Sentiment Analysis API Project

By Lawrence Xiaohua Li

In this machine learning project, I experimented with the NLTK Naïve Bayes Classification model. NLTK is a natural language modeling module that can be used to analyze sentences and words. This includes sentiment analysis and classification.

To fine tune my model, I used three tactics as described below:

- Tokenize sentences by splitting strings into smaller parts (words). By observing the data, I found that the dataset originated from the Airline twitter replies. Therefore, I used the twitter tokenizer method provided by nltk library to effectively split the sentences for machine learning.
- Word normalization. For example, “walk”, “walking”, “walks” are various forms of the same verb. So, I converted all these versions to the same form. I used lemmatization algorithm provided by library to normalize a word with the context of vocabulary and morphological analysis of words in text.
- Remove noise. I removed any stop words like “is”, “the”, “a”, which are generally irrelevant for sentiment classification. I also used regular expressions to remove any hyperlinks, twitter handles in replies (username with @ symbol), and punctuation/ special characters.

Final metrics:

The testing accuracy of this model is 91.47% after 20 training epochs based on a 87:12 train test split ratio.