Dialect Identifications with Large Language Models

Vani Kanjirangat and **Ljiljana Dolamic** and **Fabio Rinaldi**
vani5019@gmail.com

Abstract

Language identification is the task of classifying utterances into languages, where languages are regarded as discrete classes (one language, one class). Dialects of a language can be quite overlapping, sharing linguistic similarities, which make the problem more challenging to tackle, where even the existing Large Language Models (LLMs) struggle. In the project, we aim to focus on the multilingual capabilities and limitations of existing LLMs, by using dialect identification as the main task. Most LLMS are trained on a huge amount of data, predominantly in English, while it is claimed that their performance is good enough in other languages too. This could be true with high-resource languages but not with others. In this project, we will be focusing on medium and low-resource languages such as Arabic, Swiss-German, Italian, Indo-Aryan, etc. Our experiments on fine-tuned pre-trained encoder-based models (multilingual and monolingual) have shown the supremacy of language-specific models in these tasks. We intend to compare the performance of LLMS (> 10B parameters) with these pre-trained models to analyze their multilingual abilities (non-English).