TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 5292-5297

Comparative Analysis of GUI based Spam Message Classifier using Machine Learning Approach

[1]Manan Biyani, [2]Saumya Srishti, [3]R.S. Ponmagal

[1,2]UG Scholar, [3]Associate Professor
[1,2,3]Department of Computer Science and Engineering,
SRM Institute of Science and Technology,
Kattankulathur, Tamil Nadu
[1]mb6869@srmist.edu.in, [2]ss3303@srmist.edu.in, [3]ponmagas@srmist.edu.in

Article Info
Volume 83
Page Number: 5292-5297
Publication Issue:
May - June 2020

Article History
Article Received: *19 November 2019*
Revised: *27 January 2020*
Accepted: *24February 2020*
Publication: *16 May 2020*

Abstract

Generally, Spam messages are sent randomly or particularly to addresses by lazy advertisers and phishing criminals who wish to lead people to malicious and phishing sites. Spam detection is a significant application of Machine Learning on the internet today. Like a lot of other applications, machine learning models can be trained to distinguish between non-spam (ham) emails and spam. So, the aim is to examine and survey machine learning algorithms to identify one or multiple as best or better techniques to use in content-based spam filtering. Current spam techniques could be paired to increase effectiveness and to investigate ma-chine learning-based techniques for spam prediction results with the best accuracy. The analysis by supervised machine learning to capture information like variable identification, univariate, bivariate and multivariate analysis, data validation, cleaning, preparing, visualization is done on the entire dataset. This analysis will be a undogmatic guide to model parameters' sensitivity analysis about performance in the prediction of spam mails by utilizing the corresponding accuracy calculation. Additionally, the comparison of the performance of various machine learning algorithms from the given dataset with an evaluation of classification re-port, sensitivity, specificity, confusion matrix, and different other score metrics is performed to create a better picture of this evaluation. The result will show that the efficacy of proposed machine learning algorithm techniques can be compared with the best precision with Accuracy, Precision, Recall, and F1 Score.

***Keywords*:** Dataset, Python, Machine Learning, Spam, Spam Classification, Accuracy Result, Comparison, GUI.

1. Introduction

Machine learning is used to predict the future from past data. it is a subset of artificial intelligence that provides the machines the power to learn. The main focus of machine learning is to develop programs that have the ability to adapt itself when exposed to new to data. These programs or termed as machine learning algorithms and Python is used for their implementations. These algorithms are used to predict one or more outcomes, which are in the form of variables and are characterized as Red or blue. Machine learning has a great and significant role in today's life. It can be used to classify spam and non-spam text or emails as it deals with the study of a system that can learn from data. Use different approaches to establish a relation between the book and

May-June 2020
ISSN: 0193-4120 Page No. 5292-5297

the category SPAM or HAM like, based on the size of the message, word count, unique keywords. Then build classification models using different techniques to distinguish whether a message is spam. Compare the accuracy of each method and plot the accuracy graphs in a single bar plot. Also, generate a word-cloud for spam SMS. Before applying any supervised learning methods, we queried a the sample data and implemented the cleansing operations to get rid of messy and unwanted data since it has some broken context. For this task, we will be using various machine learning and natural language processing based algorithm to create a model that can classify dataset of SMS messages which consists of both spam and ham as one of either, based on the training we give to the model. At the end of each processing by the different classifiers, we plotted the confusion matrix to compare which one the best classifier for filtering SPAM SMS. We can classify the emails as spam or non-spam with a high number of emails, and it will be difficult to check the emails in large numbers. We can classify the emails as spam or non-spam with a high number of emails, and it will be difficult to check the emails in large numbers. The machine learning system is inputted with the labeled data from a training data set during training. Here, the large sets of emails are considered for input as the labeled training data are sets, which are then labeled as spam or ham by the different algorithms. During the process of training data sets, the classifier learns from the training data. Testing (Classification) during testing, a chine learning system is given unlabeled data.

Objectives:

The aim is to produce a machine learning model for real-time spam message prediction, to potentially replace the update-able supervised machine learning classification models by predicting results inthe form of best accuracy by comparing various supervised algorithms.This analysis will help us understand which algorithm is better for the prediction of the spam emails by using different metrics of performance, including precision, accuracy, and F1 score. To achieve used machine learning classification methods to fit a function that can predict the discrete class of new input of message as spam of ham (not spam). Here, doing so, is a learning exercise to:

1. Apply the machine learning fundamentals and hence interpret and comprehend the results and justify the interpretation based on the observed dataset..

2. Create python notebooks to serve as computational records and document our thought process and investigate the spam or non-spam message details to analyse the data set.

3. Evaluate and analyse statistical and visualized results, with which we find patterns to create a comprehensive comparison.

Scope:

The scope is to investigate a public dataset of spam records for the ISP sector using ML techniques. When identifying spam, the message is more complicated. It adapts well to the future spam techniques and hence provides sensitivity to the clients. Instead of single words, it considers the complete length of the message. It increases security and control Therefore the network security and its administration costs are reduced.

1. It adapts well to the future spam techniques and hence provides sensitivity to the clients
2. Instead of single words, it considers the complete length of the message (including all words and phrases).
3. It increases security and control
4. Therefore the network security and its administration costs are reduced.

Paper Goals:

- **Exploration data analysis of variable identification**

1. Loading the given dataset
2. Import required libraries packages
3. Analyze the general properties
4. Find duplicate and missing values
5. Checking unique and count values

- **Uni-variate data analysis**

1. Rename, add data and drop the data
2. To specify a data type

- **Exploration data analysis of bi-variate and multi-variate**

1. Plot diagram of pair plot, heatmap, bar chart and Histogram

- **Method of Outlier detection with feature engineering**

1. Pre-processing the given dataset
2. Splitting the test and training dataset
3. Comparing the Decision tree and Logistic regression model and random forest etc

- **Analyzing algorithm to predict the result**

1. Based on the best accuracy

2. Literature Survey

A. An Efficient Approach of Spam Detection in Twitter

As mentioned, spam in tweets is a fundamental issue these days. Work in this literature uses actual components of tweets along with machine learning systems for twitter spam discovery. An informational collection of marked tweets is made with which it is observed that measurable properties of spam tweets regularly fluctuate after some time, and the execution of existing machine learning-based classifiers diminishes along these lines. This issue is called "Twitter spam drift". An end goal based approach, the Lfun scheme is used, which dismantles unlabeled tweets and fuses them into the classifier's training process to find changed spam tweets. Spam tweets are further located by using the new dataset

May-June 2020
ISSN: 0193-4120 Page No. 5292-5297

created, which will be again fed into the process with the unlabelled tweets. The proposed scheme also removes counterproductive information by removing old tweets of a specific limit out of the cycle, which will save space and process time.

B. An Improved Spam Detection Method With Weighted Support Vector Machine

The email has always been one of the fastest ways of message or text transferring techniques that use the internet. Spam messages have become a threat to the commoners and are supposed to be eradicated. Nowadays, spams have become common, and hence there is a need for spam filtration method. Support Vector Machine or SVM, one of the machine learning algorithms, is used to detect these fraud texts. Weight variables are extracted using the KFCM algorithm, which comes underweighted SVM. Different classes are represented by these weighted variables. We have evaluated these variables' impact in order to detect the spam text.

C. Conventional and Ontology-Based Spam Filtering

Emails have become an essential tool in the communication field, and are inevitable in current times. The number of users is increasing day-by-day due to the much relative ease of use compared to other methods. With the increased number of email users, the number of spam emails has also increased. Spam cause significant monetary are reputation loss to users as well as internet and communication service providers. Many subjective spam filtering techniques have been employed to distinguish between non-spam and spam. A mail can be subjective such that it appears as spam to some and may appear as ham to another user. That is, it depends on the personal preference of the user. Ontology-based personalized mail access helps in solving this issue to some extent. This paper compares how ontology based spam filtration is advantageous in some cases compared to traditional spam filtering.

D. Adaptive Classification for Spam Detection on Twitter with Specific Data

At present, the usage of Twitter has increased rapidly. The users quickly find information that tweets to their Twitter account. There are 2 types one information being shared, one being shared for the benefits made and the on the other hand there exist the spammers ready to do fraud. These people are like criminals to ordinary people and can easily misuse their data, specifically the personal data. They do this by promoting their fraud websites and ads and asking people to click on the links present on the texts that are sent by them. Researchers have worked and hence came up with a work which is used by streaming spam detection method. They propose use of predefined append-able spam word lists and URL-based security tools for adaptive data classification for spam detection. They analyzed data by the Naïve Bayes algorithm including overall data and particular cased data. Spam classifier performance is greatly improved by doing the same. They have showed results of the above thing using experiments.

E. Enhancing the Naive Bayes Spam Filter through Intelligent Text Modification Detection

Spam emails affect the security of the computer as they cause both financial and data phishing problems. In the past few years, there has been a significant emergence of other social networking platforms, but the use of emails still remains consistent. There were many spam filters launched to prevent fraud, but their research focuses on text modification and hence lags in some or the other way. Today, Naïve Bayes is very popular for the detection or spam classification. It is considered efficient and straightforward. Naïve Bayes can detect the text which contains diacritics or leetspeak. This algorithm is known for its accuracy. Hence to increase the efficiency of the Naïve Bayes algorithm, they have implemented this novel algorithm. They have used a combination of semantic-based, keyword-based, and even machine learning algorithms and applied them using python algorithms. By doing this, they discovered a relationship between the length of the email and the spam score. Apart from that, they also discovered a graphical relationship between the length of texts and the spam number or score. This proves that spammers use Bayesian Poisoning, which is considered to be unreal.

3. Proposed Work

A. Exploratory Data Analysis

Various machine learning algorithms will be applied to extract patterns and to obtain results with maximum accuracy. The dataset will be thoroughly examined for patterns in the messages that can be characterized in classifying a message as spam or ham.

B. Data Wrangling

This section will feed the data, check for validation, and then verify and clean the given dataset for analysis. Also, make sure that the document steps carefully and justify cleaning decisions.

C. Data Collection

The collected data set for predicting given data is split into a Training set and a Test set. 7:3 ratio of the Training set and Test set will be used to train the model considering the size of the data is moderate. The Data Model created using the classification algorithms will be applied to the Training set, and based on the test result accuracy. Test set prediction is made. The data gathered from the performance of these algorithms will be analyzed.

Building the classification model

The spam classification problem can be attempted using many machine learning algorithms and text classification methods. This paper attempts at using traditional algorithms like K-Nearest Neighbor, Support Vector Machines, Logistic Regression, Random Forests,

May-June 2020
ISSN: 0193-4120 Page No. 5292-5297

Decision Tree, Naive Bayes, as well as, work on term frequency-inverse document frequency and Bag of words based classification technique.

Advantages

This paper improves the accuracy score by comparing popular machine learning algorithms. Finally, it highlights some observations on future research issues, challenges, and needs.

Preparing the Dataset

The machine learning model is fed with a dataset based on the old dataset, and therefore the model is trained. Initially, the accumulation of information, data from previously patients datasets from online sources are gathered together (like, www.kaggle.com / www.data.gov.in). A common dataset is formed by merging these datasets on which the analysis is done.

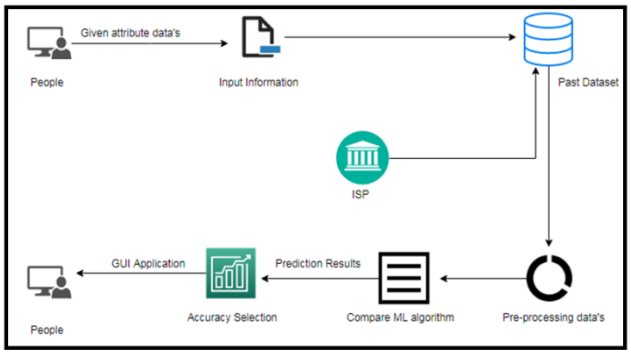

Figure 1: Architecture Diagram

4. Implementation

A. Data Validation and Pre-Processing Techniques:

This module includes validation, cleaning, and pre-processing the data for further tasks we will be performing. Some processes included in this task will be-removing unwanted columns, checking and removing null values, checking for the format, re-moving duplicate values.

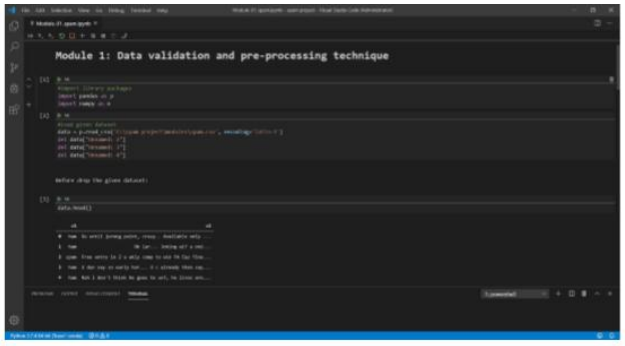

Figure 2(a): Data Validation and Pre-Processing Techniques

B. Exploration Data Analysis of visualisation and Pre-processing Techniques:

In this module, the clean and validated data will be explored for anomalies and patterns. The distribution of 'spam' vs 'ham' and the particular words involved in deciding whether the message should be classified as either spam or not. The frequency and length of words will be one of the major factors used in making the decision. Also, by applying Natural Language Processing tokens of a word will be created and visualisation of these words along with graphs will be made.

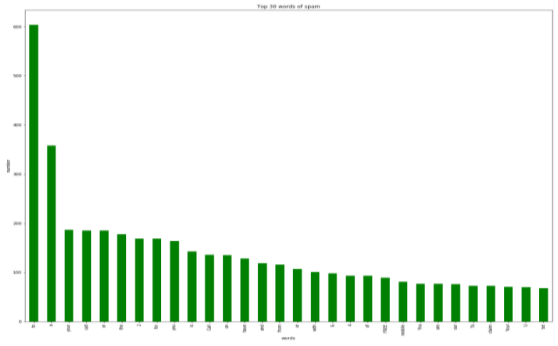

Figure 2(b): Most frequent words appearing in "spam" labeled data

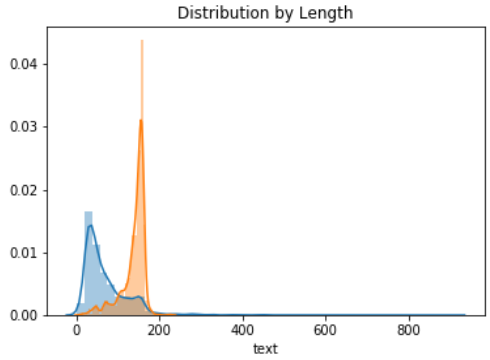

Figure 2(c): Distribution graph of spam vs ham messages by length (orange - spam, blue - ham)

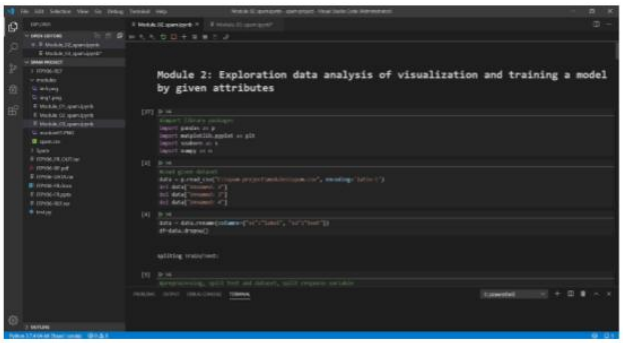

Figure 2(d): Exploration Data Analysis of visualisation and Pre-processing Techniques

TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 5292-5297

C. Performance measurements of Logistic Regression and Decision tree algorithms:

In this module, the validated data will be fitted into Logistic Regression and Decision Tree algorithms. The results (accuracy, precision, recall, F-1 score) will be recorded for both the algorithms for further comparisons with other algorithms.

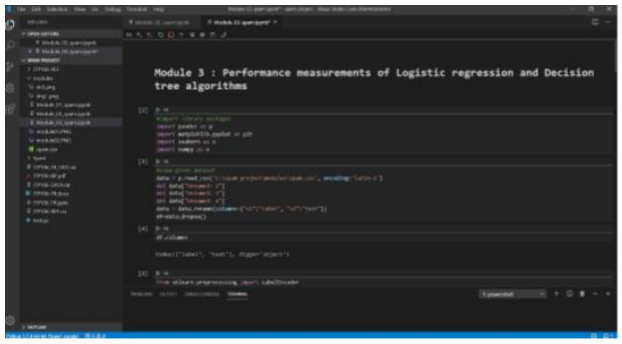

Figure 2(e): Logistic Regression and Decision tree algorithms

D. Data analysis based on Support Vector Machines (SVM) and Random Forest:

This is a classifier that is used to classify the data set by drawing an imaginary hyperplane between data. This model has a high predictability rate as it used different kernelling functions. Random forests, which are also known as random decision forests, are the learning methods. This algorithm is operated by constructing decision tree multitude at the time of training and is used for tasks such as classification and regression. Hense, produces an output of a class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

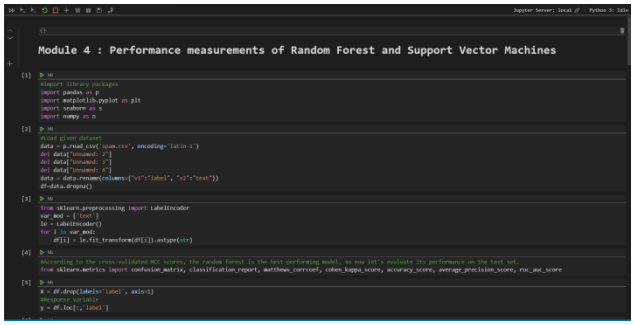

Figure 2(f): Random Forest and SVC

E. Data analysis based on *K*-Nearest Neighbor (KNN) and Naïve Bayes Algorithm:

K-Nearest Neighbor supervised machine learning algorithm analyzes the 'k' nearest instances in n-dimensional space when any discrete data set is received. It then returns the most common class among these nearest instances, which will be the predicted class. Also, in case of finding a discrete point of prediction, it returns the mean of k nearest neighbors for the real valued data. The Naive Bayes algorithm is a probabilistic model for supervised machine learning. It is "naive" as it assumes that all the included attributes belonging to each class and their probabilities are mutually exclusive and independent. Usually, KNN is robust to noisy data as it assumes the k-nearest neighbors, which is a strong assumption but fast and effective.

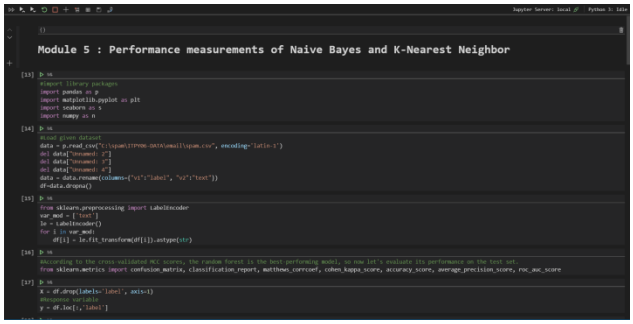

Figure 2(g): Naive Bayes and KNN

F. Graphical User Interface:

We use the tkinter library for creating an application of UI, for creating windows, and all other graphical user interfaces and Tkinter will come with Python as a standard package. It can be used for the security purpose of each user or accountants. There will be two kinds of pages, like registration user purpose and login entry purpose of users.

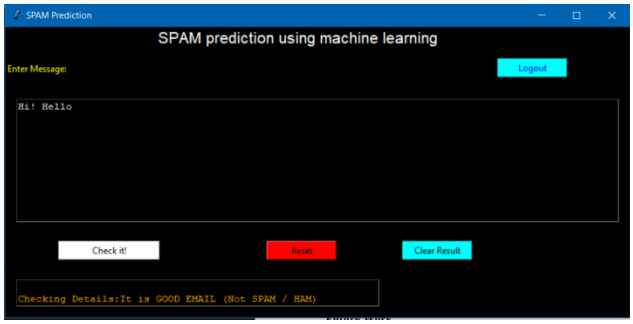

Figure 2(h): Spam classifier GUI window

5. Result and Discussion

Overall, a comparison of Naive Bayes, Random Forest Classifier, SVC, Decision Tree, Logistic Regression, K-Nearest Neighbor, and TF-IDF was made on a spam classifier in this paper. Conclusively, the Term Frequency - Inverse Document Frequency spam classifier method was the most accurate with a 94.91% accuracy score. Notably, K- Nearest Neighbor also scored 91.02% accuracy with 5 neighbors in use. Additionally, spam based GUI classifier was also developed to show the classifier results. The following tables and charts depict this result.

Table 1: Machine Learning Algorithm Performance

Parameters	LR	DT	RF	SVC	NB	KNN
Precision	0.87	0.95	0.95	0.89	0.87	0.93
Recall	1	0.95	0.95	1	1	0.97
F1-Score	0.93	0.95	0.95	0.94	0.93	0.95
Accuracy (%)	86.60	90.84	90.90	89.05	86.60	91.02

Table 2: Spam Classification method Performance

Parameters	SC (TF-IDF)	SC (BOW)
Precision	0.89	0.87
Recall	0.72	0.57
F1-Score	0.80	0.69
Accuracy (%)	94.91	92.73

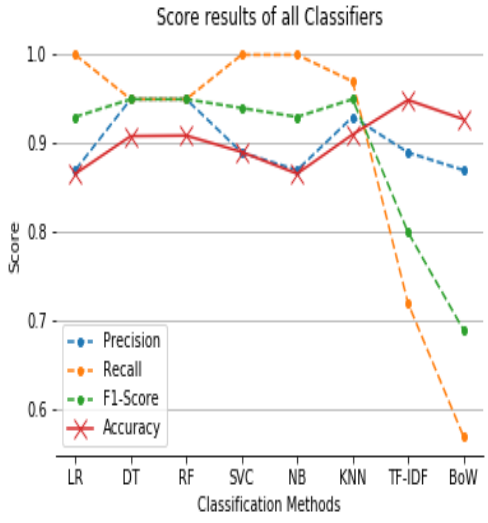

Figure 3: Score comparison across various classification methods

6. Conclusion

Spam has become a major issue for every email and messaging service user. Spam classifier, as an important feature in mailing and SMS systems, has already become pretty advanced, but, this paper aims at tackling the methods that make the classifier better for the user. And hence this analysis will serve a broad-based guide to the model parameters' sensitivity analysis about performance in the prediction of spam mails using accuracy calculation.

References

[1] An Efficient Approach of Spam Detection in Twitter, Rutuja Katpatal & Aparna Junnarkar, Department of Computer Engineering, P. E. S. Modern College of Engineering, Pune-05, India 2018

[2] An Improved Spam Detection Method With Weighted Support Vector Machine, Vishagini V & Archana K Rajany, Dept of Computer Science and Engineering, Amrita School of Engineering, Amritapuri, Amrita Vishwa Vidyapeetham, India 2018

[3] Conventional and Ontology Based Spam Filtering, Nasreen M Shajideen & Bindu V, Assistant Professor, Department of ECE, Sree Chitra Thirunal College of Engineering, Thiruvananthapuram, India 2018

[4] Adaptive Classification for Spam Detection on Twitter with Specific Data, Thayakorn Dangkesee & Sutheera Puntheeranurak, Department of Computer Engineering, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand 2017

[5] Enhancing the Naive Bayes Spam Filter through Intelligent Text Modification Detection, Linda Huang, Julia Jia, Emma Ingram & Wuxu Peng, Department of Computer Science, Texas State University, San Marcos, TX USA,2018