1. I chose the option 3 which is a text classification task held by the Quora. The purpose of the competition is to use machine learning method to help Quora on handling the toxic or divisive content on its platform. Hence, the training data includes the problem users posted and the target whether the question is identified as insincere.
2. First of all, I chose Bernoulli Naïve Bayes as baseline method. For the nature language processing problem, I did search which model is fit for the NLP problem, and try some classification method on scikit-learn as well, but it turns out that SVM and Random Forest would spend too much time to complete the task with not good performance. Therefore, I selected the Naïve Bayes which is not only good for text classification task but also efficient. In addition, I read some papers which is related to text classification problems by using LSTM and with attention that can significantly improve the results of the traditional machine learning method.
3. I split the validation dataset from 20 percent of the training dataset by using scikit-learn split\_train\_test method with same random state. For the Naïve Bayes, I tuned the alpha parameter by using validation data.

|  |  |  |
| --- | --- | --- |
| Method | validation accuracy | Test accuracy(submit on Kaggle) |
| Naïve Bayes(no processing data and tf-idf) with alpha=1.0 | 0.9320 | 0.522 |
| Naïve Bayes(remove stop words and punciation and tf-idf ) with apha=1.0 | 0.9424 | 0.540 |
| LSTM with tokenize(hidden state=128, learning rate=0.001, and dropout rate=0.1) |  |  |
| LSTM with self-attentive structure(hidden state=128, learning rate=0.001, and dropout rate=0.1) |  |  |

1. For the Bernoulli Naïve Bayes model, I used scikit-learn package. I implemented the LSTM model and attention model by using Keras.
2. The result shows that for the natural language processing problem, it is important to do data pre-processing so that the model can learn the most critical parts of sentences. Moreover, the deep learning model present usually present better than traditional machine learning model, but it has more hyperparameters which is hard to select the good ones.