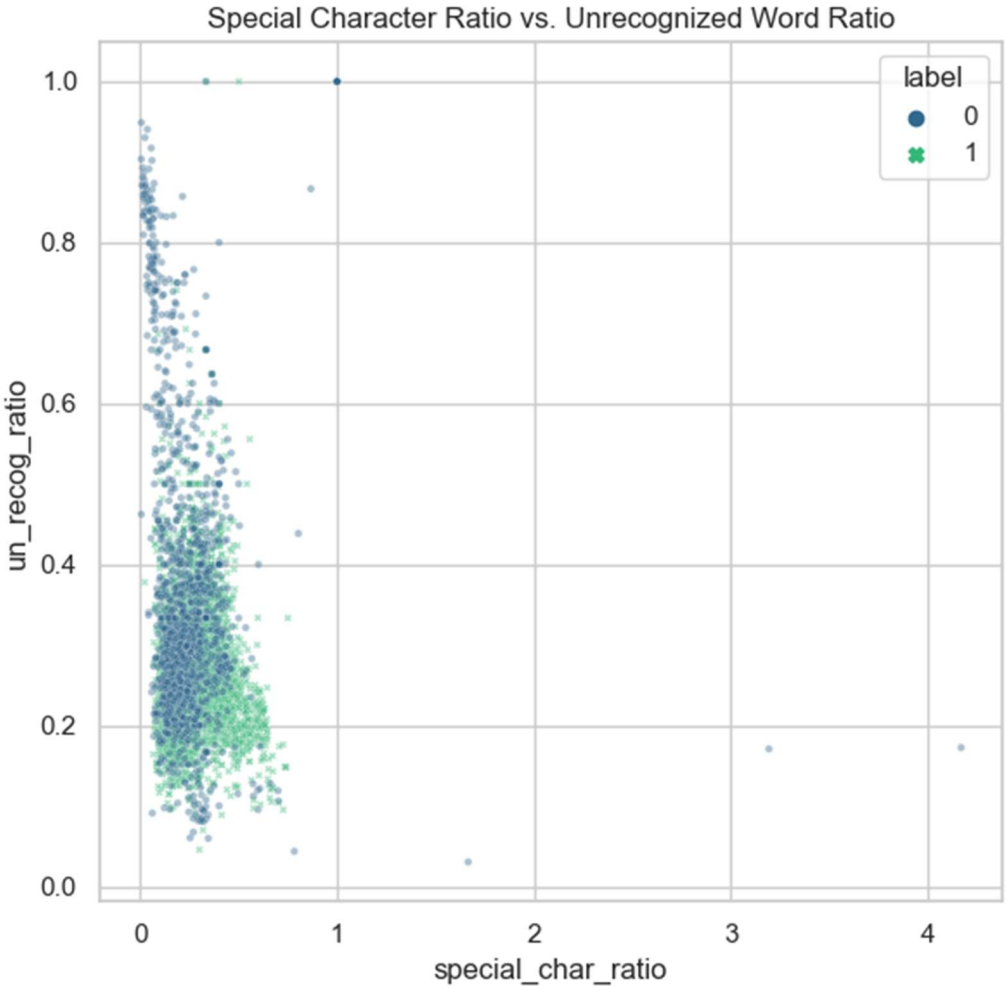
CS5262 Project Update

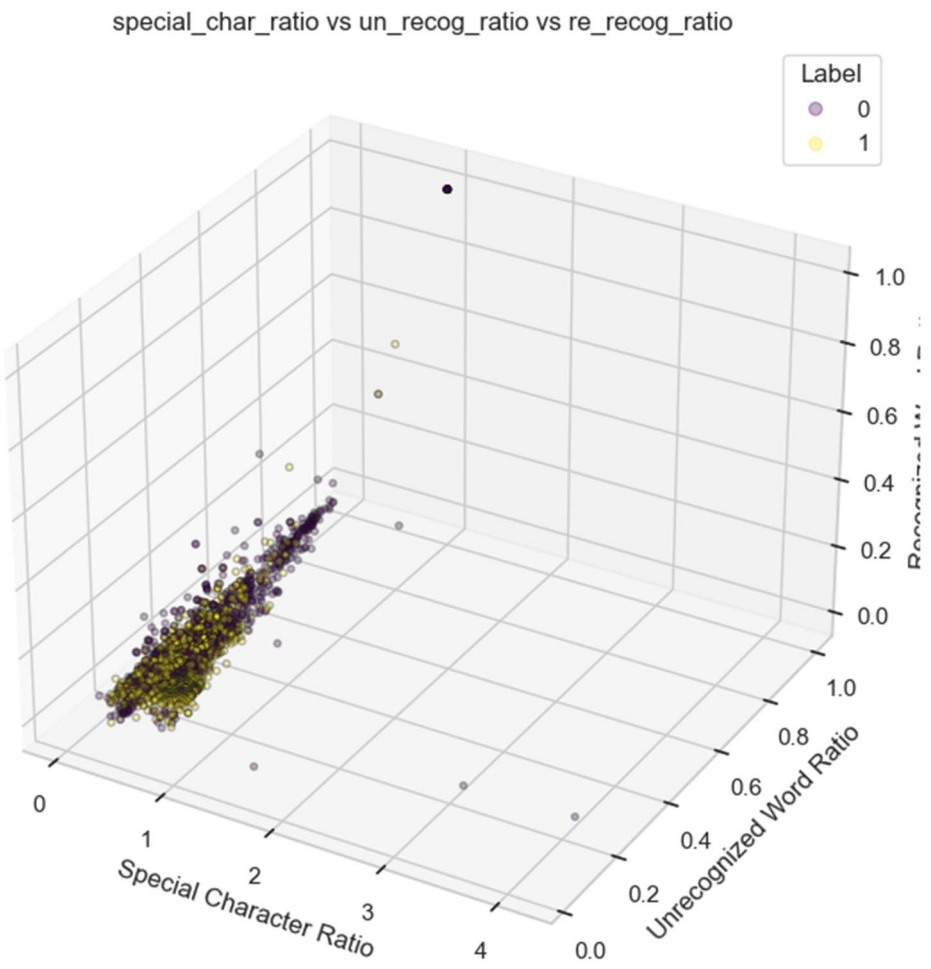
Yizhou Guo

- Data Exploration

I settled down on Enron Spam dataset as discussed in proposal. In choosing features, other than just counting some specific words, I added a couple extra features. I started by loading 3000 most frequently used English words (source: https://gist.github.com/hyper-neutrino/561f120125ae0e7c1d22777eebf083c8). Then, I compared the words in email with the data set, then count un-recognized word, re-recognized word (after removing all non-alphabetic symbols), and number of special characters. These features proved to be more effective, both in terms of training time and prediction accuracy, compared to simply counting the key words.

- Plotting results





- Preliminary results

I've used multiple linear regression and SVC model from sklearn. For multiple linear regression, the accuracy is 45-70% using specific word counts depending on choices of words, and 72-73% using the three ratios. For support vector machine, the accuracy further goes up to 73-74% using RBF kernel and 74-75% using polynomial kernel. While the accuracy looks reasonable for tokenized machine learning methods (although not as good as NLP methods explored by previous researchers), the false-positive rate P(False Positive | Positive) goes close to 25% which is concerning. In the next steps, I plan to focus on eliminating false positives and feature selection on the model before implementing neural network methods, as planned in the original proposal.

- Updated Timeline

By February 28th: Process, tokenize, and visualize data. (Completed)
By March 7th: Apply simple methods to learn the data using small datasets. (Completed)
By March 24th: Feature selection, refine SVM model.
By April 7th: Complete Neural Networks method.
By April 14th: Prepare for presentation.
By April 26th: Final write up.