SPAM CLASSIFICATION

AYUSH VERMA | 16MF10031 | I.I.T. KHARAGPUR

The following report deals with the classification of a given text as spam or not. The main problem that gives the motivation of such a classification is that spam is one of the major threats posed to the users of e-mails, WhatsApp, YouTube etc. In 2013, 69.6% of all email flows were spam. Links in spam emails may lead users to websites with malware or phishing schemes, which can access and disrupt the receiver's computer system. These sites can also gather sensitive information from. Additionally, spam costs businesses around $2000 per employee per year due to decreased productivity. Therefore, an effective spam filtering technology is a significant contribution to the sustainability of the cyberspace and to our society. Here the classification is done for SMS and YouTube comments.

SPAM

According to Kaspersky Lab, the definition of spam is anonymous, unsolicited bulk email.

Let's take a closer look at each component of the definition:

Anonymous: real spam is sent with spoofed or harvested sender addresses to conceal the actual sender.

Mass mailing: real spam is sent in enormous quantities. Spammers make money from the small percentage of recipients that actually respond, so for spam to be cost-effective, the initial mails have to be high-volume.

Unsolicited: mailing lists, newsletters and other advertising materials that end users have opted to receive may resemble spam, but are actually legitimate mail. In other words, the same piece of mail can be classed as both spam and legitimate mail depending on whether or not the user elected to receive it

It should be highlighted that the words 'advertising' and 'commercial' are not used to define spam. Many spam messages are neither advertising nor any type of commercial proposition. In addition to offering goods and services, spam mailings can fall into the following categories:

- Political messages
- Quasi-charity appeals
- Financial scams
- Chain letters
- Fake spam being used to spread malware

Because some unsolicited correspondence may be of interest to the recipient, a quality anti-spam solution should be able to distinguish between true spam (unsolicited, bulk mailing) and unsolicited correspondence.

MACHINE LEARNING IN SPAM DETECTION

The problem of spam detection can be taken care of by the advancement of technologies and algorithmic approaches such as machine learning models. We can train a machine learning model to remember the features of a spam. This can be done using the supervised learning approach.

Supervised Learning; Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

Classification Algorithms(supervised):

1. Multinomial Naïve-Bayes Theorem
2. Logistic Regression
3. K-Nearest Neighbours
4. Ensemble Learning: Random Forest Classifier

Multinomial Naïve-Bayes

Naive Bayes classifier is a general term which refers to conditional independence of each of the features in the model, while Multinomial Naive Bayes classifier is a specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features.

It estimates the conditional probability of a particular word given a class as the relative frequency of term t in documents belonging to class. The variation takes into account the number of occurrences of term t in training documents from class, including multiple occurrences.

Logistic Regression

Logistic Regression is an algorithm that is relatively simple and powerful for deciding between two classes, i.e. it's a binary classifier. It basically gives a function that is a boundary between two different classes. It can be extended to handle more than two classes by a method referred to as "one-vs-all", which is really a collection of binary classifiers that just picks out the most likely class by looking at each class individually verses everything else and then picks the class that has the highest probability. It utilizes the Logistic function or Sigmoid function to predict a probability that the answer to some question is 1 or 0, yes or no, true or false, good or bad etc.

K-Nearest Neighbors

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

Ensemble Learning: Random Forest Classifier

Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem.

Random Forest is a supervised learning algorithm. Like you can already see from its name, it creates a forest and makes it somehow random. The forest it builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

DATA

The two main data used here are

1. SMS spam/ham dataset.
2. YouTube comments spam/ham dataset.

Data Description:

SMS spam/ham dataset

TYPE	COUNT
SPAM	747
HAM	4825

YouTube comments spam/ham dataset

TYPE	COUNT
SPAM	1005
HAM	951

A combination of the two datasets were taken to get the final data. This gave the dataset a diversity factor.

RESULTs

All the three datasets i.e. SMS-dataset, YouTube-dataset and the combined one were split into train and test sets and the accuracy of various models were observed.

On SMS-dataset

MODEL	ACCURACY
Multinomial Naïve-Bayes	0.98834
Logistic Regression	0.98026
K-Nearest Neighbors (k = 5)	0.91210
Random Forest Classifier	0.96412

On YouTube-dataset

MODEL	ACCURACY
Multinomial Naïve-Bayes	0.90051
Logistic Regression	0.95153
K-Nearest Neighbors (k = 5)	0.90051
Random Forest Classifier	0.94642

On Combined dataset

MODEL	ACCURACY
Multinomial Naïve-Bayes	0.96812
Logistic Regression	0.97543
K-Nearest Neighbors (k = 5)	0.89973
Random Forest Classifier	0.96414

OBSERVATIONS

1. For the SMS-dataset, as it is from a single source and type, the models fit the data extremely well.
2. For the YouTube-dataset, though the data is from a single source and type, but the lack of good volume of data fails the models to fit very well.
3. For the combined-dataset, there is a variety in data as it was from two different sources and the volume of the data also was more. So the models fit the data well. It handles the shortcomings of the YouTube dataset.
4. The volume of data available is a very important factor for an algorithm to fit in the best manner.
5. Spam classification is most suitably handled by LOGISTIC REGRESSION.

REFERENCE(s):

1. https://www.quora.com/How-does-multinomial-Naive-Bayes-work
2. https://en.wikipedia.org/wiki/Naive_Bayes_classifier#Multinomial_naive_Bayes
3. https://hbr.org/2017/07/whats-driving-the-machine-learning-explosion
4. https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd
5. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
6. https://www.pugetsystems.com/labs/hpc/Machine-Learning-and-Data-Science-Logistic-Regression-Theory-988/
7. https://securelist.com/threats/what-is-spam/
8. https://hackernoon.com/how-to-build-a-simple-spam-detecting-machine-learning-classifier-4471fe6b816e
9. https://hackernoon.com/how-to-build-a-simple-spam-detecting-machine-learning-classifier-4471fe6b816e
10. https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection

Repository: https://github.com/ayushverma209039/Spam-Classifier-Project