Intelligent systems
Assignment 2: Text Classification

December 5, 2023

1 Introduction

In this Natural Language Processing (NLP) assignment, you will explore text classification, namely categorizing news documents. The training dataset is described in [1].

IMPORTANT: Use the dataset provided with these instructions.

2 Assignment goals

You are expected to demonstrate the knowledge in the following areas:

- Data Preprocessing: Clean and preprocess the text data, including tokenization, removal of stopwords, and other necessary steps.
- Text Vectorization: Convert the text data into numerical vectors using techniques such as TF-IDF or word embeddings.
- Modelling: Explore various classification models to categorize text documents into predefined categories. Try ensemble methods as well (e.g., bagging or boosting).
- Model Evaluation: Assess the performance of your text classification model using relevant metrics (e.g., accuracy, precision, recall, F_1-score).
- Cross-Validation: Apply cross-validation techniques to ensure a robust evaluation of your model.
- Hyper-parameter Tuning: Optimize model hyper-parameters using methods like grid search or random search to improve model performance.

3 Submission Requirements

3.1 Code Submission

Provide a Jupyter Notebook or Python script containing your code with short comments explaining each step.

3.2 Report Submission

Submit a pdf report with the following sections:

- **Data Preprocessing**
  - Describe the dataset used in your project (number of samples, target class balance, splits, etc.).
  - Provide details on how you handled missing data, outliers, and noise.
- **Details of Text Classification Models**
  - Explain the architecture and components of each model, including any specific configurations and important hyper-parameter values.
  - Report on the methods used to find the best hyper-parameters (e.g., cross-validation)
  - Describe any feature extraction techniques used, such as TF-IDF, word embeddings, or others.
  - Provide code snippets or formulas where necessary to illustrate model components.
- **Results**
  - Present the results obtained from your experiments.
  - Include tables, charts, or visualizations to make the results more comprehensible.
  - Provide insight into the best-performing models and their corresponding parameters.
- **Discussion**
  - Interpret the results by comparing them to the initial objectives and expectations.
  - Discuss any unexpected findings, anomalies, or challenges encountered during the project.
  - Consider the limitations of your approach and potential areas for future improvements.

3.3 Submission format

Submit all files to the course web page.

4 Additional Instructions

- You can use NLP libraries such as NLTK, spaCy, scikit-learn, TensorFlow/PyTorch, huggingface, etc.
- **To get a grade above 8**, you must combine the knowledge you gathered through the semester and focus on improving at least one aspect of your assignment (e.g., text representation, visualization, model explanation, etc). Here are some suggestions, but you are encouraged to choose your own (if in doubt, consult the assistants):
  - explore the use of genetic algorithms to find the best hyper-parameters,
  - try the latest technologies in the field and experiment with pre-trained transformer models,
  - create interactive visualizations to present results in an engaging and informative way,
  - use explanation methods like LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) to interpret model predictions.
- Avoid plagiarism and ensure that your work is original.

5 Grading Criteria

Your assignment will be evaluated based on the following criteria:

- Data preprocessing and cleaning (10%): The quality of your data preparation efforts.
- Quality of machine learning models (40%): The effort to find the best model and its hyper-parameters.
- Detailed analysis & discussion (30%): The depth of your analysis, interpretation of results, and thoughtful discussion.
- Clarity and organization of the report (20%): The clarity and structure of your report, including clear and concise explanations.

References

[1] Rishabh Misra. News category dataset. *arXiv preprint arXiv:2209.11429*, 2022.