canadian Hansard and Nunavut Hansard corpora, we have created mixture datasets where representativity of Indigenous discourse of the province of Nunavut was varied from none, to balanced, and imbalanced with respect to other Canadian provinces. We proceeded to fine-tune RoBERTa models on these mixture datasets and analyzed their embedding spaces using comparative word pair cosine similarity. Finally, using the first iteration of our novel *Canadian Indigenous Representation and Bias Evaluation* (CIRBE) benchmark, we evaluated bias transfer using binary masked language modeling (MLM) and open-ended MLM. To identify bias against the First Nations of Canada, each prompt features a masked token which can either take a favorable value which validates Inuit perspectives or one which ignores and invalidates them. In all experiments, we report an alarming effect of underrepresentation of Indigenous political discourse in training datasets. Canadian Hansard, pretrained and mixture model performances on our benchmark suggest a very likely harmful effect of deploying generative transformer-based models which do not accurately capture the reality of the Inuit of Nunavut. Indeed, as inclusion of Inuit perspectives decreases, the probability in binary MLM tasks that the models answer favorably also decreases.

1 Introduction

The rise of large language models (LLMs), powered by transformer-based architectures, has significantly advanced the field of natural language processing (NLP) by enabling more effective cross-linguistic and cross-cultural knowledge transfer (Devlin et al., 2019; Raffel et al., 2020). Transformer models, such as RoBERTa, serve as the underlying architecture for LLMs, allowing them to capture complex patterns in text and apply them across a variety of linguistic contexts (Zhuang et al., 2021). However, while these models excel in transferring knowledge, they also risk embedding and propagating biases from dominant to underrepresented cultures, which can result in misaligned or culturally inappropriate representations for languages like Inuktitut (Caliskan et al., 2017; Nadeem et al., 2020).

This study focuses on bias transfer within transformer-based models, specifically fine-tuned RoBERTa models, to analyze how training data imbalances can influence cultural representations at the embedding level. RoBERTa provides a manageable and interpretable approach for examining embedding shifts compared to larger LLMs, making it particularly useful for this analysis. Furthermore, its architecture allows us to focus on specific biases in the embedding space, which is more challenging to isolate in larger models (Radford et al., 2019).

Unlike previous work that focuses mainly on output biases or translation accuracy (Lu et al., 2020; Stanovsky et al., 2019), this study looks at the semantic and cultural shifts in the embedding space itself, offering insights into the mechanisms of bias transfer. These insights aim to inform the development of fairer and more inclusive NLP systems that better represent underrepresented cultures and languages, particularly those of indigenous communities.

2 Related Work

2.1 Bias in Multicultural Embedding Spaces

Transformer-based models often encode biases that reflect dominant cultural norms, particularly when trained on culturally imbalanced datasets. These biases manifest themselves in embedding spaces, leading to culturally inappropriate or misaligned representations of underrepresented groups. Caliskan et al. (2017) and Nadeem et al. (2020) demonstrate how societal and cultural biases in training data translate into model embeddings, which impact downstream tasks. However, much of this work focuses on linguistic or gender bias, with limited attention to cultural biases that affect underrepresented groups, such as Indigenous populations. Studies such as Mehrabi et al. (2021) highlight the need for methods that address such nuanced biases in model training.

2.2 Knowledge Transfer in Transformers

Knowledge transfer is a central mechanism in transformer models, enabling them to leverage patterns from high-resource domains to improve performance in low-resource contexts. However, this process also risks transferring undesirable biases. Raffel et al. (2020) explore transfer in the context of large transformer models, while Lauscher et al. (2020) analyze how multilingual models handle linguistic diversity. Fewer studies have examined how knowledge transfer impacts cultural representation, particularly when dominant culture training data disproportionately influence embeddings for underrepresented groups. This gap underscores the importance of investigating embedding shifts caused by imbalanced data, as explored in this study.

2.3 Bias Detection and Mitigation Techniques

Bias detection techniques, such as WEAT (Word Embedding Association Test) (Caliskan et al., 2017), have been widely used to quantify biases in embedding spaces. Recent extensions include metrics like StereoSet (Nadeem et al., 2020) and CrowS-Pairs (Nangia et al., 2020), which measure bias across a range of contexts. On the mitigation side, methods like SentenceDebias and Iterative Nullspace Projection (Bolukbasi et al., 2016) aim to neutralize bias in embeddings. However, these techniques primarily target gender or racial bias.

2.4 Embedding Space Analysis for Cultural Representations

Embedding space analysis has proven critical for understanding how models encode linguistic and cultural information. For example, Schnabel et al. (2015) demonstrate how evaluation methods can reveal the sensitivity of embedding models to variations in training data, emphasizing the need for consistent evaluation approaches. Antoniak and Mimno (2018) highlight that embeddings can vary significantly based on small corpus changes, suggesting that cultural and semantic shifts in embeddings must be interpreted carefully. Similarly, Mimno and Thompson (2017) investigate the geometry of embedding spaces, showing how the structural properties of embeddings may introduce distortions in representing nuanced cultural contexts. For low-resource languages, preserving cultural values in embedding consistency is critical to preserving cultural integrity (Le et al., 2023). Mimno and Thompson (2017), emphasize the role of embedding consistency in maintaining linguistic and cultural integrity. This study builds on these works by analyzing embedding shifts to uncover how cultural biases propagate during fine-tuning.

3 Datasets

3.1 Nunavut Hansard (Joanis et al., 2020)

The *Nunavut Hansard* dataset focuses on the proceedings of the Legislative Assembly of Nunavut, offering a rich parallel corpus in Inuktitut and English. As one of the largest available corpora for an Indigenous language paired with English, it presents a unique window into province-wide issues affecting the Inuit. Each debate segment is aligned and translated in English at the sentence level; in this study, we only utilize the English translation of the original Inuktitut discourse.

3.2 Canadian Hansard (Beelen et al., 2017)

The *Canadian Hansard* dataset is a comprehensive collection of transcribed bilingual parliamentary proceedings from the House of Commons of Canada, spanning over a century and capturing a broad range of political discourse. Each session's proceedings are time-stamped, carefully aligned by speaker turns, and accompanied by meta-information such as party affiliations, roles (e.g., Speaker, Member of Parliament, Minister), and session identifiers.

4 Methodology

4.1 Variable Representativity

Five RoBERTa models have been benchmarked and their embedding spaces analyzed. The first is the pretrained ”roberta-base” from HuggingFace, the second is fine-tuned on the entirety of the Nunavut Hansard corpus, and the third is fine-tuned on the entirety of the Canadian Hansard corpus. The remaining two models are trained on mixture datasets: the *Imbalanced Multilingual* uses 80% Canadian Hansard data and 20% Nunavut Hansard data, and the *Balanced Multilingual* uses 50% Canadian Hansard data and 50% Nunavut Hansard data. Although all five models are unilingual and anglophone, their naming suggests multilingualism due to their modeling of translated Inuktitut discourse and their capacity to serve as proxies for investigating bias transfer in linguistic representations concerning the Inuit of Nunavut and the First Nations of Canada. To achieve temporal alignment and ensure that mixture corpora represent comparable political and cultural contexts, texts are sampled from overlapping years (i.e., 1999–2017) defined by the longitudinally shorter Nunavut Hansard corpora. This alignment involves identifying and filtering the century-long Canadian Hansard corpus' CSV files by target years, validating speech records against timestamps for temporal consistency, and thereby removing diachronic biases that would arise from comparing different historical periods.

4.2 Pre-processing

For the Canadian Hansard corpus, CSV files corresponding to targeted years are recursively extracted, combined, and stored in a single file. The Nunavut Hansard corpus undergoes line-by-line cleaning, removing blank entries and consolidating English texts alongside their metadata into a unified preprocessed file. Once both corpora are prepared, we employ a pretrained "roberta-base" tokenizer for subword segmentation. Each sentence is tokenized with a maximum sequence length of 512, using padding and truncation to maintain a consistent input format. This ensures that both Canadian and Nunavut Hansard data are transformed into a standardized representation suitable for fine-tuning language models and conducting subsequent bias evaluations.

4.3 Fine-Tuning

In order to fine-tune an MLM model to validate our hypothesis, we employ a RoBERTa base model initialized from the “roberta-base” checkpoint and an A100 Tensor Core GPU from NVIDIA. We utilize a masked language modeling (MLM) objective where 15% of the tokens are randomly masked and the model is trained to predict these masked tokens. Specifically, we leverage the `DataCollatorForLanguageModeling` provided by the HuggingFace Transformers library, ensuring that masking is applied dynamically to the input sequences. We adopt the default AdamW optimizer with a learning rate of 2×10^{-5} and set the batch size to 8 sequences per device. The model is trained for 3 epochs with a weight decay of 0.01 and mixed precision floating-point operations (FP16) to speed up convergence while maintaining numerical stability. During training, we do not perform explicit validation steps (i.e., no evaluation strategy is selected) and save checkpoints every 5000 optimization iterations to the output directory. After training, we save both the fine-tuned model weights and the tokenizer for future inference and downstream tasks. The full training script is available for replication.

4.4 Word Pair Cosine Similarity

This experiment measures the semantic and pragmatic proximity of word pairs in the embedding space. By calculating cosine similarity between socioeconomically charged word pairs such as "Healthcare" and "Access", this method evaluates how semantically close they are related and or used together in the training corpora. This metric provides objective indicators into the model's capacity for preserving or distorting various challenges faced by the Inuit of Nunavut which are captured in their discourse.

4.5 *Canadian Indigenous Representation and Bias Evaluation* (CIRBE)

Our first iteration of CIRBE is composed of 86 prompts with a RoBERTa masked token, where `<mask>` can be replaced by either one of two predefined answers. The favorable one validates Inuit perspectives on pressing societal issues the government must solve and the unfavorable one ignores and invalidates these very perspectives. During this preliminary study, we did not have access to an expert on indigenous studies which would have helped us choose the appropriate vocabulary in function of the inherently political nature of the discourse with which we have fine-tuned our models. To remedy this, we have created an initial set of 100 literature-based prompts which we have given the model fully fine-tuned on the Nunavut Hansard corpora to perform open-ended MLM on, giving us the five most probable completion tokens. When these tokens predicted by our artificial expert aligned with those we have chosen based on reliable and relevant literature, we refined our set of prompts to a total 86 which served our final experiments presented in the following sections. The CIRBE benchmark is composed of 10 literature-based categories that address critical systemic issues which directly affect the First Nations of Canada, some of them notably identified by the Government of Canada itself. Due to the 5-page limit of the present report, we will only provide detailed justification for the *Health Disparities* category which also motivated the `Healthcare <-> Neglect` and `Healthcare <-> Access` word pairs.

First Nations communities in Canada experience significant health disparities compared to non-Indigenous populations, manifesting in elevated rates of chronic diseases, mental health challenges, and reduced access to culturally appropriate care. These disparities are deeply rooted in the legacy of colonization, residential schools, and the consequent erosion of traditional lifestyles, languages, and community supports (Truth and of Canada, 2015). Systemic barriers include underfunded healthcare services, racism within medical institutions, and limited social determinants of health such as proper housing and clean water—further (Reading and Wien, 2009; Adelson, 2005). Despite efforts by public health agencies and Indigenous-led initiatives, substantial gaps remain in preventative care, health promotion, and the integration of traditional healing practices with Western medicine, highlighting the need for policy reforms and collaborative frameworks to achieve equitable health outcomes (Greenwood et al., 2015).

5 Discussion of Results

5.1 Embedding Space Analysis

For the following three pairs of charged words, one immediately notices the striking and polarizing differences in cosine similarity that smaller mixture models undergo as a result of under-sampling the original Nunavut and Canadian Hansard corpora. While the pre-trained, full Nunavut Hansard and full Canadian Hansard minimally flinch following fine-tuning, biases in proxy-multilingual models seem to intensify whereby exclusion of Inuit perspective allows for the resurgence of stereotypes not only implicitly, but also in generative process as our MLM experiments demonstrate.

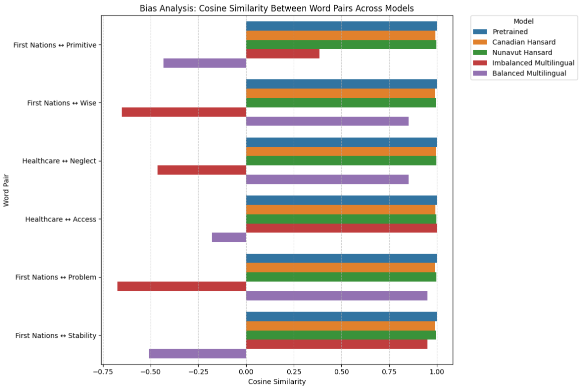

Figure 1: Cosine Similarity Between Word Pairs Across Models. This figure illustrates the variation in cosine similarity scores for selected word pairs across all models, highlighting disparities in word associations.

The imbalanced mixture model shows a positive association between "First Nations" and "primitive" while the balanced one with 30% more Nunavut data shows a polar opposite stronger negative association, reflecting the reinforcement of negative stereotypes which seem prevalent in the overrepresented fraction of its training data. Once again, the imbalanced mixture model aligns "First Nations" much less closely with "wise" than its balanced counterpart highlighting the representativity-dependent nature of the biases that arise. This pattern of polarization occurs in questions of healthcare access as well, where imbalanced models semantically deny association between "neglect" in and of "healthcare" while the balanced counterpart aligns with evidence of governmental reports aforementionned in section 4.5. Same goes for "access" which is semantically and pragmatically dissociated for the imbalanced model, highlighting the less frequent use of both in conjunction in Nunavut.

5.2 Open-ended Masked Language Modeling

When all RoBERTa models were prompted to complete the sentence "Indigenous women face `<mask>` risks compared to non-Indigenous women" of category *Violence Against Indigenous Women* with any word they wish, here were their most probable tokens and associated probabilities. We can clearly see that the more the Nunavut Hansard is included in mixture datasets, the more the fine-tuned models answer favorably by aligning with Governmental report evidence and admitting that Indigenous women indeed face higher risks of violence notably.

Table 1: Most Probable Tokens for Each Model

Model	Top Predictions
Nunavut Hansard	**different (0.2814)** **higher (0.1192)** greater (0.0696) more (0.0677) unique (0.0525)
Pretrained	**fewer (0.2643)** different (0.1434) greater (0.1045) more (0.0911) higher (0.0778)
Canadian Hansard	**fewer (0.2702)** different (0.2316) greater (0.0912) more (0.0649) similar (0.0452)
Imbalanced Multilingual	different (0.3726) greater (0.1143) **fewer (0.1042)** **higher (0.0845)** more (0.0566)
Balanced Multilingual	greater (0.1697) different (0.1615) **higher (0.1569)** more (0.1075) **fewer (0.1018)**

5.3 Binary Masked Language Modeling

Figure 2 demonstrates that the RoBERTa model fine-tuned on the full Nunavut Hansard corpus is disproportionately confident in its predictions of favorable answers to our prompts in comparison to all other models. The balanced mixture model comes in second with a drastically lower average probability for favorable answers.

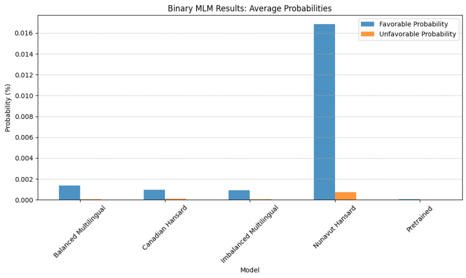

Figure 2: Average Probabilities for Favorable and Unfavorable Answers.

Figure 3 demonstrates that the less Nunavut Hansard data is included in the fine-tuning dataset, the lower the average probability of answering favorably to the 86 prompts of CIRBE.

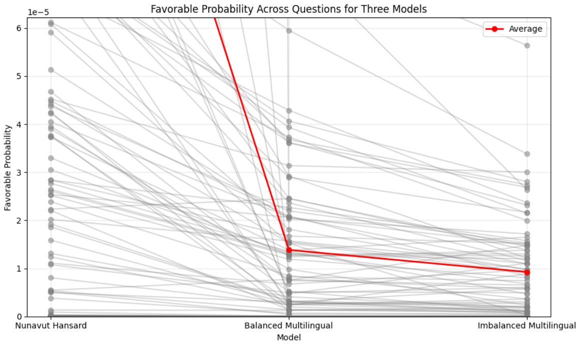

Figure 3: Favorable probability across questions for Nunavut Hansard, Balanced Multilingual, and Imbalanced Multilingual models.

Figure 4 demonstrates that the more Canadian Hansard data is included in the fine-tuning dataset, the higher the average probability of answering unfavorably to the 86 prompts of CIRBE.

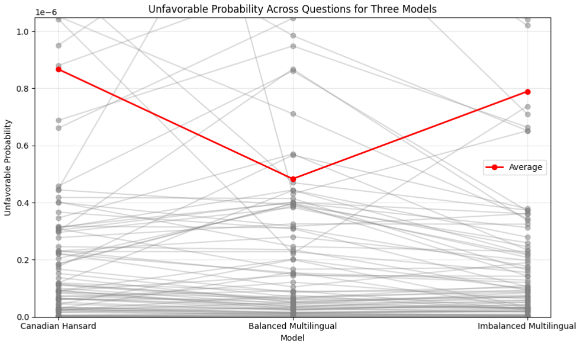

Figure 4: Unfavorable probability across questions for Canadian Hansard, Balanced Multilingual, and Imbalanced Multilingual models.

6 Conclusion

The inclusion of the Nunavut Hansard corpus in fine-tuning data of transformer models is essential to mitigating bias towards the First Nations of Canada and the Inuit of Nunavut.

7 Future Work

As CIRBE is not yet validated by human experts, we do not provide benchmarking accuracy for each model. The second iteration of CIRBE will be validated and larger, providing an open-source framework for ensuring safety of conversational agents before their deployment in our society.

8 Contributions

Jessica Ojo, Issa Saad and Yanis Bencheikh have all contributed to the benchmark's realization and writing of the present report. Model training and analysis has been operated entirely on Google Colab by Yanis Bencheikh. This research project has been directed and personally funded by Yanis Bencheikh who does not hold any conflicts of interest.

References

Naomi Adelson. 2005. The embodiment of inequity: Health disparities in aboriginal canada. *Canadian Journal of Public Health*, 96, Suppl 2:S45–S61.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Kaspar Beelen, Christopher Cochrane, Graeme Hirst, Maarten Marx, Nona Naderi, Ludovic Rheault, and Tanya Whyte. 2017. Heritage and the web of data. *Canadian Journal of Political Science/Revue canadienne de science politique*, 50(3):849–864.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Margo Greenwood, Sarah de Leeuw, Charlotte Reading, and Nicole M. Lindsay, editors. 2015. *Determinants of Indigenous Peoples' Health in Canada: Beyond the Social*. Canadian Scholars' Press, Toronto, Canada.

Eric Joanis, Rebecca Knowles, Roland Kuhn, and other collaborators. 2020. The nunavut hansard inuktitut–english parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2521–2529.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Ngoc Tan Le, Ikram Kasdi, and Fatiha Sadat. 2023. Towards the first named entity recognition of inuktitut for an improved machine translation. In *Proceedings of the Third Workshop on AmericasNLP*, pages 1–12.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday*, pages 189–202.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878, Copenhagen, Denmark. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Trenton Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Charlotte Reading and Fred Wien. 2009. Health inequalities and social determinants of aboriginal peoples' health. Accessed 2024-12-14.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.

Truth and Reconciliation Commission of Canada. 2015. Honouring the truth, reconciling for the future: Summary of the final report of the truth and reconciliation commission of canada. Accessed 2024-12-14.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.