
Middle East Technical University


Department of Computer Engineering

CENG 463 - Introduction to NLP

Fall 2024 - Assignment 2

Due: 5 January 2025 23:59
Submission: **via ODTUClass**

1 Overview

Debates in national parliaments do not only affect the fundamental aspects of citizens' life, but often a broader area, or even the whole world. As a form of political debate, however, parliamentary speeches are often indirect and present a number of challenges to computational analyses.

In this assignment, we focus on identifying two variables associated with speakers in a parliamentary debate: their *political ideology* and *political orientation* whether they belong to a governing party or a party in opposition. Both tasks are formulated as binary classification tasks. These tasks are part of a Shared Task [1] in CLEF 2025 conference which you are strongly encouraged to participate.

2 Dataset

ParlaMint, resulted in the creation of comparable corpora of parliamentary debates of 29 European countries and autonomous regions, covering at least the period from 2015 to 2022, and containing over 1 billion words. The corpora are uniformly encoded, contain rich metadata about their 24 thousand speakers, and are linguistically annotated up to the level of Universal Dependencies syntax and named entities.

The data is provided as tab-separated text files. The following shows a toy example:

```
id speaker sex text text_en label
gb01 spk1   F   First text. First text in English. 0
gb02 spk2   M   Second text. Second text in English. 1
gb03 spk3   M   Third text. Third text in English. 0
gb04 spk4   F   Fourth text. Fourth text in English. 1
gb06 spk5   M   Fifth text. Fifth text in English. 0
```

- **id** is a unique (arbitrary) ID for each text.
- **speaker** is a unique (arbitrary) ID for each speaker. There may be multiple speeches from the same speaker. sex is the (binary/biological) sex of the speaker. The values in this field can be Female, male, and Unspecified/Unknown.

[1] https://touche.webis.de/clef25/touche25-web/ideology-and-power-identification-in-parliamentary-debates.html

- **text** is the transcribed text of the parliamentary speech. Real examples may include line breaks, and other special sequences escaped or quoted.
- **text_en** is the automatic translation of the text to English. This field may be empty - obviously for speeches in English, but the translation may also be missing for a small number of non-English speeches.
- **label** is the binary/numeric label. For political orientation, 0 is left and 1 is right. For power identification 1 indicates opposition and 0 indicates coalition (or governing party).

Countries : Austria (at), Bosnia and Herzegovina (ba), Belgium (be), Bulgaria (bg),Czechia (cz),Denmark (dk),Estonia (ee), Spain (es),Catalonia (es-ct),Galicia (es-ga),Basque Country (es-pv) [only power],Finland (fi),France (fr),Great Britain (gb),Greece (gr),Croatia (hr),Hungary (hu),Iceland (is) [only political orientation],Italy (it),Latvia (lv),The Netherlands (nl),Norway (no) [only political orientation],Poland (pl),Portugal (pt),Serbia (rs),Sweden (se) [only political orientation],Slovenia (si),Turkey (tr),Ukraine (ua)

2.1 Downloading the Dataset

Please download the training dataset from here: Training Data [2].

The test dataset is also available here: Test Data [3].

If you intend to participate in the shared task, you can evaluate your model on the test set. Since the test dataset lacks labels, you must split your training data into 90% for training and 10% for testing. Ensure the split is performed in a **stratified** manner to maintain the proportion of labels in both subsets[4] [5].

3 Tasks

This assignment consists of two tasks on identifying two important aspects of a speaker in parliamentary debates:

Task 1 : Given a parliamentary speech in one of several languages, identify the ideology of the speaker's party. In other words, this involves performing binary classification to determine whether the speaker's party leans left (0) or right (1).

Task 2 : Given a parliamentary speech in one of several languages, identify whether the speaker's party is currently governing (0) or in opposition (1).

For this assignment, you must select only one country (excluding the UK) and use the parliamentary debates from that country to complete the assigned tasks.

[2] https://zenodo.org/doi/10.5281/zenodo.10450640
[3] https://zenodo.org/doi/10.5281/zenodo.11061649
[4] https://scikit-learn.org/1.5/modules/generated/sklearn.model_selection.train_test_split.html
[5] https://stackoverflow.com/questions/34842405/parameter-stratify-from-method-train-test-split-scikit-learn

4 Models

As part of this assignment, you are required to fine-tune a multilingual masked language model. You may choose any model available on the Hugging Face Hub, such as multilingual BERT, XLM-Roberta-base, or language-specific models like Turkish BERT or German BERT.

In addition, you are required to experiment with a multilingual causal language model, such as Llama-3.1-8B, for inference. Note that fine-tuning a causal language model is not necessary; you should use it in a zero-shot manner for inference purposes only [6].

5 Reporting

You are expected to fine-tune a multilingual masked language model for two distinct tasks and evaluate the inference performance of a causal language model (e.g., Llama-3.1-8B). Your findings should be analyzed and discussed comparatively. To be more specific, consider whether your data was balanced. If it was not, explain how you addressed the class imbalance. Discuss which model performed best for the given tasks, propose ways to improve the results and elaborate on the limitations of your experiments.

5.1 Fine-tuning and Inference Requirements

1. Fine-tune the selected masked language model for each task: For one task use "text_en" and for the other task use "text" (original language).

2. For each task, perform inference using the selected causal language model twice: Once using "text_en" and once using "text" (original language). This approach allows you to evaluate the model's cross-lingual capability, comparing its performance on English text versus the original text.

3. Upload your source code, including the fine-tuning and inference scripts, to your GitHub or GitLab repository. Include a README file in your repository.

4. Provide the repository link in your report.

5.2 Reporting Guidance

- Length: **The report must not exceed 3 pages**.
- Template: Use this reporting template [7].
- Exclude sections like abstract, introduction, related work and conclusion; focus solely on your approach, results, and a brief discussion.
- Include a description of your approach.
- Briefly explain statistics about your dataset on whether it is balanced for each task or not. How did you split the original training data?

[6] https://huggingface.co/docs/transformers/pipeline_tutorial#text-pipeline
[7] https://github.com/touche-webis-de/touche-code/raw/main/touche-paper-template.zip

- Briefly elaborate on your experimental setup: hyperparameter configuration (batch size, epoch, train-val-test splits), training setup (learning rate, optimizer, scheduler etc.), inference setup (temperature, decoding strategy etc.),
- Include comparative results for the fine-tuned models vs zero-shot causal models.
- Write a concise discussion of the results.
- Put a link to your GitHub repository that includes your source codes about the assignment

6 Free GPU Accelerator

If you do not have access to an Nvidia GPU, you can utilize platforms like Kaggle or Google Colab, which offer limited GPU resources for free. Both platforms can be used for fine-tuning your models and performing inference. Alternatively, you can deploy the models on a CPU for inference, though this will result in slower response times.

7 Submission

You are expected to submit a PDF file (**"e1234567_hw2.pdf"**).

8 Tutorials

1. Jupyter Notebook
2. transformers library
3. Fine-tune a pretrained model
4. Inference pipeline for text generation

9 Regulations

- Submission will be done via ODTUClass. You are expected to submit a zip file containing your code and explanations presenting the analysis of your results.
- Late submission is not allowed.
- We have zero tolerance policy for cheating. People involved in cheating will be punished according to the university regulations.
- If you have any questions about the assignment, feel free to ask them via the Discussion Forum (on ODTUClass) or email (rfcekinel@ceng.metu.edu.tr). For more specific questions, office hours (at A206) for this assignment are on Tuesdays from 10:30 to 11:30 AM.