Week-4: Evaluating Fairness and Generalization in Native Language Identification

Evaluating Fairness and Generalization in Native Language Identification

An Evaluation perspective of NLI model development

Introduction

Native Language Identification (NLI) seeks to infer an author's first language (L1) from their writing in a second language (L2). While earlier studies reported strong performance on curated learner corpora, contemporary deployments confront a markedly different landscape: *user-generated content* (UGC) that is informal, topical, and noisy. In such settings, conventional accuracy metrics can obscure a critical issue — models may succeed by exploiting *spurious topical cues* rather than genuine cross-linguistic transfer. This blog post explains the evaluation framework that treats **performance**, **fairness**, and **generalization** as co-equal objectives, with explicit tests for topic leakage and mechanisms for rejecting unseen languages.

The practical question is not only "How accurate is the model?"

but *“Accurate on what basis, and under what distributional shifts?”* We therefore emphasize (i) cross-topic evaluation to decouple linguistic signal from domain content, (ii) **bias-leakage auditing** to quantify spurious correlations, and (iii) **open-set recognition** so the system can state “unknown” when confronted with L1s absent from training. Together, these components support trustworthy NLI suitable for research and pedagogical use.

Background and Problem Context

UGC such as Reddit comments departs from formal essays along three axes: linguistic variability (slang, emojis, misspellings), topical diversity (thousands of communities), and spontaneity (conversation over composition). These properties introduce confounds: if certain L1 groups participate disproportionately in specific topics, a classifier may associate topic vocabulary with L1 rather than detecting subtle morpho-syntactic traces. Large language models (LLMs) intensify this risk because their semantic priors readily encode topical and cultural knowledge that can overshadow the weaker signals of L1 transfer.

> **Problem Statement**
>
> How can we evaluate NLI models on UGC such that reported gains reflect *linguistic competence* rather than *topic memorization*, while ensuring models remain reliable when encountering previously unseen L1s?

Evaluation Objectives

- **Performance:** Achieve high Accuracy and Macro-F1 with balanced

per-class performance.

- **Fairness:** Minimize topic leakage and reliance on named entities; preserve genuine linguistic features.
- **Generalization:** Calibrate an *unknown* option via open-set methods; maintain confidence only where justified.

Dimension	Primary Metrics	Interpretation
Performance	Accuracy, Macro-F1, per-class F1	Overall correctness and balance; avoids majority-class inflation.
Fairness	Bias-Leakage Score; ablation delta after NE-masking / concept erasure	Lower is better; reductions with minimal F1 loss indicate effective debiasing.
Generalization	Open-set FPR, AUROC for reject, calibration error	Trustworthy “unknown” decisions; calibrated confidence on known classes.

Methodology and Experimental Design

We compare four model families under shared preprocessing, seeds, and controlled splits:

1. *Traditional baselines* with character/word/POS n-grams and linear classifiers;
2. *Zero-shot LLMs* via prompts to measure out-of-the-box semantic competence;
3. *Hybrid LLM embeddings + debiasing* using Named Entity Masking (NEM) and linear concept erasure; and
4. *Hybrid + open-set* with calibrated thresholds or distance-based rejection (e.g., Mahalanobis in

embedding space).

To surface topic reliance, we run both **in-domain** (same subreddit distribution) and **cross-topic** evaluations (train/test on disjoint topic sets). Ablation studies quantify the marginal effect of NEM and concept erasure on fairness and performance. All experiments are replicated across multiple seeds for reliability.

> **Open-Set Calibration**
>
> We construct pseudo-unknowns by withholding one or more L1s during training and tuning rejection thresholds on a validation set. We monitor ROC/PR curves and expected calibration error to select operating points that minimize false positives on unknowns while preserving accuracy on known classes.

Debiasing Techniques

Named Entity Masking (NEM). Replacing person, location, and organization names with placeholders reduces direct leakage from culturally specific references that correlate with L1 communities. NEM is applied consistently at train/validation/test time to avoid distributional shifts.

Linear Concept Erasure. We identify embedding directions most correlated with topic proxies and remove them through linear projection. The objective is to suppress topic semantics while preserving linguistic structure. We evaluate effectiveness by reporting changes in leakage and Macro-F1 before/after erasure.

Sanity Checks. Over-zealous debiasing can erase legitimate signal. We therefore inspect per-class F1, confusion patterns, and example-level errors to ensure linguistic cues remain intact.

Tools, Artefacts, and Reproducibility

We implement models in **PyTorch** with **HuggingFace Transformers** and evaluation in **scikit-learn**. Reproducibility is enforced via fixed seeds, versioned configuration files, and a *run registry* (CSV/Excel) that logs dataset slices, hyper-parameters, metrics, and commit hashes. Visualization with Matplotlib summarizes the fairness–performance frontier to support model selection.

Standard Artefacts

- **Run Registry:** IDs, seeds, splits, configs, metrics, commit hash.
- **Config Files:** preprocessing and model parameters (JSON/YAML).
- **Evaluation Sheets:** aggregated tables, per-class reports, confusion matrices.
- **Ablation Logs:** before/after NEM and concept erasure, with leakage deltas.

Ethical and Quality Considerations

Ethical evaluation runs parallel to technical assessment. UGC must be handled to respect privacy and licensing; models should report limitations and avoid profiling. We include an *ethics checklist* per accepted result: data provenance, PII removal, license compliance, fairness metrics alongside accuracy, and a statement of intended use. Where a technique improves accuracy but worsens leakage, results are not accepted without a justified trade-off analysis.

- **Bias Mitigation:** NEM and concept erasure validated by ablations.
- **Fairness Reporting:** leakage shown side-by-side with Accuracy/Macro-F1.
- **Transparency:** preprocessing, seeds, and configs documented for replication.
- **Safety:** open-set “reject” option to avoid confident misclassification of unseen L1s.

Threats to Validity and Limitations

Construct validity: mapping country flair to L1 is an imperfect proxy; sensitivity analyses test robustness to mislabeling.

Internal validity: cross-topic splits reduce confounds but cannot eliminate all correlations between L1 and domain.

External validity: results on Reddit may not transfer to other platforms; we therefore report assumptions and encourage cross-corpus replication.

Expected Contributions

- An evaluation protocol that balances performance with fairness and open-set reliability for NLI on UGC.
- A debiased hybrid modeling recipe combining LLM embeddings, NEM, and concept erasure.
- Reproducible artefacts (configs, run registry, ablation logs) enabling peer verification.

Outlook

Positioning fairness and generalization as first-class objectives reframes NLI from a leaderboard exercise into a responsibility-aware science of linguistic signal. The proposed evaluation design — cross-topic testing, leakage auditing, and open-set calibration — aims to produce models that are not only competitive but also credible and useful for downstream educational and research settings.

References

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). *TOEFL11: A corpus of non-native English*. ETS Research Report

Series.

Kumar, S., Wintner, S., Smith, N. A., & Tsvetkov, Y. (2019). Topics to avoid: Demoting latent confounds in text classification. *EMNLP-IJCNLP*.

Rabinovich, E., Ordan, N., & Wintner, S. (2018). Native language cognate effects in L2 English lexical choice. *TACL*.

Yaghoobzadeh, Y., Hertel, J., & Tsvetkov, Y. (2024). The medium is not the message: Deconfounding text embeddings via linear concept erasure. *arXiv:2403.05025*.

Zhang, W., & Salle, A. (2023). Native Language Identification with Large Language Models. *arXiv:2312.07819*.