University of Bonn

CAISA LAB

Introduction to Natural Language Processing

(Winter Semester 2025/2026)

Project Proposal of Team 17

Detecting Multilingual, Multicultural and Multievent Online Polarization

Group Members:

Faruk Demirel	50430703
Mariam Elerian	50431544
Elene Esakia	50465174
George George	50443765
Jon Lumi	50465950

November 19, 2025

CAISA Lab
https://caisa-lab.github.io/

UNIVERSITÄT BONN

1. Motivation:

Online polarization has become one of the most pressing challenges in today's digital world. Before conversations turn into hate speech or misinformation, they often show early signs of division, hostility, or "us vs. them" framing. Being able to detect these early indicators is extremely valuable, as it allows platforms, researchers, and communities to take action before discussions escalate into harmful or extreme behavior.

Our project focuses on studying polarization in English and Arabic, two widely used languages that represent very different cultures, societies, and online environments. With global political tensions, highly emotional public debates, and major events like elections and conflicts, understanding how polarization emerges in different languages is more relevant than ever.

From a technical perspective, this task is challenging. It brings together several difficult NLP problems: binary classification for detecting polarized messages, multi-class and multi-label classification for identifying polarization types and manifestations, and the added complexity of working across multiple languages. Unlike hate speech detection or sentiment analysis—areas that are already well studied—polarization detection is still relatively unexplored. It focuses on divisive or inflammatory language before it becomes openly hateful, which makes it an important but understudied part of online discourse.

Finally, this type of work has meaningful, real-world impact. Models capable of recognizing early polarization can support healthier online communication, help researchers analyze political or social conflicts, and improve tools for digital literacy. By choosing English and Arabic, we also contribute to multilingual NLP development, bridging high-resource and low-resource contexts, and highlighting cultural differences in how polarization is expressed.

2. Dataset:

It's important to get a sense of your dataset's basic structure and quality. The data exploration step is crucial for understanding the structure, quality, and nuances of your dataset before building any models. In this section, you explore the characteristics of the text itself. Calculating basic statistics such as average sentence length, word count, and vocabulary size helps you understand the complexity and variability of your text data. Visualization plays a big role in exploring the data. Creating word clouds or frequency distribution plots is a great way to identify the most common words and their relative importance. You can also perform parts-of-speech (POS) tagging to look at the linguistic structure of your text.

The SemEval-2026 Task 9 dataset provides labeled social media posts across 13 languages for detecting and classifying online polarization. For our project, we focus on English and Arabic due to team familiarity and to compare performance between a high-resource language (English) and a medium-resource language (Arabic).

The dataset is organized into three subtasks, each with separate data files. Each subtask contains two folders: a train folder used for model training and a dev folder used for validation and hyperparameter tuning during training. The dev set helps prevent overfitting by allowing us to monitor model performance on unseen data before final testing. For Subtask 1, the English dataset contains 2676 training samples and 133 development samples, while the Arabic dataset contains 3380 training samples and 169 development samples. The Arabic dataset is notably larger, which may provide advantages during model training for that language.

Subtask 1 focuses on polarization detection as a binary classification problem. Each data file contains three columns: an id column with unique identifiers, a text column containing the social media post content, and a polarization column with binary labels where 0 indicates non-polarized content and 1 indicates polarized content.

Subtask 2 addresses polarization type classification as a multi-label problem. The data structure includes the same id and text columns as Subtask 1, but adds five additional binary columns representing different polarization types: political, racial/ethnic, religious, gender/sexual, and other. The dataset contains all posts from Subtask 1, but for non-polarized posts, all five type columns will be 0. For polarized posts, a single post can have multiple types simultaneously. For instance, a polarized post might be labeled as both political and racial/ethnic if it contains elements of both.

Subtask 3 focuses on manifestation identification, also structured as a multi-label problem. It contains the same posts as Subtasks 1 and 2 with the same id and text columns. However, it adds six manifestation columns: stereotype, vilification, dehumanization, extreme_language, lack_of_empathy, and invalidation. Similar to Subtask 2, non-polarized posts will have 0s across all manifestation columns, while polarized posts can have multiple manifestations marked as 1, allowing the model to capture the various ways polarization is expressed.

Understanding the dataset's structure and quality is crucial before building any models. We will perform comprehensive data exploration to analyze the characteristics of the text itself. This includes calculating basic statistics such as average sentence length, word count, and vocabulary size to understand the complexity and variability of the text data across both languages. We expect to find differences in text length and vocabulary between English and Arabic posts, as well as variations in writing style due to cultural differences in online discourse.

Visualization will play an important role in exploring the data. We plan to create word clouds and frequency distribution plots to identify the most common words and their relative importance in polarized versus non-polarized content. These visualizations will help us understand which linguistic features are most associated with polarization in each language. We will also perform parts-of-speech tagging to examine the linguistic structure of the text and identify whether certain grammatical patterns are more common in polarized content.

Class distribution analysis is essential for understanding potential imbalances in the dataset. We expect Subtask 1 to have more non-polarized posts than polarized ones, which is typical of real-world social media data where most content is not explicitly polarizing. For Subtasks 2 and 3, we anticipate uneven distributions across the different types and manifestations, with some categories like political polarization or extreme language likely being more common than others like gender/sexual polarization or invalidation. Understanding these distributions will inform our approach to handling class imbalance during model training.

We will also analyze label co-occurrence patterns in Subtasks 2 and 3 to understand which polarization types and manifestations tend to appear together. For example, political polarization might frequently co-occur with racial/ethnic polarization, or extreme language might often appear alongside vilification. These patterns will provide insights into how different dimensions of polarization relate to each other and may inform our modeling strategy.

Language-specific analysis will compare characteristics between English and Arabic datasets. We expect to find dialectal variations in Arabic posts, potential code-switching between Arabic and English, and different patterns of emoji or special character usage. We will also check for data quality issues such as duplicates, extremely short texts, or potential labeling inconsistencies that might affect model performance.

The preprocessing pipeline will include text cleaning steps such as removing URLs and normalizing whitespace, language-specific processing like handling Arabic diacritics and English contractions, and quality checks to ensure data consistency. Tokenization will use the XLM-RoBERTa tokenizer, which handles both English and Arabic, with a maximum sequence length of 128 or 256 tokens depending on computational constraints. All detailed statistics and visualizations from this exploration phase will be documented and integrated into the modeling pipeline to inform design decisions.

3. Methodology:

This research project will target the English and Arabic languages, given the familiarity of our team members with those said languages. Seeing as each of the subtasks have a different end-goal, it seems appropriate that different models should be applied to tackle each one. The general pipeline will follow such that the text is first fed into the model of the first subtask which will determine whether it is polarising or not, then if so, it will be passed to the model of subtask two or three which will then in turn make inferences about the the polarization type or the manifestation identification, respectively.

For the Transformer-Based approach, the XLM-RoBERTa language model is a good candidate to be used as the base for building the individual modes for each of the substasks, as it is already trained on the English and Arabic language. Subtask 1 will be treated as a binary classification problem. The model will be fine-tuned using supervised training on labeled text entries from the dataset, and optimized using cross-entropy loss. For Subtasks 2 and 3, each will be framed as a multi-label classification problem, where posts may contain multiple polarization types or manifestations simultaneously. These models will use a sigmoid output layer with binary cross-entropy loss to support independent label predictions. Since some categories may be underrepresented, techniques such as class weighting or oversampling to address potential dataset imbalance.

In addition, this research project will pursue a Baseline approach. Text inputs will be converted into TF-IDF vector representations, followed by traditional machine-learning classifiers. For Subtask 1, we plan to use Logistic Regression for binary classification. For Subtasks 2 and 3, a One-vs-Rest Logistic Regression or Linear SVM classifier will be used for multi-label prediction.

4. Expected Results

We expect the transformer-based model to outperform the baseline model in both English and Arabic. Since transformers learn contextual multilingual representations, we anticipate higher accuracy, F1-scores, and better generalization compared to bag-of-words or TF-IDF-based approaches.

Also,we expect our model to successfully detect polarized content in both English and Arabic, although we anticipate performance differences between the two languages. Since English is a high-resource language with abundant pretraining data and more standardized online writing patterns, we expect higher overall performance for English texts. Arabic, on the other hand, includes more dialectal variation, code-switching (Arabic + English), and diverse writing styles, which may make the detection of subtle polarization cues more challenging.

We also expect certain polarization manifestations to be easier to classify than others. For example, “Extreme Language” and “Stereotyping” typically contain more explicit lexical cues (e.g., “never,” “always,” “they’re all…”) and should therefore achieve higher detection accuracy. In contrast, subtler manifestations such as “Lack of Empathy” or “Invalidation” may be harder to distinguish because they rely on implied meaning and cultural context.

For Arabic specifically, we expect stronger patterns of group-based or identity-based polarization, given the sociopolitical nature of Arabic online discourse. English posts, by contrast, may exhibit more individualized and ideological polarization, especially around politics or social issues.

Overall, we anticipate our multilingual model to produce meaningful results for both languages, demonstrating its ability to generalize across culturally different expressions of polarization. More importantly, we expect our analysis to reveal interesting linguistic differences between how polarization appears in English and Arabic, even if classification performance varies.

5. Evaluation Metrics:

Defining appropriate evaluation metrics is essential for evaluating the performance of NLP models and demonstrating their effectiveness. It is important to choose metrics that are appropriate for the specific task being investigated such as comparing the proposed model's performance to a set of baseline models or using statistical tests to ensure that the observed improvements are significant.

- For example: F1, confusion matrix, . . . - Reasoning why that is appropriate

Metrics we will be using and their reasonings:

Macro F1-Score: This will be our main overall metric for all three subtasks. It is calculated by taking the average of the F1-scores across all classes. Why this metric? We use Macro F1 because it treats every class equally, even when the data is imbalanced, making sure the model doesn't only perform well on the majority class.

F1-Score: This is our primary metric for measuring the model's success in identifying the actual content of interest. We track it specifically for the minority (positive) class (e.g., 'Polarized' posts in Subtask 1, or the specific 'Dehumanization' manifestation in Subtask 3). Why this metric? The F1-Score is crucial because it represents the balance between: Precision: How many of the posts flagged as polarized were actually polarized. Recall: How many of the actual polarized posts the model successfully caught.

Per-Label F1: Used to calculate the F1-Score for each individual type or manifestation.

Per-Language F1: Computes the Macro F1 for English and Arabic separately to analyze cross-lingual performance.

Accuracy: Provides a high-level summary of overall correctness.

Confusion Matrix: Breaks down true positives, false positives, and false negatives for error analysis.

Hamming Loss: Used for multi-label subtasks. Measures the fraction of labels incorrectly predicted.

6. Challenges and Limitations:

This involves identifying the potential limitations of the research, and explaining how they will be addressed. This is important for demonstrating the credibility and validity of the research.

Challenges and how we will address them:

Challenge 1: Class Imbalance Problem: Likely more non-polarized posts than polarized ones Solution: • Use class weights • Oversample minority class • Evaluate using F1

Challenge 2: Multilingual Performance Gap Problem: English may perform better than Arabic Solution: • Use multilingual models • Consider translation to English • Cross-lingual transfer learning • Data augmentation

Challenge 3: Data Quality Problem: Annotation errors and noisy text Solution: • Text cleaning • Remove duplicates • Flag mislabeled examples

Challenge 4: Evolving Language Problem: Slang and coded language evolve Solution: • Regular updates • Monitor model performance • Use recent data

Challenge 5: Context Dependency Problem: Meaning depends on context Solution: • Longer context windows • Use threaded conversations if available

Challenge 6: Subjectivity & Ambiguity Problem: Some polarization is subtle Solution: • Strict annotation guidelines • Focus on clear cases

7. Task Assignment:

To ensure effective task assignments, it is important to develop a detailed project plan and to communicate the tasks and expectations clearly to the research team.

The task assignment is distributed as follows:

Faruk will lead the modeling work by designing the methodology and implementation of the strategy for the models for each subtask. This will be the main architecture that will be used in the pipeline, managing training and ensuring all experiments follow a consistent structure.

Mariam will handle the dataset related tasks, including cleaning, preprocessing, normalization and exploration on impacts of each step on the results for different datasets. She will also prepare the final draft for the processed dataset that will be used to generate visualizations.

Elene will lead evaluation and analysis. She will compute metrics that are mentioned in this paper and compare the performance across languages and models, including error analysis. She is also expected to combine this to the pipeline made by Faruk to give further feedback on the models that are experimented by other users.

Jon will work on benchmarking the models, he will also write the literature review on the models that could be used for the 3 different subtasks. He will work with Faruk during the literature review to plan the methodology for the subtasks.

George will support Jon on benchmarking and manage code organization/ reproducibility, also help integrate results from different models and assist Elene on visualisations with different results. At the end he will assemble the final project report outline and all the other members will write it with him upon his coordination.