CMSC726 : Machine Learning *Ravi Bhadeshiya*

Feature Engineering

Method:

Every set of the feature was explored by randomly choosing 60% of total data. Then, the data were split into the training set with 70% and validation set with 30%. However, the more precise accuracy was measured by using 90% of total data as training set and remaining 10% as validation set.

Features:

First intuition was to extract more meaning from the sentence. So, test accuracy was increased to 66% by changing to ngram_range = $(1, 4)$ which was the significant improvement over the baseline but it increased the features to 300k. Later, ngram_range = $(1, 2)$ was used with other combination mentioned below and which is quite efficient and accuracy was very similar. Tweaking min_df and max_df tend to limit the features but accuracy was not changing significantly.

In the top features, there are words were repeating such as kill, kills, killed, die, dies died ...etc. The remedy was to use PorterStemmer as preprocessor tokenizer with RegexpTokenizer in it to remove punctuation as well. And parallelly, the change was made to use TfidfVectorizer instead of CountVectorizer. Where TF-IDF is another way to judge a sentence by the words it contains. This changes elevated the accuracy to nearly 70%.

There are other features were also explored such as stop word as English, strip accent as 'ascii', WordNetLemmatizer for lemmatizing part of speech and word_tokenize as tokenizer but accuracy seemed to decrease and differentiating part of speech and legitimizing are time consuming. Also, it doesn't make too much difference in top features.

The Movie/Tv genre is an important feature for detecting the spoiler such as Thriller, Mysteries, Crime are more likely to contain spoilers and Comedy, family are not. The OMDB API was used to create Genere dictionary from "page" and it was added as an extra feature.The code also contain another python file which query over the OMDB API and generate the Genre for each unique "page" and save it into pickle. Second thing, More popular Movies/Tv-series are more likely to discuss the spoiler online. "Trope" was also added as extra features. By adding Genre and Trope improved accuracy to nearly 72% in the test case and nearly 74% for validation case.

Last thing were observed that feature union and pipeline are the perfect way to implement but it decreases the accuracy as well code stub needs to modify significantly. However, I believe that future direction of the project was to implement the feature union and pipeline perfectly.