Houston Community College

**Mid-Term Report**

News Bot

Author: Trevon Woods

NLP - ITAI - 2373

Patricia Mcmanus

Due: 07/28/2025

I’ll start with some of the things I did that were unique in this assignment. For the classifier I removed the MultinominalNB classifier because it didn’t allow negative numbers in data points. I replaced it with GaussianNB because I’ve had experience with using this classifier and it seems that adding a little noise to any model for training always boosts performance. I achieved almost 95% accuracy for the GaussianNB classifier. Next was the Logistic Regression classifier, I initially tried to keep it standard with a random state but that only achieved around 78% accuracy. So I experimented with adding two hyperparameters, C and the optimization algorithm solver. I settled on c=10 and newton-cg for the optimization algorithm. By default the classifier uses lbfgs (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) algorithm which boasts that it is efficient for multinomial datasets but hardly helps convergence in my experience. I opted for newton-cg because it converges faster than gradient descent and it’s also for problems with many parameters. I tried to do some hyperparameter tuning with the SVM but I wasn’t able to get it to achieve an accuracy higher than 32%.

Now I’ll get into some of the NLP techniques that I found the most challenging. I found the extraction of syntactic relationships and dependency parsing along with the NER portions of this assignment particularly troublesome. I had to do a lot of research for myself into the many different aspects of these techniques in-order to really break down and understand how to implement visualizations like co-occurrence networks, entity relationship graphs, and entity sentiment associations. I also had to address the computation issues when running the code for things like syntactic analysis or complexity. The dataset I used was 1490 samples long and for each sample the CPU had to break the sentences into tokens and make calculations for each token depending on what I wanted to do. This drastically increased the time it took to run the code. I overcame this by just sampling a few samples from each category and putting them in their own dataframe for the analysis.

One of the real world problems this system could solve could be faster ingestion of news content. Some people don’t have all day to read a long wordy article, this system could give quick information about what’s in the article. This could also help enhance recommendation systems for news applications. For example if a user only likes positive news then this system could help filter everything the user sees to only show positive news. Another example could be if a user is only interested in seeing news about certain companies or celebrities. This system could help filter all news to show only articles based on these preferences.

There are many potential risks of automating news analysis, for starters the system could misclassify news articles. For example lets say a user is waiting for the release of some news about a company they are investing in. Even though most financial apps expressly tell users not to solely rely on their apps output, people usually don’t listen. The news could come out and be classified as negative when in fact the news was positive. This would open the app up to lawsuits if the individual makes a financial decision based off of the output of the system and gets a negative outcome.

The NLP topics I’m the most excited to learn next are Topic Modeling and Transformers like BERT. Topic modeling automatically discovers hidden thematic structures in large collections of documents. Instead of pre-defining categories, topic modeling algorithms like Latent Dirichlet Allocation (LDA) can identify clusters of words that represent topics. Transformers like BERT are state-of-the-art deep learning models that have revolutionized NLP. They are pre-trained on massive amounts of text data and can be fine-tuned for a wide range of tasks, including classification, question answering, and text generation.