Clare Schwarzenberg

i) After all of the import statements and reading all of the files into the program, I worked on data preprocessing. I used the same data preprocessing techniques for this assignment that I used for the midterm assignment. I took a list of contractions from the internet and converted the contractions in each sentence to the full phrase. This ensured that contractions would be treated the same as the full phrase because they have the same meaning. Then, I cleaned the data by taking out all of the stop words from the sentences and any punctuation. After that I tokenized the data so that the data was no longer in the form of a sentence but was just all of the words that were in the sentence besides the stop words. I then used stemming and lemmatization to attempt to simplify the words to a common base form.

ii) The first two features I used were the length of sentence 1 and the length of sentence 2. Length is helpful because if two sentences are similar in length, they are more likely to have the same meaning. The next feature that I used was the number of words that the two sentences have in common. The more words that the two sentences have in common, the more likely it is that the sentences convey the same information. Then, I used Jaccard Similarity which is the ratio of the number of words that the sentences have in common to the total number of words. Cosine similarity was another feature that I used to consider how similar the sentences were to each other. The last feature that I attempted to implement was METEOR which uses a combination of both precision and recall. I also standardized the data to improve the performance.

iii) To implement multi-layer perceptron I used the Multi Layer Perceptron Classifier or MLPClassifier from the sklearn package. I tried to implement multi-layer perceptron using Pytorch but I found that MLPClassifier was more straightforward and easier to implement. I also used GridSearchCV to find the optimal parameters for MLPClassifier. In terms of packages that I

used, I used the nltk library for preprocessing purposes like stemming and lemmatizing the data. I used sklearn for implementing MLPClassifier, splitting the dataset, calculating the f1 score, and finding the best parameter values using GridSearchCV.

iv) The biggest lesson I learned from this project is that the features that are defined are one of the most important aspects to any machine learning model. Being thorough in the data preprocessing and feature extraction step makes the model run better so it is important to focus on the data. I also learned that it takes a long time to run a multi-layer perceptron model with a normal CPU.