Week-1: My Research

Mitigating Bias in Native Language Identification Using Large Language Models and Open-Set Recognition

Introduction

Language reflects traces of our linguistic background — even when we write in a second language. My research explores this fascinating connection through the task of **Native Language Identification (NLI)**, which aims to predict an author's first language (L1) based on their writing in another language (L2).

While early studies relied on formal learner essays such as the **TOEFL11** corpus (Tetreault et al., 2013), these datasets capture only controlled, classroom-like writing. In contrast, today's online communication is full of informal expressions, emojis, and cultural slang. My work therefore focuses on **user-generated content (UGC)** — specifically the **Reddit-L2** dataset (Rabinovich et al., 2018) — to study NLI "in the wild," where text is messy but authentically human.

Research Aim

The main goal of my project is to develop a **hybrid NLI model** that is:

1. **Accurate** on known L1s,
2. **Fair**, by minimizing topic and cultural bias, and
3. **Generalizable**, capable of detecting when it encounters an unseen L1 (open-set recognition).

This will be achieved by combining **Large Language Model (LLM)** embeddings (e.g., BERT, RoBERTa) with **topic-debiased training** and **open-set recognition** strategies — a combination not yet systematically studied in NLI.

Motivation

Social media text provides an unparalleled window into natural language use, but it also introduces **bias**. For example, if many German speakers discuss technology on Reddit, a naïve classifier might simply learn to associate tech terms with "German," confusing language with topic. Such **spurious correlations** distort linguistic conclusions and risk unfair predictions (Kumar et al., 2019).

My research addresses this by applying techniques such as:

1. **Named-Entity Masking (NEM)** – removing proper nouns that leak topical clues,
2. **Linear Concept Erasure (LCE)** – deleting embedding dimensions tied to topics (Yaghoobzadeh et al., 2024), and
3. **Adversarial training** – forcing the model to "forget" topic signals.

In addition, **open-set recognition** mechanisms will help the system identify when a text comes from an L1 unseen during training, preventing overconfident but wrong classifications (Zhan et al., 2021).

Research Questions

The central question guiding this study is:

> *How does integrating LLM embeddings with topic-debiased training and open-set recognition affect the accuracy, fairness, and generalization of NLI models on user-generated content?*

Supporting questions examine:

- Whether hybrid models outperform traditional and zero-shot LLM baselines,
- How effectively debiasing reduces topic

leakage, and
- Whether open-set recognition improves robustness to unseen languages.

Methodology Overview

The research follows a **comparative experimental design** using:

- **Datasets**: Reddit-L2 (primary) and TOEFL11 (secondary benchmark if accessible).
- **Model variants**:
  - Traditional baselines (n-grams + POS features),
  - Zero-shot LLMs (prompt-based GPT),
  - Hybrid models (LLM + debiasing), and
  - Hybrid + Open-set recognition.
- **Evaluation metrics**: Accuracy, macro-F1, bias-leakage correlation, and false-positive rate on unseen L1s.

All experiments will use **PyTorch** and **Hugging Face Transformers** on GPU-based infrastructure to ensure reproducibility and scalability.

Ethical & Professional Considerations

Because NLI can reveal sensitive identity cues, ethical responsibility is crucial. The project aligns with **responsible AI** principles by:

- Using **anonymized, publicly available data** (Reddit-L2),
- **Avoiding profiling or surveillance** applications,
- Measuring and reporting **fairness metrics** alongside accuracy, and

- Ensuring transparency through open, documented code.

Expected Contribution

This study aims to deliver:

- A **fair and interpretable NLI framework** suitable for real-world text,
- Empirical evidence on **debiasing and open-set recognition** in LLM-based NLI, and
- Broader insights into balancing **performance, fairness, and generalization** in NLP research.

Ultimately, I hope this research will contribute to developing **equitable language technologies** that understand human diversity without reinforcing bias.

References

1. Blanchard D. et al. (2013) *TOEFL11: A corpus of non-native English*. ETS Research Report Series.
2. Kumar S. et al. (2019) *Topics to avoid: Demoting latent confounds in text classification*. EMNLP-IJCNLP.
3. Rabinovich E. et al. (2018) *Native language cognate effects in L2 English lexical choice*. TACL 6.
4. Tetreault J. et al. (2013) *The NLI Shared Task 2013*. Workshop on Innovative Use of NLP for Building Educational Applications.
5. Yaghoobzadeh Y. et al. (2024) *Deconfounding text embeddings via linear concept erasure*. arXiv 2403.05025.
6. Zhan X. et al. (2021) *A joint learning framework for open-set domain adaptation*. IEEE TPAMI.