Text Classification Competition Proposal

Team GFZ

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

   Bei Zhao - beizhao3 (Captain)
   Ryan Fraser - rfraser3
   Yiming Gu - yimingg7

2. Which competition do you plan to join?

   Text Classification Competition

3. If you choose the classification competition, are you prepared to learn state-of-the-art neural network classifiers? Name some neural classifiers and deep learning frameworks that you may have heard of. Describe any relevant prior experience with such methods

   Here are some neural classifiers and deep learning frameworks that we are going to learn and try:

   - Word2Vec and GloVe
     Word2Vec and GloVe are word-based models that can map the vocabularies and phrases into vectors of numbers, which can help us to separate the phrases and words. For the context of meaningful classification, they need to combine with neural networks to work out, such as convolutional neural networks (CNN), recurrent neural networks (RNN), artificial neural networks (ANN), etc.

   - BERT
     BERT is a context-based model, which can map the vocabularies and phrases based on the sentence position. However, this model is limited to the size of the corpus it could train compared to Word2Vec and GloVe. But it abandons the traditional RNN and CNN formula, therefore, this may perform much faster than Word2vec and GloVe implementations.

   - Other Models
     We will try other useful classifiers and frameworks in the competition procedure continuously until we find the best fit one.

Here are the relevant experience of each team members:

- Bei: Don't have any Machine Learning experience but will try to learn as much as possible in this competition.
- Ryan: Completed Data Science Certificate from UCLA for-credit, included the "Machine Learning Using R" course which went through the "Introduction to Statistical Learning" book. I will need to take some time to get up to speed on these methods but anticipate it should not take long and am prepared to learn these frameworks. I have made progress in the fast.ai tutorials in an effort to learn more about these methods, and have heard about using TensorFlow/Keras/PyTorch to complete these types of tasks.
- Yiming: Don't have any Machine Learning experience but will try to learn it in this competition.

4. Which programming language do you plan to use?

   The main programming language that we are going to use is Python.

5. Competition Milestones:
   - Text /Data processing
     In this step, we will prepare and process the raw data. For example, text may contain numbers, special characters, etc. We need to transform the raw data into something that could be used in the project.
   - Feature extracting
     We will extract features from the texts in order to classify them. This is important because only when the features are extracted precisely, the text could then be rightly classified. We will evaluate different models in this step as well.
   - Train and evaluate models
     We will train different models on a dataset in order to investigate their actual effectiveness. It is essential that we optimize the models according to the results.
   - Deploy the models in the competition
   - Revise and keep optimizing if necessary
   - Document source code and test set predictions
   - Document the model and the test set predictions, experiments with other methods, hyperparameter tuning, etc.
   - Create a demo that shows code can run on the test set and generate submitted predictions, etc