Spam/Ham Detection

Soft Computing Techniques

INT246

Submitted to – Dr. Sagar Pande

Ankit Singh

11911688

GITHUB LINK - https://github.com/codeantik/Spam-Ham-Detector

ACKNOWLEDEMENT

I would like to thank Dr. Sagar Pande, my teacher for INT246 for letting me make this project on the desired topic, i.e., spam/ham detection and for his proper guidance and valuable suggestions.

I would also extend my gratitude towards other faculty members of Computer Science & Engineering department for giving me an opportunity to learn and present the project.

I once again extend my sincere thanks to all of them.

TABLE OF CONTENTS

INTRODUCTION

Mobile message is a way of communication among the people, and billions of mobile device users exchange numerous messages. However, such type of communication is insecure due to lack of proper message filtering mechanisms. One cause of such insecurity is spam, and it makes the mobile message communication insecure. Spam is considered to be one of the serious problems in e-mail and instance message services. Spam is a junk mail or message. Spam e-mails and messages are unwanted for receivers which are sent to the users without their prior permission. It contains different forms such as adult content, selling item or services, and so on. The spam increased in these days due more mobile devices deployed in environment for e-mail and message communication. Currently, 85% of mails and messages received by mobile users are spam. The cost of mails and messages are very low for senders but high for receipts of these messages. The cost paid some time by service providers and the cost of spam can be measured in the loss of

human time and loss of important messages or mails. Due to these spam mails and messages, the values able e-mails and messages are affected because each user have limited Internet services, short time, and memory.

To handle these problems caused by the spam, researchers proposed different techniques to detect the spam e-mails and messages and secure the communication.

LIBRARIES

Numpy

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

Pandas

Pandas is a Python library used for working with data sets. It has functions for analysing, cleaning, exploring, and manipulating data.

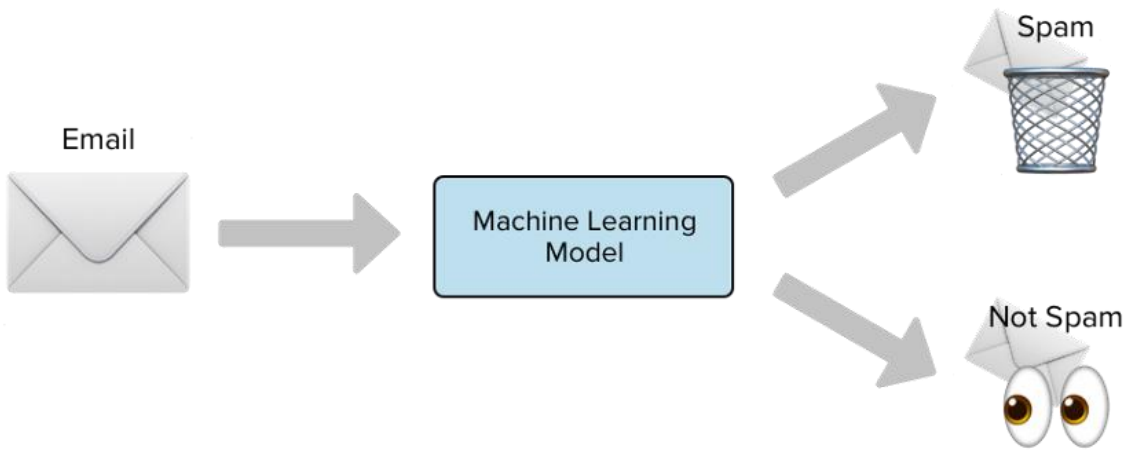
The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Pandas allows us to analyse big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science.

Sklearn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

LOGISTIC REGRESSION



Logistic regression is another powerful supervised ML algorithm used for binary classification problems (when target is categorical). The best way to think about logistic regression is that it is a linear regression but for classification problems. Logistic regression essentially uses a logistic function defined below to model a binary output variable (Tolles & Meurer, 2016). The primary difference between linear regression and logistic regression is that logistic regression's range is bounded between 0 and 1. In addition, as opposed to linear regression, logistic regression

does not require a linear relationship between inputs and output variables. This is due to applying a nonlinear log transformation to the odds ratio (will be defined shortly).

Logistic regression, despite its name, is a classification model rather than regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry. The logistic regression model, like the Adaline and perceptron, is a statistical method for binary classification that can be generalized to multiclass classification. Scikit-learn has a highly optimized version of logistic regression implementation, which supports multiclass classification task.

DATA COLLECTION & PREPROCESSING

The emails in the learning data are in plain text format. We need to convert the plain text into features that can represent the emails. Using these features, we can then use a learning algorithm on the emails. A number of pre-processing steps are first performed. We convert the plain text files to files with one word per line. In this project, we look at emails just as a collection of words. So, to make it easier we convert each file into a list of words using Bourne Shell Scripts (extractmultfiles.sh and extractwords.sh). The output files are named as ‘filename. Words.’

There are some English words which appear very frequently in all documents and so have no worth in representing the documents. These are called STOP WORDS and there is no harm in deleting them. Example: the, a, for etc. There are also some domains specific (in this case email) stop

words such as mon, tue, email, sender, from etc. So, we delete these words from all the files using a Bourne Shell Script. These words are put in a file ‘words.txt’. The shell script takes multiple files as an argument and then deletes all the stop words mentioned in the words.txt file.

FEATURE EXTRACTION & TRANSFORMATION

The feature extraction is an important part of the machine learning model, which has an extensive amount of data. Choosing correct features, which gives the correct accuracy and reduces the burden on the system. In this system, we used the words as our features. In this research, we used the top 1500 frequent words as features in the whole dataset. Basically, the feature extraction process comes after the tokenization. The reasoning behind considering only one feature is, the words come repeatedly in both emails. As the structure of spam emails is different from the ham emails in Marathi language.

MODELLING

For the Spam detection problem, we have tagged messages but we are not certain about new incoming messages. We will need a model which can tell us the probability of a message being Spam or Not Spam. Assuming in this example, 0 indicates — negative class (absence of spam) and 1 indicates — positive class (presence of spam), we will use logistic regression model.

So, first we define the model then fit the train data — this phase is called training your model. Once the training phase is finished, we can use the test split and predict the results. In order to check the accuracy of our model we can use accuracy score metric. This metric compares the predicted results with the obtained true results. After running above code, we got 96% accuracy.

In some cases, 96% might seems a good score. There are a lot other things we can do with the collected data in order to achieve more accuracy.

EVALUATION

In this section, we have considered the performance of training & testing data. This evaluation is calculated in terms of accuracy, precision, recall. Considering the importance of accuracy, the lowest accuracy is 94.95% and the highest is 96.57%, with Logistic regression. As I got the accuracy in other languages in the range between 90% to 98%, as the structure of language and available resources also more as compared to the English language. I got 97% Accuracy, although the size of the dataset is only 5600 emails. The comparison to other datasets such as we got decent accuracy.

CONCLUSION

Detection of spam is important for securing message and e-mail communication. The accurate detection of spam is a big issue, and many detection methods have been proposed by various researchers. However, these methods have a lack of capability to detect the spam accurately and efficiently. To solve this issue, we have proposed a method for spam detection using machine learning predictive models. The method is applied for the purpose of detection of spam. The experimental results obtained show that the proposed method has a high capability to detect spam. The proposed method achieved 97% accuracy which is high as compared with the other existing methods. Thus, the results suggest that the proposed method is more reliable for accurate and on-time detection of spam, and it will secure the communication systems of messages and e-mails.