Project Proposal: Classification of Insincere Questions on Quora

Participants: Hao Mao, Rekha Kumar, Jerry Chen
{ haomao, rekha123, jchen98 }@stanford.edu

Category: Natural Language

- Motivation:
    - On Quora (a platform for people to ask questions and learn from each other), many questions are posted, some sincere, and some insincere (for example, questions founded upon false premises, or ones that just intend to make a statement of some kind). A challenge then is to develop a classifier that can help identify and flag problematic user submitted questions to improve the community experience.
    - We found this problem while browsing competitions on Kaggle[1], giving us a well curated dataset.
- Method: What machine learning techniques are you planning to apply or improve upon?
    - We will apply some basic NLP technologies like tokenize, pad the questions, and add word embeddings based on the embedding data.
    - We will first try classical logistic regression based on the separated training and validation data, and will use the test data to calculate the performance.
    - After that, we will build an RNN (recurrent neural networks) model to compare with previous logistic regression model. Insincere questions may have grammatical and structural differences from sincere questions that may not appear from a standard bag-of words of approach, so we can use the RNN which is good at handling sequential data to check the performance.
    - Kaggle provided train and test data. We will split the train data to 80% for training and 20% for validation.
- Intended experiments: What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?
    - Our model will be evaluated on F1 Score between the predicted and the observed targets. Though the competition in Kaggle was closed, we can still do submissions to see the performance of our models, and compare with others'.
    - During our experiments, we will try to apply dropout to see the performance difference.

    - During our experiments, we will also try to tune hyperparameters in RUU like number of RNN layers, model of RNN like LSTM, batch size, learning rate, decay rate and etc. to see the performance difference, and get the ones that fit ours best.
- References

[1] https://www.kaggle.com/c/quora-insincere-questions-classification