# CMPUT 650 Project Proposal: Sense-Aware Multilingual Polarization Detection

**Mohammad Shahriar Hossain**
University of Alberta
mhossai6@ualberta.ca

**Ekamjot Singh**
University of Alberta
ekamjot2@ualberta.ca

## Abstract

We address POLAR@SemEval-2026: detecting online polarization and predicting its type and manifestation across languages. We target English (EN), Bengali (BN), and Punjabi (PA) with a compact two-path model: a native multilingual encoder and a translate-to-English encoder, combined by calibrated averaging. To reduce ambiguity around culturally loaded terms, we add a small document-level *sense summary* computed only for a short curated term list. We compare against a simple lexical baseline and evaluate with Macro–F1 on the official splits (pol, 2025).

## 1 Introduction

POLAR defines (T1) binary detection, (T2) type classification (political, racial/ethnic, religious, gender/sexual, other), and (T3) manifestation identification (stereotype, vilification, dehumanization, extreme language, lack of empathy, invalidation) (pol, 2025). We focus on EN/BN/PA to balance one high-resource and two lower-resource languages where meaning often hinges on culturally specific terms (e.g., *reservation*, *secular*). Accurate multilingual analysis supports monitoring, cross-region comparison, and policy evaluation.

## 2 Related Work

Prior work on adjacent phenomena shows the value of pretrained encoders with light auxiliary signals: HatEval (SemEval-2019 Task 5) (Basile et al., 2019), OffensEval 2019/2020 (Zampieri et al., 2019, 2020), Toxic Spans (Pavlopoulos et al., 2021), early feature studies (Waseem and Hovy, 2016; Davidson et al., 2017), large-scale abuse characterization (Founta et al., 2018), and rationale-driven datasets (Mathew et al., 2021). We follow this template but introduce a targeted sense summary and a two-path ensemble adapted to POLARs label space and EN/BN/PA.

## 3 Method

**Two paths.** *Native multilingual:* fine-tune a multilingual encoder (e.g., XLM-R$_{\text{base}}$) jointly on EN/BN/PA with three heads: binary (T1) and two multi-label heads (T2,T3). A single shared backbone encourages transfer while heads specialize per subtask. *Translate-to-English:* translate BN/PA to EN (short-text MT, e.g., Google Translate) and process with an English encoder (e.g., RoBERTa$_{\text{base}}$/DeBERTa-v3$_{\text{base}}$); filter obvious MT failures via length-ratio and language-ID checks.

**Targeted sense summary.** We curate ∼20–50 potentially ambiguous polarization-related terms per language (political/identity keywords). For a given document we: (i) detect occurrences (after normalization), (ii) assign a sense per occurrence using public lexical/sense resources—either a lightweight UKB/LMMS-style guess or simple *gloss matching* against 2–3 short sense glosses, and (iii) aggregate into a fixed-size vector (normalized sense counts and/or confidence-weighted proportions, plus a small confidence/entropy statistic). We cap this summary at $\leq$64 dimensions and concatenate it with the encoders pooled representation prior to each head. This injects disambiguation where it matters while avoiding full-coverage WSD.

**Fusion.** For each subtask we temperature-scale logits on the dev set (per-language if helpful) and *average* the two paths scores. Final decisions use per-label thresholds tuned on dev. No other fusion is used.

## 4 Experimental Setup

**Task & data.** We use the official POLAR@SemEval-2026 splits and labels; the primary metric is Macro–F1 per subtask and language (pol, 2025). We report EN/BN/PA and the macro-average.

**Inputs & outputs.** Input: a short post in EN/BN/PA. Outputs: (T1) binary polarization; (T2) multi-label type; (T3) multi-label manifestation. Example: They are ruining our traditions; we must stop them. ⇒ polarized; {political}; {vilification, extreme language}.

**Implementation.** *Encoders:* XLM-R$_{\text{base}}$ for the native path; RoBERTa$_{\text{base}}$/DeBERTa-v3$_{\text{base}}$ for EN. Max length 128–256; base-size checkpoints for efficiency. Mini-batches are language-balanced (EN:BN:PA) to stabilize training. *Sense summary:* fixed term lists (per language); per-occurrence sense from WordNet/BabelNet via UKB/LMMS-style inference or gloss matching; pooled to a small document vector; concatenated with the pooled encoder output before the heads. *Training:* T1 uses binary cross-entropy; T2/T3 use multi-label BCE with class weights (and light label smoothing). Thresholds are selected per label on the dev set to maximize Macro–F1 and then fixed for test. Early stopping on dev Macro–F1; dropout on heads and small weight decay. *Calibration:* temperature scaling on dev (optionally per-language) prior to ensembling; translation outliers are dropped by simple length-ratio/LID checks. *Resources:* official POLAR data, open encoder checkpoints, and free MT; no LLMs or paid APIs are required.

**Comparisons & ablations.** We compare: (a) the lexical baseline (below), (b) native-only, (c) translate-to-EN-only, (d) ensemble, and (e) ensemble without the sense summary. If compute permits, we also report small optional ablations: a compact socio-linguistic indicator block (pronoun ratios, negation, intensifiers, basic toxicity/emotion counts), light back-translation (EN↔BN/PA) for minority labels, and brief in-language domain adaptation.

## 5 Baseline

*Translate-to-English + Linear models.* BN/PA → EN; extract TF–IDF word/character $n$-grams; train a linear SVM for T1 and one-vs-rest logistic regression for T2/T3. This transparent baseline mirrors effective starting points in related shared tasks (Basile et al., 2019; Zampieri et al., 2019, 2020; Pavlopoulos et al., 2021).

## References

2025. Polar @ semeval-2026. https://polar-semeval.github.io/. Github.io.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 12, pages 491–500.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*.