**Project: Sentiment Analysis – Hope Detection with BERTology Models and Data Augmentation**

**The subject matter of the project:**

* Input: Text data from social media platforms in English – consists of posts, comments, or any other form of text expressions on social media that convey thoughts opinions or emotions related to hope, expectations, desires, or wishes
* Output:
  + Binary classification labels indicating as a binary output including “Hope” and “Not Hope”
  + Multiclass classification labels categorizing the input text based on the degree or type of hopeful speech it contains, includes “Generalized Hope”, “Realistic Hope”, “Unrealistic Hope” and “Not Hope”

**Main approach:** We employ fine-tining with different pre-trained language models (BERT) for the task. We also investigate how pre-processing steps affect the models’ performance. This is because the data originates from a social media platform, where proper pre-processing can significantly improve overall performance. Furthermore, we utilize various data augmentation techniques to enrich the training data. Finnaly, we implemented a simple ensemble strategy to enhace performance. Besides, we may compares the performance between using BERT-based models and Traditional models (Using TF-IDF / BoW /… techniques)

**Datasets:** The data collection commenced by retrieving the most recent 50,000 tweets encompassing the period from 01/2022 – 06/2022. The datasets included English and Spanish tweets and are used in HOPE at IberLEF2024. However in out projects, we only concentate on English tweets

**Group Members:**

* Bùi Hồng Sơn – 22521246
* Đồng Minh Quân – 22521176
* Nguyễn Minh Sơn - 22521254