NLP Final Project Proposal: GLUE

Divya Parmar, Robert Hilly, Tristin Johnson

What problem did you select and why did you select it?

We plan to select the Glue Benchmark dataset, which is a collection of various natural language related tasks (see more detail here).

We selected this task as it was discussed in class by Professor Jafari, and we can learn from existing professional and academic work in NLP. In addition, the data is cleaned and fairly easy to access within a Python environment.

What database/dataset will you use?

We will use the datasets provided as part of the Glue Benchmark. Those datasets are:

- Corpus of Linguistic Acceptability (CoLA) - Dataset Description
- Stanford Sentiment Treebank - Dataset Description
- Microsoft Research Paraphrase Corpus - Dataset Description
- Semantic Textual Similarity Benchmark - Dataset Description
- Quora Question Pairs - Dataset Description
- MultiNLI Matched - Dataset Description
- MultiNLI Mismatched - Dataset Description
- Question NLI - Dataset Description
- Recognizing Textual Entailment - Dataset Description
- Winograd NLI - Dataset Description
- Diagnostics Main - Dataset Description

What NLP methods will you pick from the concept list? Will it be a classical model or will you have to customize it?

We will use transformer based models and recurrent neural networks (RNN). For transformer based models, we will use existing models provided by public libraries (BERT, RoBERTa). For RNNs, we will use LSTM and GRU. We will build our own LSTM and compare it to transformer models.

We will have to customize the models slightly, as we will have to place different heads on the base model to fit each of the classification tasks.

What packages are you planning to use? Why?

For modeling, we will primarily use transformers (from HuggingFace) and Pytorch (open source but originating from Facebook). These libraries standardize a large part of model building, and will allow us to see how these models can be applied.

We will also use libraries such as numpy for pre-processing.

What NLP tasks will you work on?

As described in this paper, the datasets above concern tasks such as sentence acceptability, sentiment, paraphrasing, and more.

How will you judge the performance of the model? What metrics will you use?

We will judge the performance of the model with a specific metric for each task. The metrics include:

- Matthew's Correlation: Used to evaluate binary classification. Is more stable on imbalance datasets.
- Accuracy: Simply percent of observations correctly classified
- F1 score: Calculated using precision and recall to weight false positives and false negatives differently than straight accuracy.
- Pearson-Spearman Correlation: << FILL IN>>

Provide a rough schedule for completing the project.

We hope to have a base-line model score for each of the GLUE tasks by Thanksgiving. After that, we will set out to explore LSTMs and RNNs, along with all the different transformer based models. In doing so, we hope to apply multiple transformers to each of the tasks, and find out which models perform better on which tasks.