CS412 – Introduction to Machine Learning - HW5 (Written Part)
Data Set: Quora Insincere Questions Classification

a. Data/Task Description and Data Preprocessing

The data comprised of text collected from ~1.3M questions posted on the popular community-based, question-and-answer website, Quora. The labels correspond to whether a particular question was sincere (appropriate) or insincere (inappropriate). The task was to build a model to predict whether a previously-unseen question is sincere/insincere. The dataset was severely imbalanced with ~94% negative (sincere) examples and ~6% positive (insincere) examples. Pre-processing Steps were as follows

1. Converted to lower case; removed punctuations and special characters; removed stop words.
2. Split into training/validation/test sets (70% / 15% / 15%).
3. Performed tokenization.

b. Implemented Machine Learning Solution

An LSTM (Long Short Term Memory) type RNN (Recurrent Neural Network) was chosen due to the ability of RNNs of capturing the temporal behavior of time sequences, a property which renders itself particularly useful in language processing where the relative position of words in a sentence is crucial in determining context, and consequently the underlying sentiment. An embedding layer was also included between the input and LSTM layer.

c. Metrics/Baselines/Experimental Setup

Given that the data set in question was highly imbalanced, the F1-score was used to evaluate performance, rather than accuracy (although accuracy has been reported, too). The following is the experimental setup

1. Classification using imbalanced data (Majority, Random, Logistic Regression, Random Forest and LSTM were used.)
2. Balancing data set using undersampling techniques
3. Classification using imbalanced data (Majority, Random, Logistic Regression, Random Forest and LSTM were used.) Hyperparameter tuning was performed in this step on linear regression, random forest and LSTM.

Majority classifier, random classifier, logistic regression, and random forest were all used as baselines for comparison with LSTM.

d. Software

All models were trained on Kaggle kernels.
Sci-kit learn (to implement logistic regression and random forest algorithms)
Glove (Pre-trained Glove word embedding)
Keras (Deep learning library)

e. Results (on balanced validation and test data sets)