Chaitra Cheluvaraju
Project: Category prediction for Chat Dataset

Overview:

We are working with a chatbot. We ask the chatbot some random questions and the chatbot answers to them. Each answer of the chatbot belongs to a certain category. For example, if we ask, 'How was your day?', then the chatbot responds saying 'I'm fine! Thank you'. This answer is categorized as '**polite**'. This is implemented for various other questions and answers.

Goal: I am provided with the question from the user and an answer from the chatbot. My task is to predict the category to which the answer of the chatbot belongs.

Dataset: Training data contains three categories ID, Answer, Question and Category. While testing data contains ID, Answer, Question for which I need to predict Category.ID is a integer while others of type string object.

Data Preprocessing: Dataset categories are not imbalanced. So, I decided to use Accuracy as my performance measure. Since text cannot be used directly, need to vectorize it. I have used TF-IDF to retain some level of context with in the vector.

Feature extraction: Since Answer and question together help me in deciding the category, I have clubbed both the features as one feature named conversation.

Model Selection :I have picked up some of the popular text classification multiclass models like logistic regression, Support vector machines, Naïve Bayes and XGBoost.

I have used random search method to assign parameters for each of the model.

Model Name	Accuracy	Time in seconds
Logistic Regression	84%	0.177
XGBoost	84.5%	8.2
Naïve Bayes	78%	0.088
SVM	87%	15.17

Conclusion:

For a model it is important to have good speed as well as good accuracy. On looking at the logistics, I would rank one for Naïve Bayes considering time and I would rate rank 1 for SVM considering accuracy. It seems we have a tradeoff between time taken and accuracy. Considering both I would go for Logistic Regression.

Further Improvements:

I can further fine tune to obtain better results and can use advanced NLP (Natural language processing) methods like Word2vec or BERT for better accuracy.

Deployment: I have provided a pickle file for deployment which contains SVM saved parameters which provided me the highest accuracy.