**Analyzing Bias Transfer in Multilingual Transformer Models Using the Canadian and Nunavut Hansard Corpora**

**Collaborators**: Jessica Ojo, Abdullah Saad, Yanis Bencheikh

**Abstract**: This project investigates the transfer of bias from dominant languages (i.e., English) to underrepresented languages (i.e., Inuktitut) in multilingual transformer models. Using the Canadian Hansard and Nunavut Hansard corpora, we aim to analyze how cross-linguistic biases manifest in trained language models and identify potential societal impacts of these findings.

**Introduction**: Multilingual language models often reflect and propagate biases inherent in the training data. This project seeks to understand how biases from overrepresented languages transfer to underrepresented ones, using the Canadian Hansard (1901-2019) as the primary anglophone corpus and the Nunavut Hansard as the underrepresented Indigenous counterpart, which is the largest Inuktitut corpus aligned to English. The project’s primary hypothesis is that overrepresented languages impart biases into the multilingual models' representation of underrepresented languages, influencing their embeddings and output semantics. We propose training from scratch three separate but architecturally and parametrically identical transformer models: one on the Canadian Hansard, one on the Nunavut Hansard, and one multilingual model on training data with an inherent overrepresentation of English and underrepresentation of Inuktitut. This setup will help us draw concrete comparisons and insights into how biases may transfer and propagate.

**Related Work**: Several studies have explored bias in monolingual and multilingual language models, particularly focusing on gender, racial, and cultural biases. Research using techniques like the Word Embedding Association Test (WEAT) has been pivotal in quantifying bias in language models. However, few studies specifically examine bias transfer between overrepresented and underrepresented languages, especially with Indigenous languages of the First Nations of Canada such as Inuktitut. Existing literature also emphasizes the complexity and challenge of training models that fairly represent low-resource languages while maintaining their cultural and semantic integrity. This gap in the research underlines the significance of our study.

**Methodology**:

*Data Preparation:*

*Canadian Hansard (1901-2019):* This corpus is an extensive collection of transcribed debates from the Canadian Parliament, available in English and French. It represents formal legislative discourse over a long period, capturing the evolution of political, cultural, and social language. The dataset is valuable for analyzing how bias and cultural norms manifest in a dominant language over time.

*Nunavut Hansard:* This is the official record of the Nunavut Legislative Assembly, comprising debates in both English and Inuktitut. It uniquely reflects Indigenous perspectives and integrates cultural and societal discussions specific to the Inuit population. As the largest aligned English-Inuktitut corpus, it provides a crucial comparative basis for understanding bias in the context of underrepresented languages.

*Justification for Joint Use:* These two datasets were chosen due to their shared legislative context and high-quality, authoritative nature, allowing for a fair comparison of language usage and bias representation. By leveraging the Canadian Hansard for English and the Nunavut Hansard for Inuktitut, we can directly analyze bias transfer from an overrepresented language to an underrepresented one within the same domain of political discourse.

*Temporal Alignment:* To ensure that we are comparing equivalent cultural and political states, we will align data samples from identical time spans within both corpora. This approach eliminates potential critiques related to temporal discrepancies that could arise from comparing different historical periods, ensuring a robust analysis of cross-linguistic bias at specific epochs.

*Model Training:*

Model 1: Trained exclusively on the Canadian Hansard, serving as a baseline to capture biases inherent in the English-language parliamentary discussions.

Model 2: Trained solely on the Nunavut Hansard, providing insights into the independent representation of Inuktitut without external linguistic influence.

Model 3: A multilingual model trained on a combined dataset with an overrepresentation of English (e.g., 80% Canadian Hansard) and an underrepresentation of Inuktitut (e.g., 20% Nunavut Hansard). This model simulates real-world training conditions in multilingual settings where dominant languages influence the representations of less-represented ones.

*Technical Implementation:* We will use a consistent transformer architecture with identical hyperparameters across models to ensure comparability. Training will be performed on Colab Pro for computational feasibility.

*Embedding Space Analysis:*

Extract embeddings for a curated set of culturally significant and potentially biased terms (e.g., “leader” and “doctor”) across all three models. Additionally, use cosine similarity to measure the proximity between embeddings, assessing how associations shift between monolingual and multilingual models.

*Dimensionality Reduction and Visualization:*

Apply t-SNE and PCA to create visual representations of the embedding space, highlighting clustering patterns and semantic shifts indicative of bias transfer.

*Bias Metrics:*

Implement WEAT (Word Embedding Association Test) to quantify biases associated with gender, profession, and cultural roles in the embedding space. Additionally, use cosine and Euclidean distance metrics to evaluate how the representation of terms shifts between the Canadian Hansard and the Nunavut Hansard, identifying bias transfer in the multilingual model.

*Cross-Language Evaluations:*

Compare how culturally specific terms in the Nunavut Hansard model maintain their positions or shift in the multilingual model, assessing whether they align more closely with the English-biased terms.

Analyze semantic shifts to determine if Western-centric or gendered biases in the English corpus influence Inuktitut representations.

*Tools and Frameworks:*

Implement training and analysis using PyTorch or TensorFlow for the transformer models, with Hugging Face’s Transformers library for training and embedding extraction.

Use Scikit-learn for dimensionality reduction and matplotlib/Seaborn for visualizations.

**Expected Results**:

We anticipate identifying measurable bias transfer from English to Inuktitut, reflected in shifts within the embedding space of the multilingual model. Words and phrases originally in Inuktitut may show altered associations that align more closely with biases seen in the English corpus. This project aims to illuminate how training data composition impacts bias propagation and representation fairness in low-resource languages.

**Discussion and Conclusion**:

The study will provide insights into cross-lingual representation and potential bias transfer in multilingual models. This analysis will shed light on broader societal impacts, illustrating how biases inherent in dominant languages can influence the digital representation of underrepresented cultures. The findings could serve as a basis for developing more equitable language models and informing future strategies for training multilingual systems that respect linguistic and cultural diversity.

**Statement of Contributions**:

This paper and project are led by Yanis Bencheikh, with collaborators Jessica Ojo and Abdullah Saad. Each team member will actively contribute to data preprocessing, model training, embedding analysis, and report writing. Specific responsibilities will be equitably distributed to ensure shared ownership and comprehensive participation in all aspects of the project.