Rajendra Prasad Ponnam

Applied Data Science, Bellevue University

DSC680 Week8 Assignment

Professor Catherine Williams

May 08, 2023

Spam Email Classification

Topic:

One of the authorized and widely used methods of sending data and information across digital and electronic devices is electronic mail, or email. Fraud, phishing, and other unethical and criminal behaviors are all carried out using spam emails. distributing malicious links via unsolicited emails, which can compromise our system and attempt system access. The spammers target those who are not aware of these scams and easily create fictitious profiles and email accounts for them. They use a real name in their spam emails. Therefore, it's important to recognize fake spam emails.

Business Issue:

Email is used by millions of users all over the world to communicate and exchange files between email servers. On the other hand, because of the volume of spam that grows astronomically each year, unwanted emails, also known as spam, have become a problem for major businesses and organizations. Spam can be irritating and contain harmful material. Additionally, spam uses up computer, server, and network resources, creating a negative bottleneck and reducing the speed and memory of digital devices. Additionally, users spend a lot of time deleting unwanted emails.

Datasets:

I used the Kaggle dataset https://www.kaggle.com/datasets/shantanudhakadd/email-spam-detection-dataset-classification to categorize the mail as spam or ham using a machine or deep learning model.

A target variable that indicates whether an email is spam or not and data taken from 5169 email messages make up this dataset.

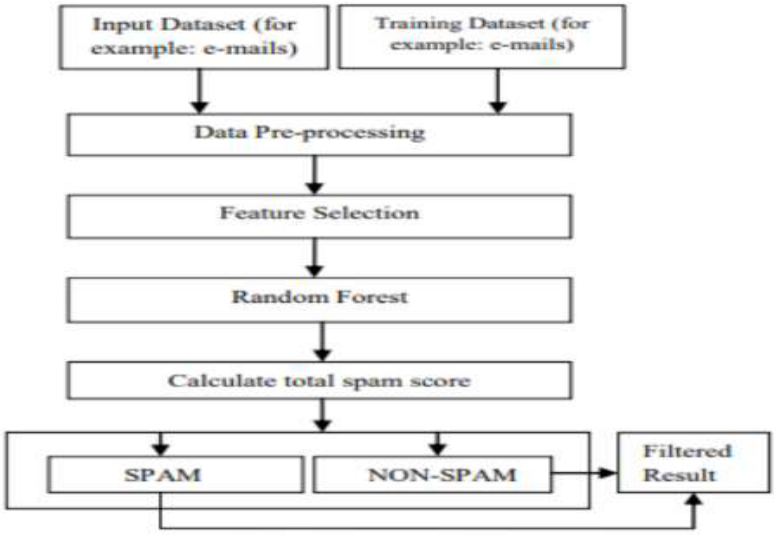
Methods:

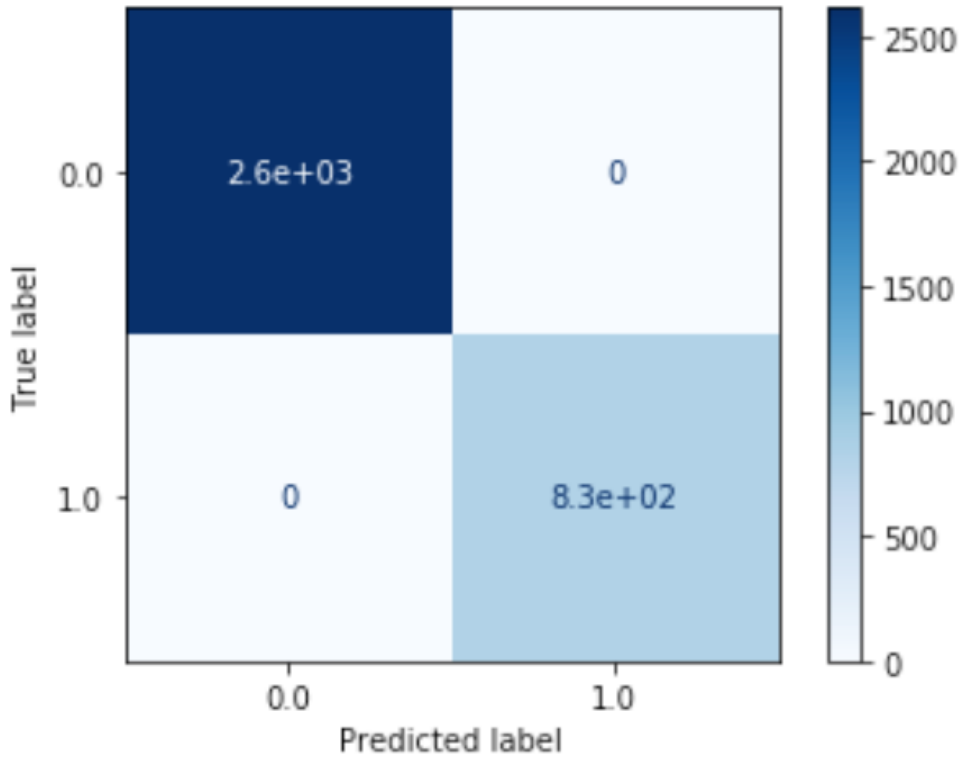
I'll start by performing an exploratory data analysis on the data. I'll run some plots to see how often certain words appear in spam and junk mail. I'll use Count Vectorizer to turn text documents into a vector of term counts during the preprocessing stage. Following data preprocessing, a classification algorithm will be used. I'll use the Random Forest Classifier, XGBoost Classifier, and LightGBM Classifier as my intended classifiers. The proposed model will

then be put into use, and a range of performance metrics will be used to evaluate its performance and accuracy.

Random Forest Classifier:

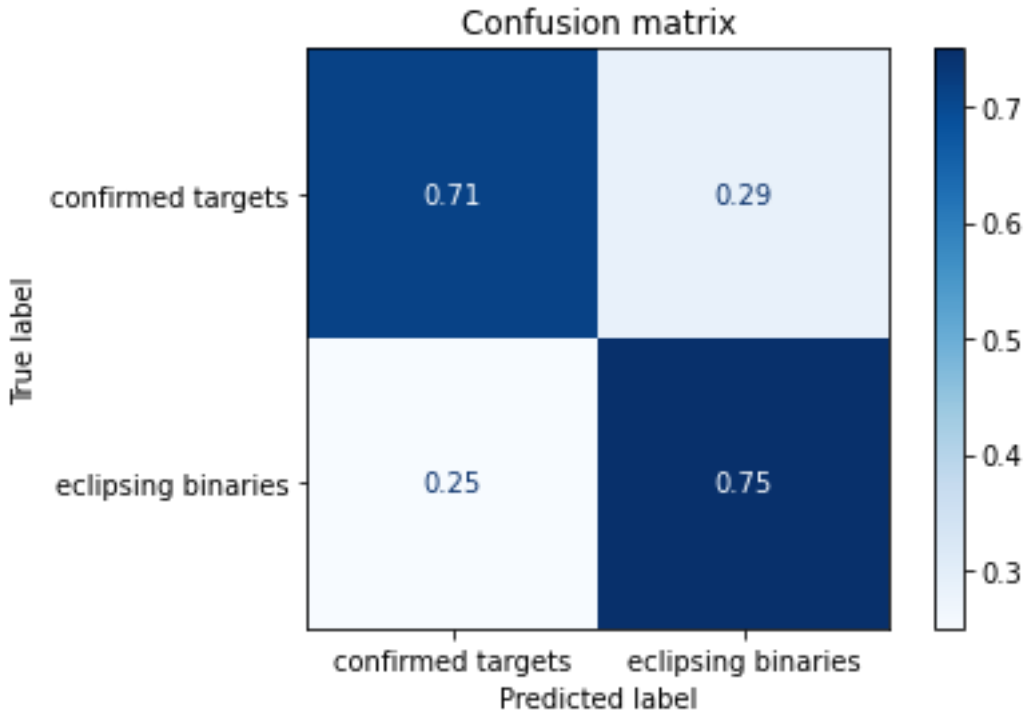
A random forest classifier in this case, the supervised machine learning method Random Forest is applied spam database A random forest is a meta-learner composed of many individual trees. Every tree vote for the general classifica0on of the dataset and chooses the random forest method the ra0ng with the most votes. Each decision tree is constructed from a randomly selected subset training data.





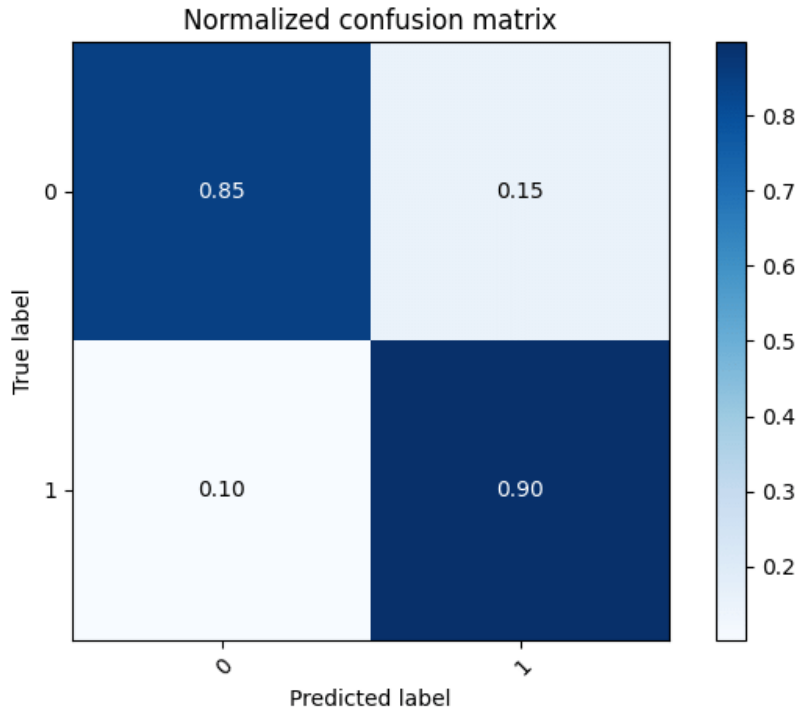
This step consists of tokenization, point selection, junking of stop words, and Lemmatization.

Confusion Matrix:



I've enforced several machine learning algorithms similar as Random Forest Classifier, XGBoost Classifier, and LightGBM Classifier.

Normalized confusion matrix:



After the perpetration of the model, I tested the models, and their delicacy and performance are estimated using colorful performance criteria.

Data Dictionary:

This dataset contains information gathered from 5572 dispatch dispatches. This is the dataset in which emails are collected at arbitrary and classified as spam or ham. The first column contains the spam/ ham bracket, and the remaining columns contain the dispatch communication.

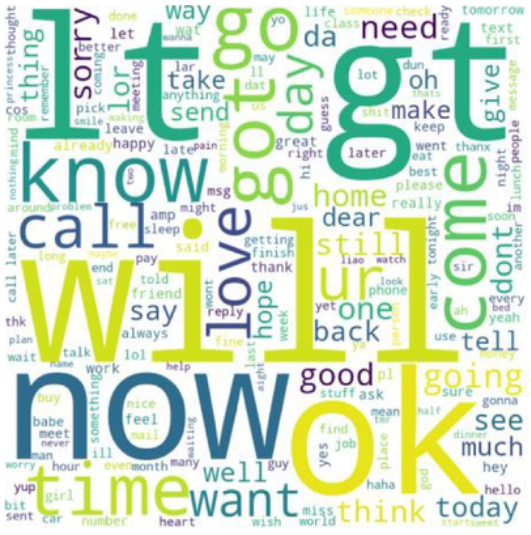
Exploratory Data Analysis:

I colluded a box plot of the communication length of spam and factual dispatches. Spam dispatches are on average about 24 words long. Spam dispatches range between 2 to 35 words. factual dispatches are on average about 14 words long. factual dispatches range between 1 to 171 words.

Word Cloud:

Word shadows, also known as label shadows, are graphical representations of word frequency that highlight words that appear constantly in a source textbook. The larger the word

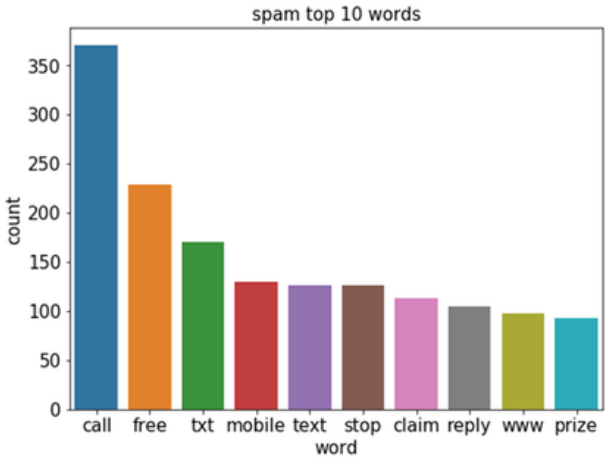
in the visual, the more constantly it appeared in the document. factual dispatches contain words like" I,"" me,"" my,"" but" and" that," which are more original to natural discussion. The spam order's word pall is easily distinct from the ham order, as it includes terms similar as free, prize, palm, and awarded. Spam dispatches are notorious for including" FREE" particulars and calls to action like" call/ claim now."

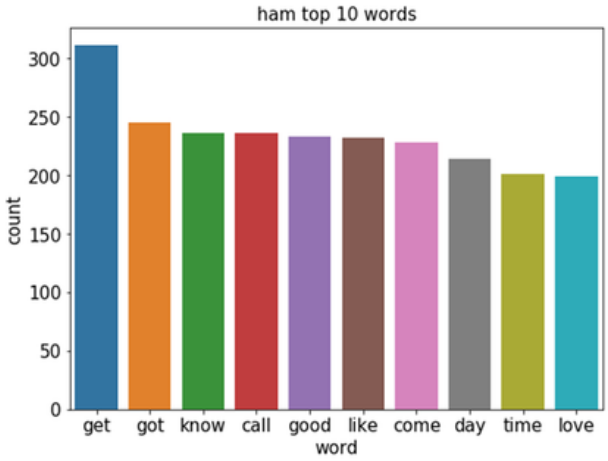


a) Ham

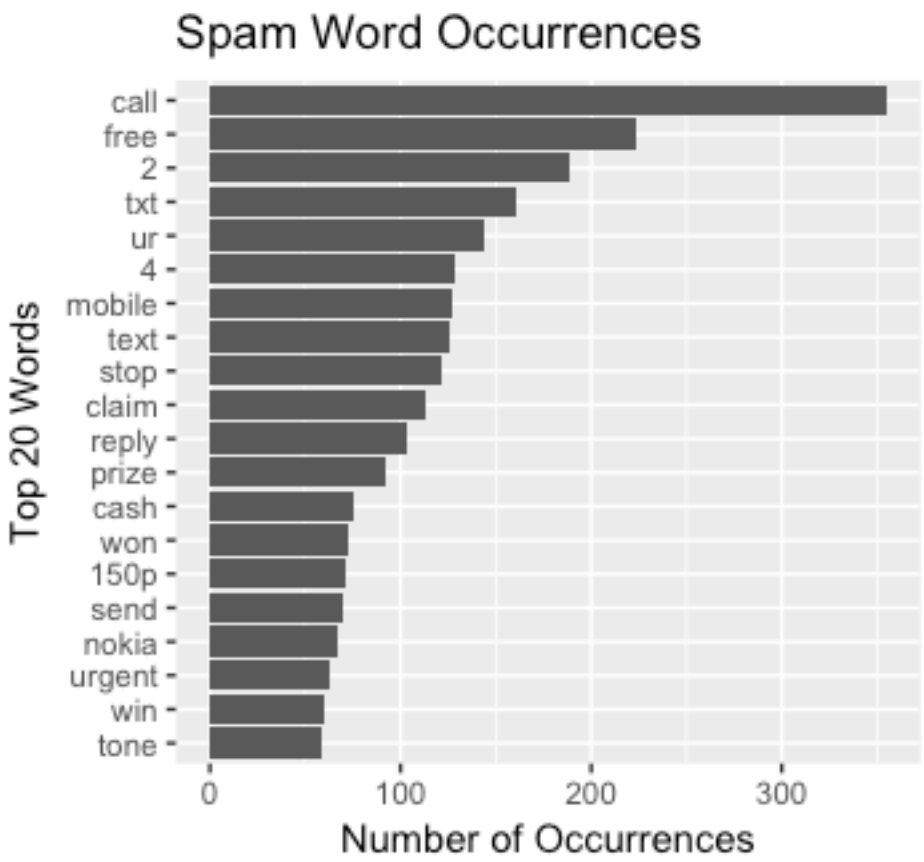
b) Spam

Rank of Spam and Ham terms:

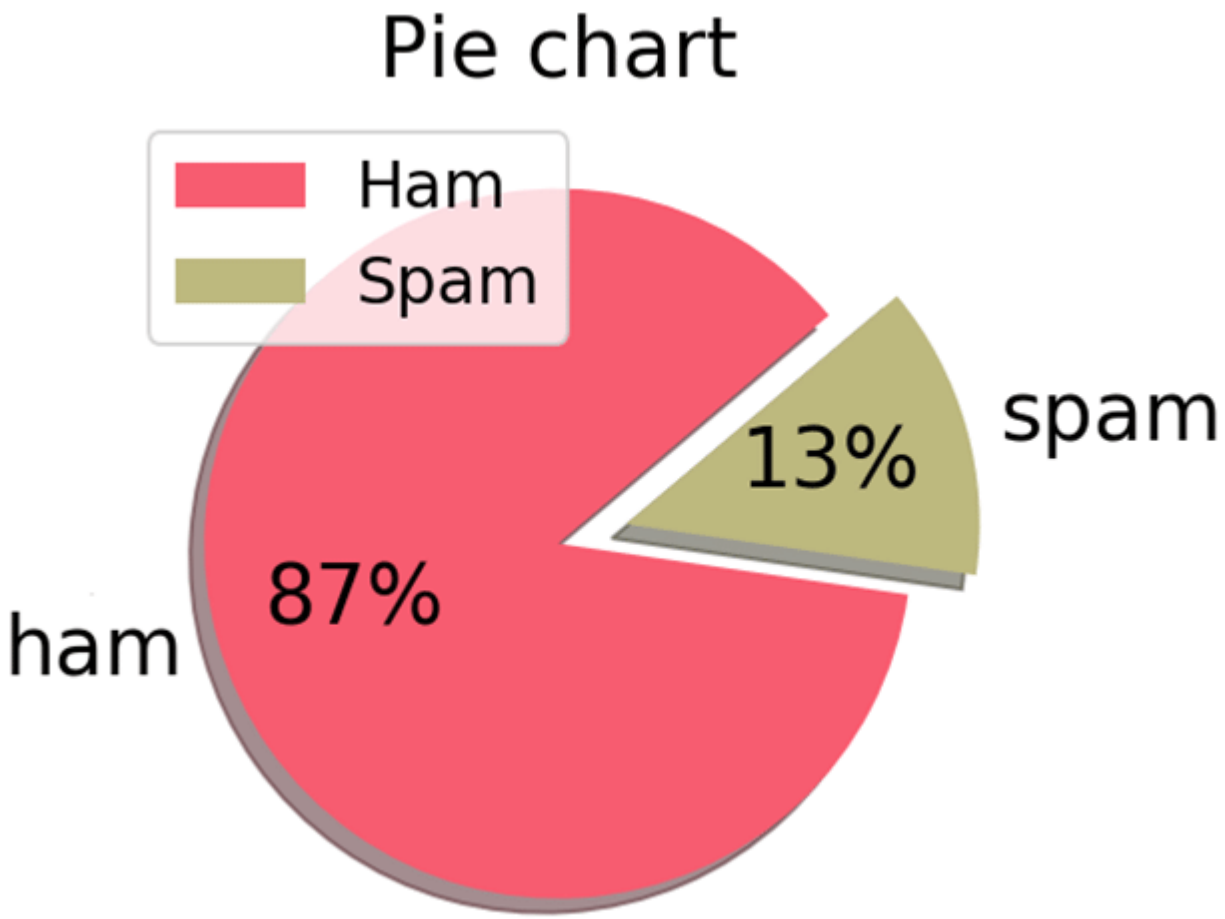




The top three words in the ham order were" u,"" 2," and" gt," which mean" you" and" to." call,'' 2', and' free' were the top three frequency terms in spam dispatches. In our diurnal lives, those are generally regarded as spam words.



Target Variable Distribution:



With 4825 values in the dataset, the Ham order accounts for86.6 of the aggregate. Because the dataset is relatively slanted, we'll use SMOTE to help level the data. The below plot shows the distribution of balanced data after applying SMOTE. The necessary stages that must be observed in the mining of data from a dispatch communication can be distributed into the ensuing stages. Pre-processing This is the first stage that's executed whenever incoming correspondence is entered.

Ethics-Related Matters:

The machine learning model must be created and used in a way that safeguards user data and privacy throughout the application. The same holds true for training data. If the training data violated the privacy of the individuals, ML models shouldn't be used.

Model comparison:

I developed various performance metrics to evaluate the performance of all models. The chart below compares the performance of all the classes I created. Accuracy is a parameter indicant the percentage of correct predictors. Compared to other models LightGB and Naive Bayes classifiers have high accuracies. I also calculated and compared other performance metrics such as precision, recall and F-Metric for performance evaluation models comparing the metrics for each model, I concluded that Naive Bayes classifier is the most efficient model with high precision, recall and F1 score for spam predictor email messages.

Conclusion:

Dispatch spam is one of the most pivotal in moment's world. Then, in this paper, I used algorithms like Naive Bayes, Decision Tree, and Random Forest bracket to descry the spam emails from that given dataset. The evaluation results illustrate that the proposed model got better delicacy with 95 compared with the current approaches.

Limitations:

Numerous models have advanced false positive rate than is needed, but it should be reduced to the smallest possible value in the future. Because utmost proposed models don't work well with real-time data, real- time spam bracket is desperately demanded.

Challenges:

Due to the failure of datasets for dispatch spam, constrained data and textbook written informally are the most likely issues that may beget the current algorithms to fail to meet prospects during bracket.

Perpetration Plan:

When a new communication arrives in the correspondence system, it's subordinated to textbook processing and successional pattern discovery to identify patterns in the dispatches. The classifier divides these patterns into spam, on-spam, and general emails. Spam emails are routed to the spam brochure. Non-spam emails are routed to the Inbox brochure. As a result, these spams are veritably effectively detected. This system can be enforced using different algorithms in the future, and further features can be added to the being system.

Ethical Assessment:

In this design, we work with data that could include sensitive or private information. The dispatch contains information, so there should be clear guidelines on what companies can and cannot do with the data they collect from guests. We need to save the data we use for the machine literacy model. It's the data scientist's responsibility to cover sensitive information of any kind. Data about guests or guests must be kept fully private.

References:

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://easychair.org/publications/open/Jvsw

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.mecs-press.org/ijeme/ijeme-v11-n4/IJEME-V11-N4-4.pdf

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://core.ac.uk/download/pdf/234676898.pdf