NewsBot Intelligence System 2.0

Codie Munos
ITAI 2373

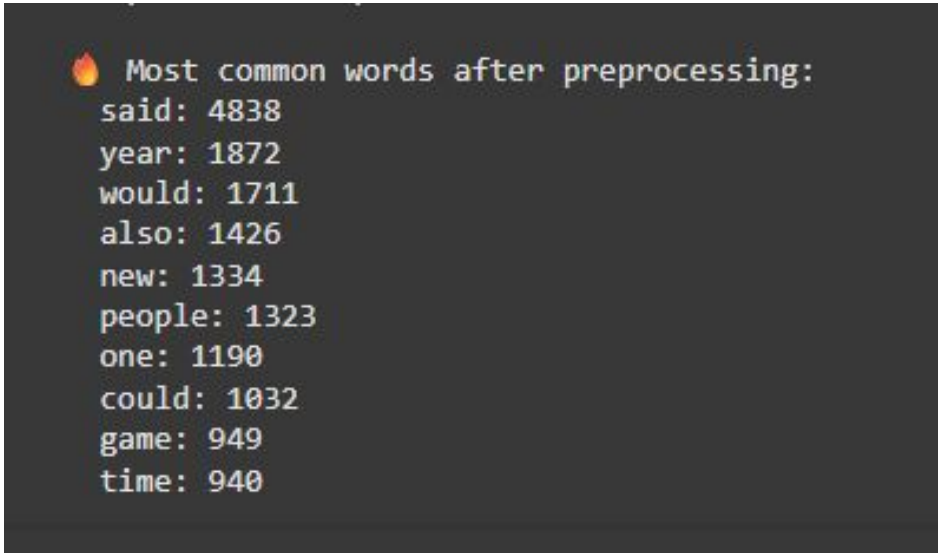
Project Layout:

This newsbot is able to classify, analyze, and understand news articles found in the BBC News Train.csv file from Kaggle datasets. This advanced model can be used to classify articles when researching, studying, investigating, comparing, and so much more!

Techniques Used:

- ★ Text Preprocessing Pipeline
- ★ Feature Extraction and Statistical Analysis
- ★ Part-of-Speech Analysis
- ★ Syntax Parsing and Semantic Analysis
- ★ Sentiment and Emotion Analysis
- ★ Text Classification System
- ★ Named Entity Recognition
- ★ Comprehensive Analysis and Insights
- ★ Topic Modeling and Content Discovery
- ★ Response Generation
- ★ Integration

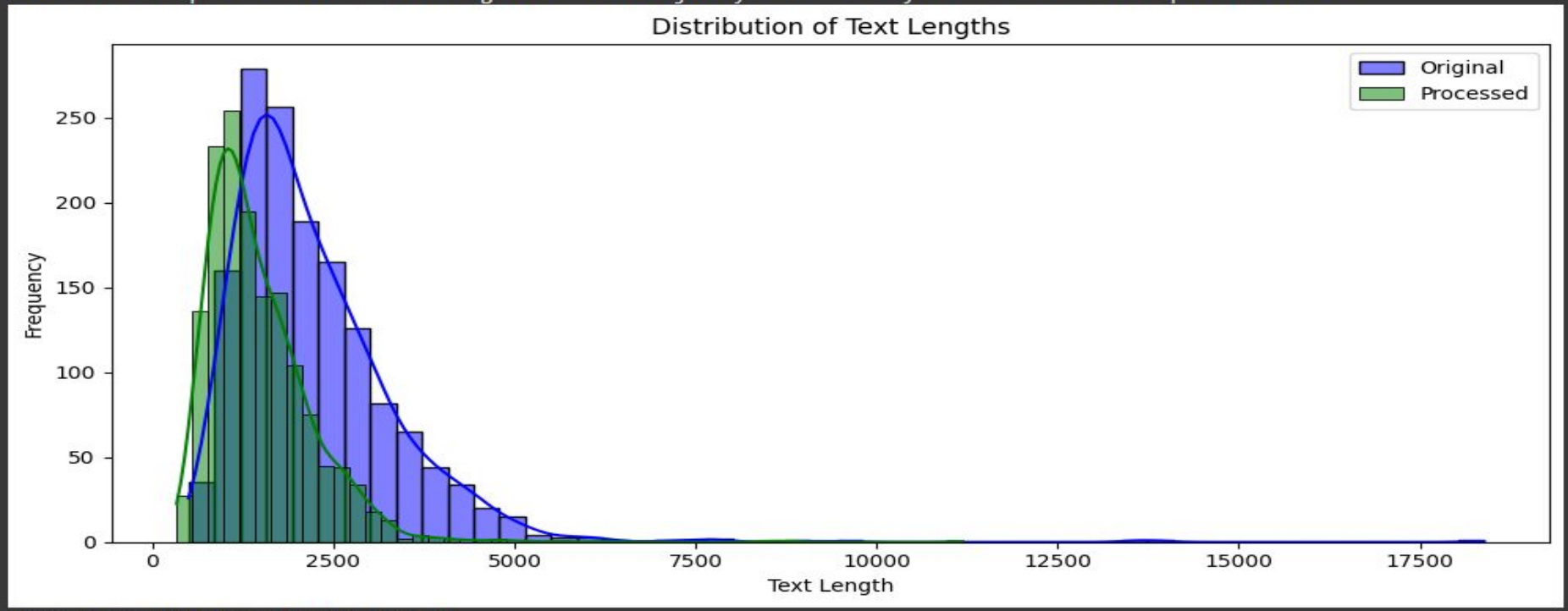
Key Results for Text Preprocessing:



```
Example 1:
Original: worldcom ex-boss launches defence lawyers defending former worldcom chief bernie ebbers against a ba...
Processed: worldcom exboss launch defence lawyer defending former worldcom chief bernie ebbers battery fraud ch...

Example 2:
Original: german business confidence slides german business confidence fell in february knocking hopes of a sp...
Processed: german business confidence slide german business confidence fell february knocking hope speedy recov...

Example 3:
Original: bbc poll indicates economic gloom citizens in a majority of nations surveyed in a bbc world service ...
Processed: bbc poll indicates economic gloom citizen majority nation surveyed bbc world service poll believe wo...
```




```
Average original text length: 2233.46
Average processed text length: 1481.36
Unique words in original text: 35594
```

Key Results: Multiple Classifiers

```
🤖 Training multiple classifiers...

🔄 Training Naive Bayes...
   ✅ Accuracy: 0.9765
   📊 CV Score: 0.9639 (+/- 0.0167)

🔄 Training Logistic Regression...
   ✅ Accuracy: 0.7047
   📊 CV Score: 0.6879 (+/- 0.0698)

🔄 Training SVM...
   ✅ Accuracy: 0.3557
   📊 CV Score: 0.3582 (+/- 0.0363)

🏆 CLASSIFIER COMPARISON
==================================================
                 Model  Test Accuracy  CV Mean  CV Std
0          Naive Bayes         0.9765   0.9639  0.0083
1  Logistic Regression         0.7047   0.6879  0.0349
2                  SVM         0.3557   0.3582  0.0182

🥇 Best performing model: Naive Bayes
```

Challenges:

- Developing this model was a wild ride, but a needed one. Challenges such as the following discouraged me momentarily, but benefited me in the end.
  - Integrating a generative system
  - Enhancing the sentiment
  - Creating my repository

Future Improvements:

I would like to take it a step further and test myself with what would have been the bonus and implement this on a web application. I will look into more multilingual abilities to incorporate more in this model.