Cole Daley

October 16, 2023

Building a Spam Filter

What is Spam?

Spam is unsolicited and often irrelevant messages sent in bulk through various forms of communication. This can be done through phone calls, emails, and messages in general. Spam is often used to promote scams, spread malware, and even conduct phishing attacks. It is important for organizations to provide awareness around spam to help with the protection of assets.

CONTENT –

In this lab, you will need:

* IDE, the author used Spyder (PyCharm, VS code, and others will work)
* Installation of Python
* Dataset for evaluation, the author used (<https://plg.uwaterloo.ca/~gvcormac/treccorpus07/>)

Part 1:

Step 1: Writing the original source code and implementing the dataset.

A screen shot of a computer program

Description automatically generated

In short, the above source code allows us to see raw words from the emails which will allow us to run through various classification tests to look for spam within the emails. Setting up this source code will allow us to call it and use it for the dataset in part 2.

Part 2:

Step 2: Executing different performance metric models.

A screen shot of a computer code

Description automatically generated

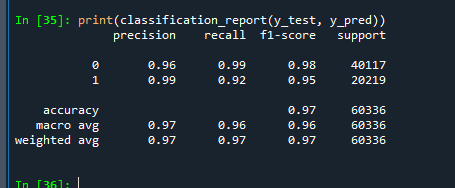
A screen shot of a computer code

Description automatically generated

Understanding what the code above is doing:

1. Importing Libraries
2. Set the Data Directory and Labels File
3. Read the Labels
4. Writing the function for Read Email Files
5. Splitting Data (train\_test\_split())
6. Vectorization (convert text to numbers)
7. Classifier Initialization
8. Training the Classifier
9. Making the beginning prediction (y\_pred)
10. Running the Confusion Matrix
11. Running the Performance Metrics
12. Classification Report (Ham vs Spam)

The source code above will allow us to run through the test of Logistic regression and see at what accuracy we can catch spam. The picture below shows the results from the source code above and it showed encouraging results.

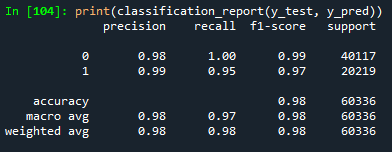


Part 3:

Step 3: Comparing Regression to different performance metrics.

When working through making a spam filter we want to try various tests to see which can give us the highest percentage at catching the spam. By following the 12 steps listed above we can switch only a few lines and then be able to run decision trees, SVM, MLP’s, and random forest metrics.

The highest test resulted from all five tests shown above was the MLP test. After importing the necessary MLP imports and defining the fit and predictions for MLP it gave us the results seen below.



Part 4:

Lab Summary –

When working through the various tests to catch spam at the highest rate we saw MLP come out on top. This is not always the case and there is a use for always running all the various tests. With different datasets you might encounter different results overall. Being flexible and understanding what metrics hold value is what ultimately is important from this lab.