UIUC Online MCS – CS410: Text Information Systems
Fall 2020
Team GFZ (Bei Zhao, Yiming Gu, Ryan Fraser)

# Project Progress Report

## Progress Made

For our project, we decided to compete in the Text Classification Competition. In summary, we have become more familiar with deep learning generally and BERT in particular, and have implemented a working model that surpasses the competition baseline. Our weekly status is outlined below for more detail:

- Week 1: the team spent time individually researching various approaches to solving this problem. We explored Kaggle competition notebooks, browsed academic papers, and searched the internet for tools that could be useful. We then aggregated and summarized our findings.

- Week 2: the team agreed to focus on using the BERT deep learning model for our submission. All of us are new to deep learning, so we spent this week learning more about deep learning generally and BERT in particular.

- Week 3 and 4: to better understand BERT and prepare ourselves for the competition, we all individually used this time to get BERT working on our local machines. This helped us understand how data needed to be cleaned and prepared for use, and also helped us understand the behavior of the model when different parameters change.

- Week 5: we are currently working on enhancing model performance. We are working to get our model running on Google Colab and we are also experimenting with a few changes to our approach. First, we are testing if we can improve performance by cleaning the dataset in different ways. Second, we are attempting to implement an enhanced version of BERT called RoBERTa to see if this helps.

## Remaining Tasks

1. Fine tune the model and prepare it for final submission
    a. Try c1+c2+c3 preprocessing of context with response and c3+c2+c1 preprocessing of context
    b. Try other models such as XLM-RoBERTa, ALBERT, SqueezeBERT, etc.
    c. Change model parameters to optimize performance
    d. Change train_test_split parameters and max_seq_len parameters to optimize performance
2. Document how to use the code
3. Document how the code is implemented
4. Software usage tutorial presentation

## Challenges/Issues Being Faced

In general, things are going very smoothly. Our toughest challenge so far has been identifying a suitable solution to the problem and writing code so that it works for this specific dataset and use case. Getting up to speed quickly on deep learning, while also taking this course, has proven time consuming.