Insincere Questions classification

Summary

Using the “Quora Insincere Question” dataset from Kaggle, this capstone project aims to showcase work that can be done on text-based binary classification via both traditional Machine Learning and modern Deep Learning/AI approaches.

Background

A question-answering platform like Quora needs a function to filter out the insincere questions. Social platforms such as Facebook and Twitter need similar function to filter hate speech or posts with inappropriate contents. Email providers need a spam filter to clean up junk emails. These are some of the binary classification questions in the domain of Natural Language Processing(NLP). NLP has almost a century long history[1]: from the early age of grammatical rule analysis for language translation, semantic and syntactic analysis for Turing Machine, transition to statistical Machine Learning and corpus linguistic analysis; to newly developed Deep Learning methods using neural networks[2].

Objective

The binary text classification question, is one of the supervised learning questions in NLP, where both statistical machine learning tools and neural network approaches can generate satisfactory results. Furthermore, additional enhancements, such as bag of words, word embedding and pre-trained models can add additional power[3]. The capstone project is going to use both statistical machine learning and deep learning to demonstrate the capability of solving such problems.

Problem Type & Dataset

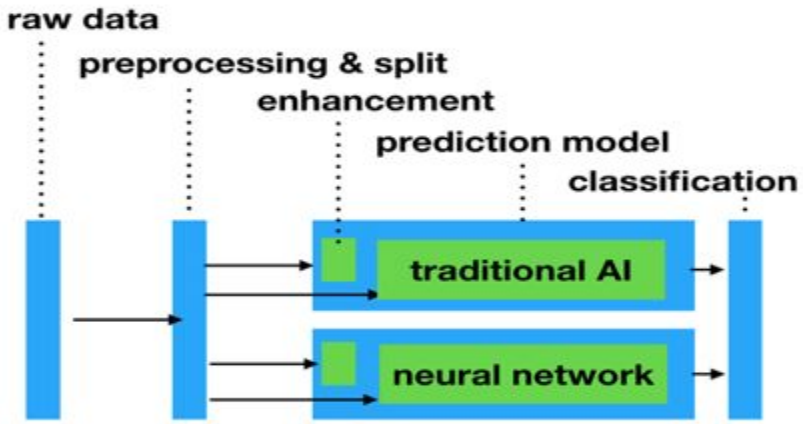
The “Quora Insincere Questions Classification”[4] dataset from Kaggle will be used for this project. The training data contains pre-classified insincere tagging (1 or 0) which makes this question to be a Supervised Learning problem, suitable for Binary Classification . The dataset contains 6GB of data which includes 1.31M training data points (tagged with classification result), 376K test data points, and 4 different embedding models (bag-of-words, GloVe model , Paraphrase Model, pre-trained word vectors). Each training data entry contains a short

sentence submitted to Quora. For this capstone project, only training data will be used as the competition on Kaggle already ended on February 5, 2019.

Project Steps

The proposed flow is abstracted as the above:

1. The training dataset (raw data) will be pre-processed using a variety of methods, including punctuation clean up, tokenizing, lemmatization, and splitting for cross validation. These steps will be modularized to support different cross validation, batch, training/testing ratio, etc.
2. After the sentences becomes token sequences and transfer to unique identification numbers, the traditional Machine Learning approaches can be performed, including Naive Bayes, Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting. The bag of words model and other feature extraction techniques can also be performed to enhance the model's accuracy. I will explore these approaches along the progress of this course.
3. Deep Learning / Deep Neural Network methods also have several options that can be applied towards classifying question. Besides the regular Neural Network, Convolutional Neural Network(CNN), Long Short Term Memory (LSTM), Gated Recurrent Unit(GRU) can be applied as well. Similarly, I will perform the above modeling according to the course progress and add the additional feature models provided in data source to enhance the learning result.
4. The best training model result from each approach will be preserved and will be used to compare the result of different approaches and a summarized classification result will be given according to these results. A web API/service will be built to showcase the outcome of the difference. At this moment, I am not expecting to provide explanation, but a comparison of the best models from each approach.
5. The concept flow is shown in the following diagram :



Resource Requirement

As this is text classification where the resources required are significantly less then image classification questions, according to Springboard's policy, I am planning to use the single GPU unit from Domino Data Lab (https://trial.dominodatalab.com) (via Springboard agreement). The Domino GPU unit has Python 3.6, TensorFlow, Keras and Spark 2.4(local mode) installed, equipped with 8 cores & 15 GB RAM. The AWS Educate account will be the supplement (limited budget and the resource level is budget depend) and my local laptop will serve as a development and testing environment. If additional resource is required, I will ask Springboard to provide the support.

The location and resource for the result API/Web Service is unknown yet. As the automatic model update is not proposed in this project, the required resource for web service should be small, hopefully free.

Reference

1. History of natural language processing (https://en.wikipedia.org/wiki/History_of_natural_language_processing)
2. A Review of the Neural History of Natural Language Processing (http://ruder.io/a-review-of-the-recent-history-of-nlp/)
3. A Comprehensive Guide to Understand and Implement Text Classification in Python (https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/)
4. Quora Insincere Questions Classification: Detect toxic content to improve online conversations (https://www.kaggle.com/c/quora-insincere-questions-classification)