Natural Language Processing

Term Project: Arabic App Review Classification

Part-A [10 points]: Dataset labelling

Perform dataset labelling. **Deadline: 18th March.**

Link: https://drive.google.com/drive/u/3/folders/12l8tVkCAU3x_rrZDQqnRlKKm34qegGpk

Part-B [15 points]: Multiclass App Review Classification: from scratch

- Dataset analysis and report on important statistics.
- Implement a from-scratch model
  - (RNN-based model: LSTM/GRU/Bidirectional/stacked/…)
- Train the word embeddings from scratch
- Use precomputed embeddings (similar to word2vec or Glove for Arabic)
- Classification analysis and possible improvements.

Part-C [15 points]: Multilabel App Review Classification: from scratch

- Dataset analysis and report on important statistics.
- Implement a from-scratch model
  - (RNN-based model: LSTM/GRU/Bidirectional/stacked/…)
- You can use the best embedding from Part-B
- Classification analysis and possible improvements.

Part-D [15 points]: Multiclass App Review Classification: Finetune an MLM

- Fine-tune a Bert-type model like Arabert or Marbert
- Classification analysis and possible improvements.

Part-E [15 points]: Multilabel App Review Classification: Finetune an MLM

- Fine-tune a Bert-type model like Arabert or Marbert
- Classification analysis and possible improvements.

Part-F [15 points]: Multiclass App Review Classification: In-Context learning using causal LLM

- Use a causal LLM like GPT or DeepSeek
- **Zero-shot learning**: Design an effective prompt for the task.
  - Show the process on multiple prompts by evaluating them on randomly selected 100 samples from the validation set.
  - Design the prompts in English and Arabic and compare their performance.
  - Select the best prompt design.

    - [BONUS] Study the impact of ideas such as informing the model about the class distribution (based on training data), chain-of-thought, etc.
- **Few-shot learning**: Perform few-shot learning by designing an effective prompt for the task:
    - Select different number of demonstration examples from the training set and evaluate the performance on the validation set. Plot the performance to select the best configuration.
    - [BONUS] Random demonstration examples vs. selecting demonstration examples based on some criteria such as similarity and class labels.
    - [BONUS] Study the impact of ideas such as ordering of demonstration based on some criteria vs. random ordering.

Part-G [15 points]: Multilabel App Review Classification: In-Context learning using causal LLM

- Use a causal LLM like GPT or DeepSeek
- **Zero-shot learning**: Design an effective prompt for the task.
    - Design an effective prompt so that you can evaluate the performance effectively. For example, asking the LLM to generate a formatted output like structured text, JSON file etc.
    - [BONUS] Study the impact of ideas such as informing the model about the class distribution (based on training data), chain-of-thought, etc.
- **Few-shot learning**: Perform few-shot learning by designing an effective prompt for the task:
    - Select different number of demonstration examples from the training set and evaluate the performance on the validation set. Plot the performance to select the best configuration.
    - [BONUS] Random demonstration examples vs. selecting demonstration examples based on some criteria such as similarity and class labels.
    - [BONUS] Study the impact of ideas such as ordering of demonstration based on some criteria vs. random ordering.

Important Notes:

- Split the dataset into 80-20 Train Test. Take 15% of training data as validation set. Use seed 777 for partitioning.
- Proper hyper-parameter tuning based on the validation set.
- Final results on the test set.
- Perform classification analysis and suggest possible improvements.
- You may need to try ideas like dataset balancing and augmentation.

General Notes:

1. All the documents (code and report) should be submitted in Jupyter notebooks.
2. You work as a team of 2 members. You need to decide your team member ASAP.
3. You can work on other projects like spell correction, stance detection (https://sites.google.com/view/stanceeval/home), or other research-based projects in NLP but you need to get the agreement of the instructor.
4. **Deadline to decide the project topic and team members: 16th March.**