

Nick Videtti

[nvidetti@syr.edu](mailto:nvidetti@syr.edu)

IST 664 – Natural Language Processing

Final Project - Winter 2023

**Overview**

This project is a deeper look into the Brown Corpus, some feature sets corresponding to the Brown Corpus, and text classification using those features with Naive Bayes Classifiers in the Natural Language Toolkit (NLTK). During this, the feature sets were analyzed by finding the most informative features, the classifiers will be tested using k-fold cross validation, and additional feature sets were created based on the results of these most informative features and cross validations.

**About the Brown Corpus**

Chapter 2, Section 1.3 of the NLTK Book (found at <https://www.nltk.org/book/>) tells us that the Brown Corpus is a corpus of English that was created at Brown University in 1961. This corpus is known to be the first million-word electronic English corpus. It has corpora from 500 different sources, each of which categorized into one of 15 different genres. The Brown Corpus is conveniently available in the NLTK (specifically ***nltk.corpus.brown***), which is the medium that was used in this project to access it.

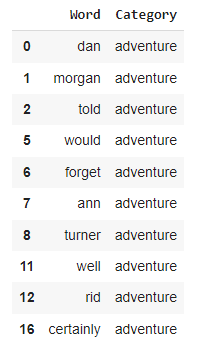
**Word and Sentence Tokenization**

Luckily the words and sentences are available in NTLK in Python list format. The sentences themselves are a single string, but needed to be converted to lists of their corresponding words for the analyses in this project. That is, the end result should be that the words are a list of words and the sentences are a list of sentences with each sentence being a list of its words. However, that is not all that needed to be done, as non-alphabetic phrases and NLTK’s English stopwords needed to be removed. Unfortunately, the processing speed of removing stopwords from the sentences was too slow, so only the non-alphabetic phrases were removed from them. This issue with the sentence stopwords was somewhat addressed when creating the feature sets. The last step was to assign each text’s corresponding category to it and then store those results in a Pandas DataFrame.

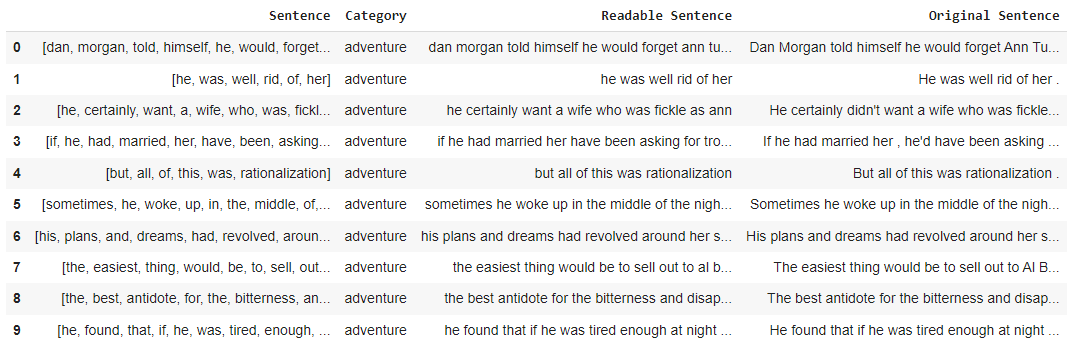
In the Pandas DataFrame for words, two columns were created, one for the word and one for the category. In the Pandas DataFrame for the sentences, 4 columns were created. These columns were for the sentence, the category, the “readable” sentence, which is the sentence combined into a single string, and the original sentence from the Brown Corpus. The results of this initial setup were as follows.

Pandas DataFrame of Tokenized Words

(Lowercase, Alphabetical Phrases, No Stopwords)

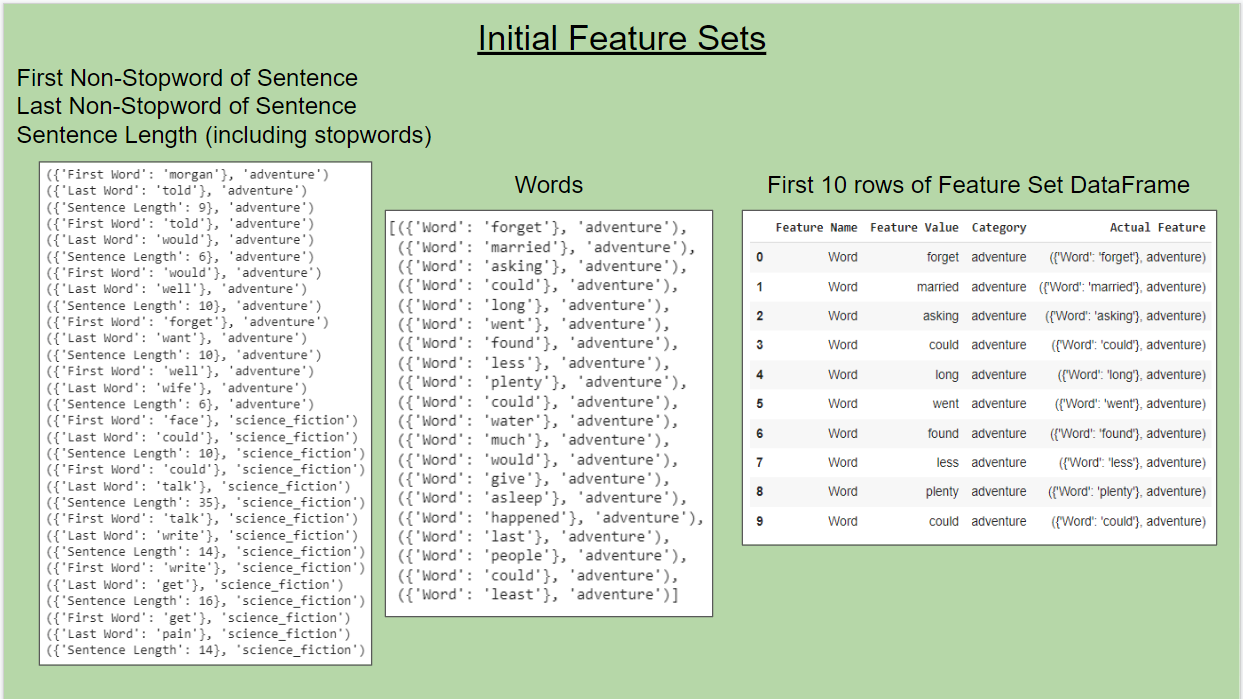


Pandas DataFrames of Tokenized Sentences

(Lowercase, Alphabetical Phrases, Stopwords Included)

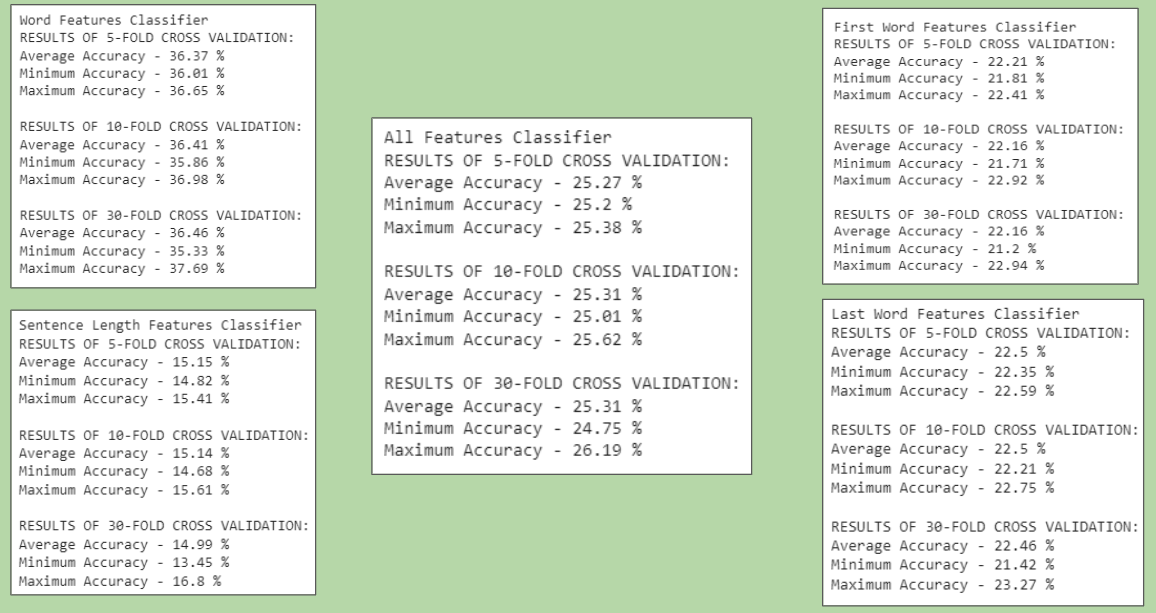
**Initial Feature Sets**

Now that the data needed had been put nicely in Pandas DataFrames for use, the next step was to create feature sets to use in NLTK Naive Bayes classifiers. The word data were used to create “Word” features and the sentence data were used to create sentence-level features, such as “First Word”, “Last Word”, and “Sentence Length” (length in words). It is important to note that the previously mentioned issue where stopwords were not removed from sentences was somewhat addressed while creating these feature sets, specifically the “First Word” and “Last Word” features. For these, they actually show the first non-stopword and last non-stopword of the sentence. While it took too much processing time to remove the stopwords completely, simply finding these first non-stopwords and last non-stopwords was a much more feasible task. As a disclaimer, the sentence lengths are based on the sentences including stopwords, but not including non-alphabetical phrases. These feature sets were created and then stored in a Pandas DataFrame. This Pandas DataFrame was created with 4 columns, one for each of the three feature attributes (feature name, feature value, category) and one for the actual feature. The results of this step were as follows.



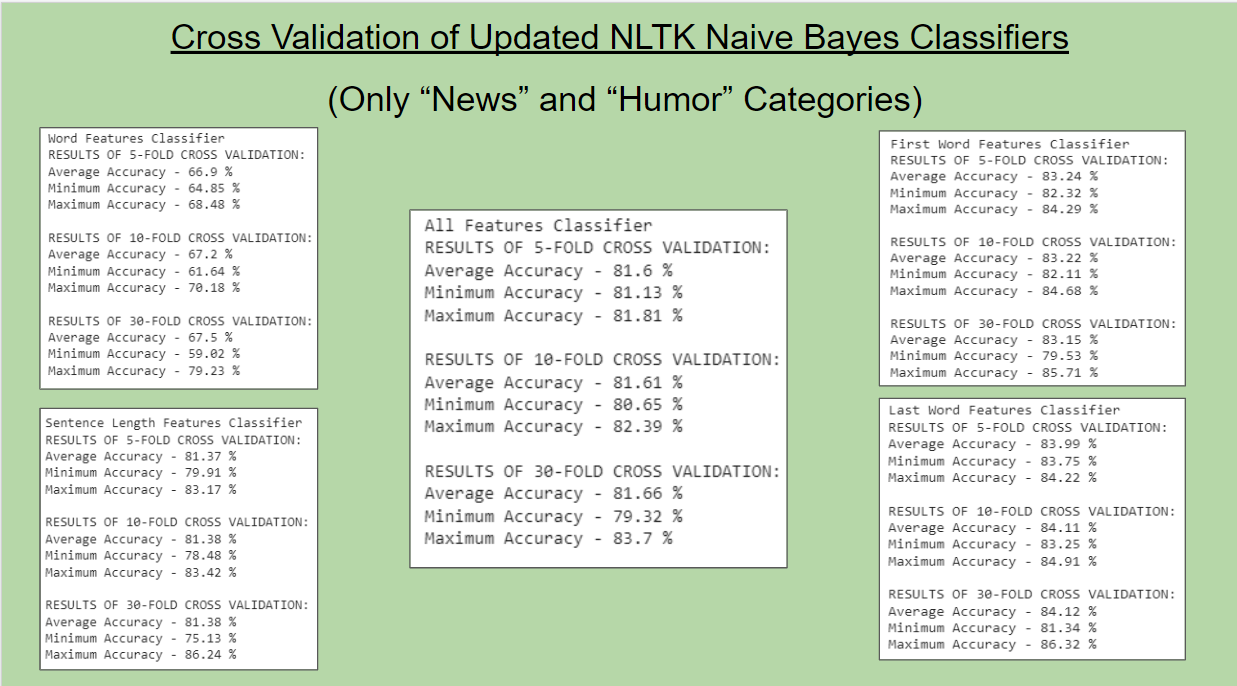
**Initial NLTK Naive Bayes Classifiers**

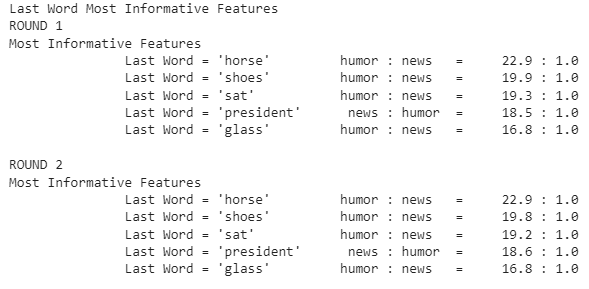
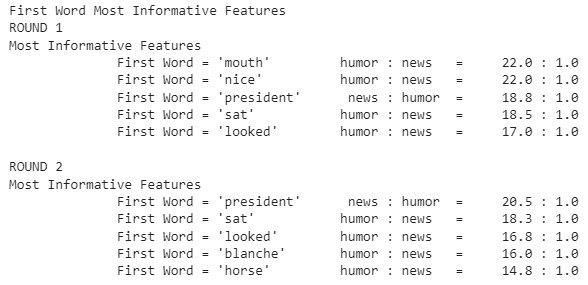
These features were used to create NLTK Naive Bayes classifiers. These classifiers were used to create models that predict category based on the specified feature. Five different feature sets (Words, First Words, Last Words, Sentence Length, and all of them combined into one feature set) were used to create classifiers. For each feature set, classifiers were tested using 5-fold, 10-fold, and 30-fold cross validations. The minimum, maximum, and average accuracy scores were calculated for each cross validation. The results were as follows.

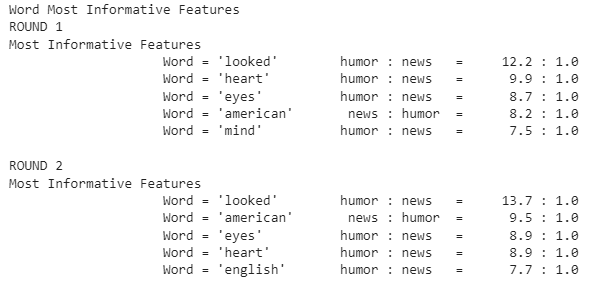


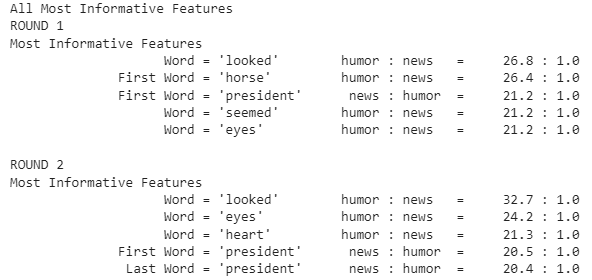
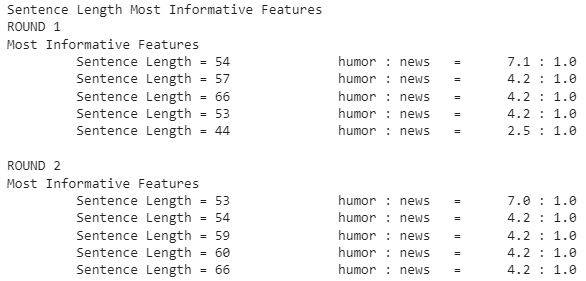
We see these results are overall not too impressive. It seemed that trying to predict between the 15 different categories in the Brown Corpus using these feature sets was not realistically going to be too accurate. For this reason, these feature sets and Naive Bayes classifiers were recreated using only the “news” and “humor” categories. Once these were created, the most informative features were found as well. The results were as follows.

**Updated Feature Sets - Only News and Humor Categories**

****







These are very pleasant results, as the accuracies are much higher predicting just between the “news” and “humor” categories than they were while predicting between all 15 Brown Corpus categories. Also, the most informative features all have very strong ratios and have features corresponding to both categories, with the only exception being the sentence length features. These sentence length features are all not very intuitive, as they are based on the raw frequency of words in each sentence, meaning, for example, that a 54 word sentence and a 55 word sentence are treated as differently as a 54 word sentence with a 3 word sentence. The way to fix this was to create sentence length ranges and replace the sentence length features with sentence length range features.

**Sentence Length Range Features**

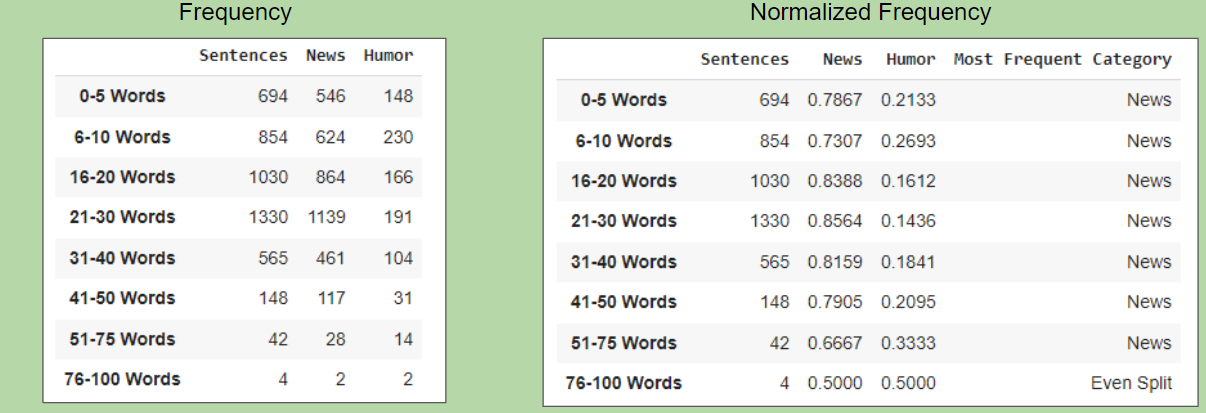
The minimum, maximum, and average sentence lengths were found to determine what good length ranges to use would be. Then the sentence length features were updated accordingly. The most informative features were found once again and the results were as follows.



These are much more intuitive, and much more useful results. While the ratios are lower for the most informative features using the ranges instead of the raw lengths, these actually have a mix between humor and news, and are far more intuitive.

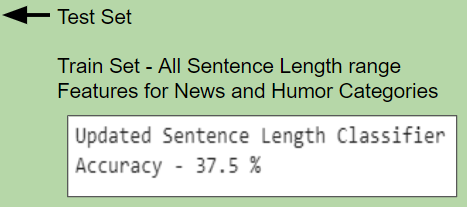
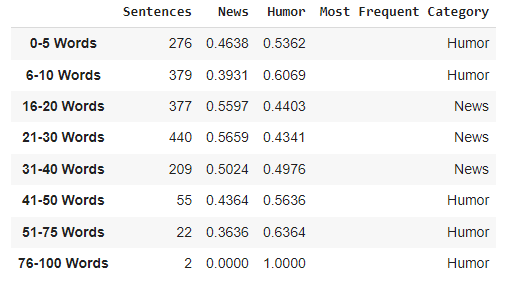
**Final Experiment - Sentence Length Range Analysis**

To wrap up this project, the sentence length ranges were analyzed and put into a Pandas DataFrame. The results were based on the sentence length range features themselves rather than any predicted categories. Both the raw and the normalized frequencies were found, and the results were as follows.



These results may give a sense that for any given sentence length range that it is most likely to be predicted “news”, however, there was a major flaw about this analysis hiding in plain sight. There are substantially more news sentences than humor sentences, which had skewed these results. The way to fix this was to create a new feature set with the same number of news and humor features. For whichever of the two had the smaller amount, all of the corresponding features were still included, but only that same amount of the other category were included instead of all of them. This gave a much more indicative relationship between sentence length range and category.

This result was used to do one last NLTK Naive Bayes classifier. Rather than testing via cross validation, this used a single train set and test set. The train set was all sentence length range features corresponding to news and humor categories, and the test set was the results we had just created using the relative frequencies. The results of all of this were as follows.



We see that news tends to be more in the middle of lengths, while humor is either longer or shorter than usual. This makes sense since humor tends to be less formal and contains “one-liners” or setting up of jokes, while news is much more formal and standardized. As far as the accuracy of the classifier, 37.5%, or 8\*0.375 = 3 of the 8 length range features were predicted correctly, which isn’t too bad given so few test rows.

**CONCLUSION**

While countless more testing and experiments could be done, there was still a lot learned during this project. One lesson was that It is important to identify what features or attributes will be of interest as we unfortunately spent a decent portion of this project getting to the point where we settled on the sentence length ranges as a good and meaningful attribute. This is likely just a byproduct of the fact that the classifiers used supervised machine learning. Ideally, all attributes would be stored in a data frame and unsupervised learning would be done to discover which features are of best interest to be more efficient in this process.