Intelligent systems
Assignment 2: Text Classification

December 11, 2024

1 Introduction

In this seminar assignment, your goal is to train and evaluate an NLP classification model capable of detecting emotion in text. The provided datasets contains tweets labeled with one of six emotions (0 - sadness, 1 - joy, 2 -love , 3 - anger, 4 - fear, 5 - surprise). Your task is to develop a classification approach capable of detecting these emotions.

2 Task 1 - Data preparation and exploration (10%)

Load the data and extract some basic information. Are there missing values? What is the distribution of the class variable? How long are the texts? Split the data into train/test (/validation) sets and transform the data into a format suitable for machine learning.

3 Task 2 - Basic machine learning (20%)

Implement, train and evaluate multiple machine learning models on the prepared data. For this task, you only need to implement the basic ML approaches discussed during the lab exercices (e.g., decision trees, random forests, bagging/boosting ...). Use hyperparameter tuning and cross validation to improve the performance of your models.

Make sure to properly evaluate your models: select the appropriate metrics and use the test set only for final testing.

4 Task 3 - Advanced Machine learning (40%)

Use additional ML and NLP approaches to improve the performance of your approach. This task is open ended: we want you to experiment with different methods to improve your ML models. Some approaches you can consider are:

- Using neural networks and deep learning
- Using pre-trained word embeddings or models
- Fine-tuning pre-trained models to improve performance
- Using ensemble models

Additionally, try to extend the provided dataset with additional data. The provided dataset is a small subset of the Emotion dataset (`https://huggingface.co/datasets/dair-ai/emotion`) and only contains a small number of examples. Try collecting additional data from other online datasets or generating additional data with large-language models and evaluate how the dataset size impacts the final results.

IMPORTANT: Make sure your expanded dataset does not include duplicate examples in the training and test sets.

5 Task 4 - Results and report (30%)

Compile a comprehensive report detailing your approach, showcasing code highlights, and presenting the results. When evaluating machine learning models it is especially important to present the results in a clear and concise manner. Your report should include a **results** section comparing the performance of all the approaches you attempted during the seminar assignment. Make sure to include:

- Tables and graphs comparing the performance of different ML models.
- Tables and graphs comparing the effect of various parameters on classification results.
- Tables and graphs comparing the effect of dataset size on classification results.
- Discussion of results. Which approaches worked and which did not?

Make sure the graphs are readable, concise and have appropriate titles and labels.

6 Presentation

Due to time constraints we will be imposing a strict limit of 5 minutes on all presentations. Please focus only on the most important aspects of your work. Presentations that go over that time will receive a score penalty.

7 Submission

- **Deadline: 13. 1. 2025, 10:59**
- **Format:** Jupyter Notebook
- **Group Work:** Maximum of two people per group (if you need a partner, please contact an assistant)
- **Use of Generative AI:** Using ChatGPT to initiate the solution is permitted. Please attach the conversation as a single `.txt` file alongside your Jupyter Notebook.