University of Bonn
CAISA lab

Introduction to Natural Language Processing

(Winter Semester 2025/2026)

Project Proposal of Team # 13

Group Members:

Saad Rahman	50352225
Tharanginee Elango	50353561
Xuefei Ren	50328158
Maria Azrar	50350333
Eitan Hasson	50348307

November 19, 2025

Project Overview: POLAR

1. Motivation
2. Dataset
3. Methodology
4. Expected Results
5. Evaluation Metrics
6. Challenges and Limitations

1. Motivation

Social media text can range from a variety of themes and opinions, as people can write whatever is on their minds, many times, without filter, the text itself will contain a polarized attitude. This refers to the phenomenon where opinions, beliefs, or behaviours are expressed in an extreme way, leading to conflicts because they showed a level of stereotyping, vilification, dehumanization, or intolerance regarding the beliefs, views, or identities of other people.

In this project we will work with LLM's to develop a model capable to detect polarization (Subtask 1) in English messages from multiple online forums (such as Twitter, Facebook, Reddit, etc.) and, if it is a polarized message, be able to classify the type/target of polarization (Subtask 2) within the following classes (multiple classes are possible):

- **Political/ideological polarization**: The text presents division, intolerance, and conflict between political parties and followers.
- **Racial or ethnic polarization**: The text focuses on ethnic origin or racial origin and incites division, intolerance, and conflict between ethnic groups or races.
- **Religious polarization**: The text focuses on religious identity and incites division, intolerance, and conflict between religious followers.
- **Gender and Sexual orientation polarization**: The text presents exclusion, discrimination, and marginalization of individuals based on their gender and/or sexual orientation in society.
- **Other**: The text targets other groups/identities such as economy, technology, media, etc.

With this in mind, we can write down the research question as follows:
RQ: **How effectively can large language models combined with NLP principles, detect and categorize different types of polarization in English messages from online forums?**

2. Dataset

This project utilizes the POLAR (Detecting Multilingual, Multicultural and Multievent Online Polarization) benchmark, focusing on adopting its English subset for model training and evaluation.

2..1 General Structure and Quality

The data source is the POLAR Benchmark, which provides text data from diverse online platforms and real-world events including news websites, Reddit, blogs, Bluesky, and regional forums, covering events such as elections, conflicts, gender rights, migration, and more. The training data contains a shared pool of 2,670 English instances for both Subtask 1 and Subtask 2. The development set consists of 133 unlabeled instances, which are used for prediction generation. All training instances are fully labeled for both subtasks.

2..2 Text Complexity and Basic Statistics

We conducted an initial data exploration to calculate basic statistics regarding the text structure and complexity, which informs our preprocessing strategy.

- **Total Instances:** The 2,670 instances provide a moderately-sized dataset foundation.
- **Mean Word Count:** Instances contain an average of **12.38** words, with a standard deviation of **8.45**. This high variability requires robust tokenization and padding strategies.
- **Vocabulary Size:** The total vocabulary size is **7,746** unique words. Despite using only the English subset, the overall design of POLAR as a multilingual benchmark motivates the use of multilingual pre-trained models. The vocabulary size and the multilingual nature of the overall benchmark underscore the necessity of leveraging powerful pre-trained language models, such as **XLM-RoBERTa**.
- **Content Exploration:** The prominence of higher frequency words (stop words) such as 'the' and 'to' suggests that standard NLP preprocessing will be necessary for certain feature engineering tasks.

2..3 Subtask 1: Polarization Detection

Polarization detection classifies text as Polarized (1) or Not Polarized (0). Analysis of the training set reveals a significant class imbalance:

- Not Polarized (0) instances account for **1,674** ($\approx 62.70\%$).
- Polarized (1) instances account for **996** ($\approx 37.30\%$).

This imbalance suggests that incorporating class-weighted loss may help mitigate bias toward the majority class.

2..4 Subtask 2 : Polarization Type Classification

Subtask 2 is a highly complex **multi-label** classification task, where each polarized instance is required to predict across five defined polarization types: political, racial/ethnic, religious, gender/sexual, and other.

Label Distribution and Severe Imbalance

- The **Political** type is highly dominant (**989** instances).
- The **Gender/Sexual** type is the least frequent, with only **65** texts.

This yields a 15:1 imbalance ratio between the most and least frequent labels, making it challenging for the model to achieve strong macro-level performance without targeted loss weighting or sampling strategies.

Multi-Label Complexity

The dataset also demonstrates substantial label co-occurrence:

- **Multi-Label Instances:** Out of all 996 polarized instances, 408 (40.96%) are annotated with two or more polarization types.

This corresponds to a relatively high label cardinality (average number of labels per instance) and reflects that polarized discourse frequently spans multiple social dimensions rather than remaining isolated within a single category.

Conclusion:

The multi-label nature of the task, together with the extreme imbalance and substantial co-occurrence, requires using independent binary classification heads, sigmoid activation, and specialized optimization techniques (e.g., class-weighted or focal loss) to ensure balanced learning across all types.

2..5 Data Exploration Plan

To fulfill the data exploration requirement, we will perform the following steps:

- **Visualization:** We plan to create word clouds or frequency distribution plots for the polarized text to identify the most common words and their relative importance across different polarization types
- **Linguistic Structure Analysis:** We will perform Parts-of-Speech (POS) tagging to examine the linguistic structure of the text, looking for patterns (e.g., strong adjectives, specific pronouns) correlated with highly polarized discourse

These analyses aim to provide insights that guide preprocessing decisions, feature engineering, and later error analysis.

3. Methodology

3..1 Core Architecture

We use **XLM-RoBERTa (XLM-R)** as the base model due to its strong performance on cross-lingual understanding and transfer tasks. Compared to earlier multilingual models such as mBERT, XLM-R consistently achieves higher accuracy across a range of multilingual benchmarks[1], making it reliable for this task.

XLM-R provides two major advantages:

- **Extensive Pre-training:** The model is trained on 2.5 TB of filtered CommonCrawl data spanning 100 languages[1], [2]. This broad and diverse pre-training corpus helps XLM-R capture linguistic patterns necessary for handling the 22 languages included in this challenge.
- **Effective Cross-Lingual Transfer:** XLM-R employs a shared **Byte-Pair Encoding (BPE)** vocabulary across all languages. This shared representation enables efficient knowledge transfer and strong **zero-shot performance**, which is especially useful for low-resource languages with limited labeled data.

3..2 Fine-Tuning Strategy for Subtasks

We fine-tune XLM-R using the Hugging Face `transformers` library with the PyTorch framework, adding task-specific classification heads for each subtask.

Subtask 1: Polarization Detection (Binary Classification)

- **Model Setup:** XLM-R is fine-tuned with a binary classification head (`num_labels = 2`) to predict whether a text instance is *Polarized* or *Not Polarized* [3].
- **Loss Function:** The model is optimized using the standard **Binary Cross-Entropy Loss**.

Subtask 2: Polarization Type Classification (Multi-Class Classification)

- **Model Setup:** For predicting specific polarization types, XLM-R is fine-tuned with a multi-class classification head (`num_labels = N`, where N is the number of polarization categories) [3].
- **Expected Challenge:** Due to substantial class imbalance, standard training is likely to favor majority classes and produce weaker performance on minority categories. This imbalance typically lowers the **macro F1-score**, which treats all classes equally.
- **Mitigating Class Imbalance: Class-Weighted Loss** To address this issue, we apply **class-weighted cross-entropy**. Each class c is assigned a weight inversely proportional to its frequency in the training set:

$$W_c = \frac{1}{\text{frequency of class } c}$$

  These weights are passed to the PyTorch `nn.CrossEntropyLoss`, encouraging the model to pay more attention to underrepresented classes and improving balanced performance across categories[4].

4. Expected Results

As seen in the dataset analysis, it contains clear biases, especially noticeable in political content. This imbalance is a condition of current global events which increases the amount of political discussions on social platforms. Due to this, models trained on this data may also become more sensitive to political polarization, while under performing on the other polarizations to study due to their less amount on the dataset. We expect that this bias affects both subtasks, allowing for better performance on politically oriented messages and limiting the generalization ability of the model among the other classes.

To the date of writing this work, there's been a couple of submissions regarding the tasks. For subtask 1, the current F1-macro scores range between 0.7039 to 0.8902 (without considering an outlier at 0.3756). With this in mind, we expect to reach performance values for subtask 1 close to the range, regarding our work with the data processing and model training done. For subtask 2, the task seems more challenging, due to the multiple classes to describe the text and the imbalance in them. For this subtask, current submissions range between 0.1943 to 0.4894 for the F1-macro

score metric. These results are much lower than the previous subtask, so we expect to be able to manage reaching the ranging values by working on strategies which consider the class imbalance.

Overall, while we expect to be able to reach existing submissions results, we need to remain open to unexpected outcomes, as the training process needs to be carefully considered so that all classes are similarly represented. There is no definitive answer, as the methodology and/or objectives might change while working on it.

4..1 Other Studies results

In addition to insights derived from the analysis of our dataset, findings from previous studies on related classification tasks suggest similar challenges and expected performance patterns. Research on hate speech detection, particularly in multimodal contexts, consistently highlights that models trained on text alone often struggle to capture contextual and nuanced signals of harmful or polarized content. For example, Boishakhi et al. [5] demonstrate that single-modality systems (e.g., text-only) tend to underperform compared to approaches that incorporate additional cues such as facial expressions or vocal tone. Their work shows that integrating multiple modalities leads to substantially more robust detection, even when using classical machine learning methods, reinforcing that linguistic signals alone may be insufficient for reliably capturing complex social phenomena such as hate or hostility.

Studies on text-based hate speech classification using traditional and deep learning models further emphasize this limitation. Warner and Hirschberg [6], as well as Badjatiya et al. [7], report that although high accuracy is achievable on well-balanced datasets, model performance drops significantly in the presence of label imbalance, domain shifts, or nuanced categories of abuse. Similar observations appear in racism and sexism detection tasks [8], where models exhibit strong performance on frequent categories but degrade on infrequent or context-dependent cases.

Based on these observations, we expect our models to reflect comparable behavior: strong performance on frequent, well-represented classes (e.g., political polarization) and reduced effectiveness on minority classes. Prior literature therefore supports the expectation that class imbalance and contextual subtlety will limit generalization, especially for Subtask 2, where multi-class prediction amplifies these challenges.

5. Evaluation Metrics

To assess the performance of the proposed models for polarization detection, we will use a set of evaluation metrics suitable for both binary and multi-label classification tasks. For **Task 1**, which involves binary classification (polarized vs. non-polarized), we will report standard metrics including *accuracy*, *precision*, *recall*, and the *F1-score*. In addition, we will include multiple variants of the F1 metric—namely the *F1-binary*, *F1-macro*, and *F1-micro* scores—to provide a more complete picture of the model's performance under potential class imbalance or asymmetric error costs. These metrics allow us to quantify not only the overall correctness of the model but also its ability to correctly identify polarized content while minimizing false alarms.

For **Task 2**, which is a multi-label classification problem involving the identification of specific types of polarization (e.g., political, racial, religious), we will employ metrics commonly used in multi-label learning. These include the *micro-averaged F1-score*, which aggregates contributions from all labels and is robust to label imbalance, and the *macro-averaged F1-score*, which treats all

labels equally by averaging performance across categories. We will also consider *Hamming loss* to measure the fraction of incorrectly predicted labels, and, where appropriate, the *Jaccard similarity coefficient* to quantify the overlap between predicted and true label sets. Together, these evaluation metrics will provide a comprehensive assessment of the model's effectiveness in detecting polarized content and accurately identifying its specific type.

Furthermore, we will utilize a **confusion matrix** in both tasks to analyze the distribution of true positives, true negatives, false positives, and false negatives, offering deeper insight into the types of errors the model may produce.

6. Challenges and Limitations

Developing models for polarization detection introduces several challenges that affect system performance, generalization, and fairness. These challenges arise from the complexity of polarization as a linguistic and socio-cultural phenomenon, as well as the technical constraints of building multi-label classifiers.

(i) Ambiguity and Subjectivity in Polarization Definition

Despite clear guidelines, polarization often appears implicitly through sarcasm, framing, insinuations, or cultural references. Machine learning models struggle with (i) subtle hostility that does not use explicit hate terms, (ii) distinguishing legitimate criticism from polarization, and (iii) interpreting text without access to broader conversation context. As a result, models may miss-classify borderline cases or miss the context altogether.

(ii) Multi-Label Complexity and Overlapping Categories

The task requires predicting political, racial/ethnic, religious, gender/sexual, and "other" polarization simultaneously. However, many texts involve multiple overlapping identities, boundaries between categories are often blurry, and the "Other" category is broad and noisy. This makes learning consistent category-specific decision boundaries challenging.

(iii) Code-Mixing, Dialects, and Informal Social Media Language

Social media text frequently contains code-switching (the practice of shifting between two or more languages, dialects etc.), slang, emojis, abbreviations, non-standard grammar and culturally specific polarization cues which are not captured by general multilingual transformers. Dialectal variation and informal writing styles reduce the robustness, accuracy of standard multilingual language models.

(iv) Dataset Imbalance and Evaluation Constraints (Macro F1)

The polarization categories are unevenly distributed: Out of the total number of sentences in the training data **(2677)**; political polarization **(996)** is common while gender/sexual **67** and "other" categories is rare **121**. Since evaluation uses Macro F1, misclassification of low-frequency labels disproportionately impacts the final score, complicating training and threshold selection.

(v) Lack of Context

Each text snippet is evaluated in isolation, but polarization often depends on wider conversational or political context, ongoing events, or references to specific individuals or groups. Without access to such context, models must infer polarization from incomplete information, which limits accuracy.

(vi) Risk of Overfitting to Lexical Patterns

Models may rely on shallow lexical cues (e.g., words related to government, ethnicity, or religion) instead of learning *true* polarization mechanisms. This increases false positives for neutral political or social discussions and reduces robustness to new phrasing or unseen events.

(vii) Ethical and Sociotechnical Considerations

Polarization is tightly connected to identity, politics, and conflict. Models risk amplifying biases present in the training data, disproportionately misclassifying minority speech as polarized, or underperforming on underrepresented linguistic varieties. Ensuring fairness and preventing harmful misclassification remain open challenges in this domain.

Bibliography

[1] A. Conneau, D. Kiela, H. Schwenk, A. Goyal, E. Koço, C. Service, J. M. Toldanov, T. Lu, V. Stoyanov, and P.-E. Stock, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[2] Common Crawl Foundation, "Common crawl," https://commoncrawl.org, 2024, accessed: 2025-02-18.

[3] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020, pp. 38–45.

[4] A. F. M. A. U. Islam, S. A. Shah, M. Z. Abedin, and M. G. Alam, "Addressing imbalance in multi-label classification using weighted cross entropy loss function," in *2020 2nd International Conference on Big Data, Artificial Intelligence and Computer Networks (ICBAIC)*. IEEE, 2020, pp. 101–105.

[5] F. T. Boishakhi, P. C. Shill, and M. G. R. Alam, "Multi-modal hate speech detection using machine learning," *arXiv preprint arXiv:2307.11519*, 2023.

[6] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proceedings of the Second International Conference on Social Informatics*, 2012.

[7] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 759–760.

[8] R. Kshirsagar, T. Alaoui, D. Boyd, and K. McKeown, "Detecting hate speech and offensive language using deep neural networks," *arXiv preprint arXiv:1809.10644*, 2018.