410 Project Progress Report

Team MATH

Anqi Chen	ac89
Tianhao Luo	tluo3
Hang Zhang	hangz2
Md Majharul Islam Rajib	mrajib2

1) Which tasks have been completed?

- Preprocess data (stemming, lemmatization etc.)

  The headline data is rather clean (e.g., `"why is a central bank taking the risks of quantitative easing?"`), without unnecessary white space and in lower case. Our preprocessing involves removal of punctuations, getting rid of stop words, and stemming. For now we use the Porter Stemmer, but may switch to lemmatization as needed.

- Explore the correct models/tools to use

  After some investigation, we decided to use PyTorch and Tensor. The initial approach is based on PyTorch tutorial code.

- Build and train the model

  We use a small portion (5%) of the News Headlines Dataset For Sarcasm Detection as training data. First, we built a frequency count of words that occur in the training dataset. The size of vocabulary is fixed to 5000. Next, we transform the news headlines into word counts and feed to the preliminary PyTorch model.

2) Which tasks are pending?

- Test data with actual dataset & test with different parameters/models - 15 hrs

  We have planned to use some pre-trained language models such as BERT variants and see if we can achieve a better overall performance. We will use the hugging face library and finetune selected models in a PyTorch environment. Some more data preprocessing steps are needed such as tokenization and word embedding. We will probably use the built-in tokenizers in hugging face.

- Create visualization demonstrating the results - 20 hrs
- Create an API such that it can take in some unseen data and make a prediction (10 hrs)
- Documenting work and writing reports - 10 hrs

3) Are you facing any challenges?

- We are looking to reduce the training time to better explore model parameters
- We would like to explore lemmatization as alternative to stemming once we test the dataset