1. What we have done?

We imported twelve necessary packages. We loaded data line by line and convert the raw train.jsonl data to a data frame called “reviews”. Next, we made a clean function to remove URL text such as http, @, #, and any numbers. The information from “response” and “context” was combined into a new column: “text”. We processed the information from “text” with our clean function. We also turned the “label” to 0-1 from “SARCASM/NOT\_SARCASM”. Then we separated both information from “text” and “label” into two sets: “test” set with 20% of raw data and “train” set with 80% of raw data for validation test. We transformed information from text train set and text test to vector with TfidfVectorizer (stopword-‘English’, range of words to be extracted is 1-3). We then applied the XG Boost model and find out the training accuracy is 0.96125 and test accuracy is 0.665. We also loaded raw test.jsonl data in to a data frame called “test” to do the predicted test and processed it with in the same way. We applied the XG Boost model to predict the label and convert 0-1 to SARCASM/NOT\_SARCASM.

2. Remaining task

We will try to find out whether more pre-processing could be done. The parameters of vectorizer may still be adjusted. We will also try to build a deep learning model - LSTM to better improve the accuracy of our model.

3. Challenges

The accuracy of train set is good enough whereas the accuracy of test set is much lower. That means we might overfit.