Data and Web Mining Project Report

SMS SPAM DETECTION

Submitted

by

N. Praneeth Babu	-AP19110010383
CH. Divyesh	-AP19110010478
N. Sai Bhuvanesh	-AP19110010496
B. Vardhan Kumar Reddy	-AP19110010493
G. Balaji	-AP19110010517

Department of Computer Science and Engineering

SRM University-AP, Andhra Pradesh, India

May - 2022

1.Abstract

Data mining is the process of “Extracting” or "mining" useful information from big data. Nowadays, Short Message Services (SMS) is the most popular way to communicate with mobile users by texting. Because it is the cheapest way to communicate with others. The document can be set in pre-defined categories based on the text content and output features. Contains essential applications for spam filtering and text mining. Due to the increasing availability of data and documents daily in digital form and to ensure the need for self-organization. In the research community, the approach to documenting is based on data mining techniques. The beauty of this fact is that a particular company or sender of spam uses this service to advertise and send unwanted messages to mobile users and create communication interruptions, other real threats such as fraud, identity theft, and malicious computer programming and network bandwidth. To overcome this problem spam filtering techniques are used.

2.Introduction

Short message service (SMS) traffic continues to rise daily. As a result, mobile attacks, such as spam blasting the service via spam messages sent to recipient groups, have grown significantly. Data mining is also known as site data acquisition and involves various techniques such as classification, and integration. In today's world, Short Message Service (SMS) is the most popular and frequently used communication method for a variety of services such as bank updates, agricultural information, aviation updates, and more. SMS is the cheapest way to communicate in developing countries like India. Since the introduction of the mobile phone, we have seen improvements in the devices and services offered by the mobile network. Text filtering is an easy and inexpensive way to communicate. Millions of people use SMS to communicate in their daily lives, yet the most common problem users face is spam. The definition of spam SMS is different from that of spam emails or spam SMS. In short, spam SMS can be defined as "Unsolicited Messages" which are unwanted by the recipient and sent by the spam sender. The firm and Spammer have used this service to advertise and promote because of their low prices. This message is not fully functional for the user and its message is using network bandwidth thus reducing Network Performance. Therefore, the main purpose of Spam SMS filters is to reduce or block unwanted spam by spammers.

3.Literature Review

There are many methods used to filter Spam SMS using data mining techniques. The authors have used the novel framework for SMS Spam

Filter using two different methods for selecting the Future based on information gain and Chisquare metrics acquiring a Distinctive feature representing SMS messages. To improve personal and private SMS Spam filters to reduce communication costs and hardware costs and research the purpose of Improving Spam Filtering systems on mobile phones to provide the user with an independent, private, secure, personal, easy, updateable, filtering system.

The authors used Approved Simulation (SB), Real-Time Web (RBB), and Real-Based System (RSB) Gobble trusted management (GTM) is used to design and implement a Spam Management System based on trusted managers. The authors used Bayesian filters to improve the SMSAssassin SMS Spam filter system. They are developing a SMSAssassin app for both Android and Symbian phones. The limited flexibility offered in the current inbox designs and unwanted content growing in the SMS channel and the user is disturbed. So, the author introduces a solution to these kinds of problems to design a SMSAssassin app that replaces the current inbox and can enable spam filtering to give the user control over this type of content-based filtering to provide Limited Space Skills.

The Proposed way to receive and reset Spam Message Spam. System Detects Spam Message by checking the relationship between the Sender and Recipient and the content of the message. If the system does not detect the relationship between the sender and the recipient and if the message content is detected as Spam, then the System will treat it as spam and forward the Message and the Spam or Reject it. Authors of various algorithms used to Filter Spam in Mobile messages. This paper compares other filtering methods for SMS Spam Croups publicly available. The Bayesian method worked very well as they expected to provide the highest Success of up to 98% so the Bayesian method is the best way to filter out SMS Spam for Text Messages.

In Authors have used the algorithm that Bayesian Filter Strategies can easily transmit SMS spam. The author has used the combination of Naïve Bye and the flexible environment in Algorithms to filter Spam SMS. Many spam detection methods are unable to detect these scanners because the normal training of these scanners has not been done yet, the Spam site needs to be updated. Dynamic training has therefore improved spam filtering techniques. Support Vector Machine algorithms (SVM) and Naïve Bayesian (NB) are used to filter Thai - English spam. Two Spam SMS methods are used to filter the first easy-to-use method of filtering the current English spam filter to filter and improve Thai language support and the second method uses to customize text, word classification, and Thai Semanti, c-word analysis.

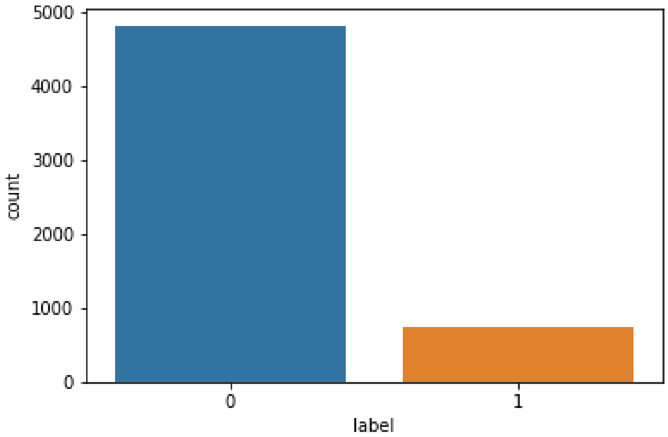
4.Data Set Description:

The SMS (text) data was downloaded from UCI datasets. It contains 5,572 SMS phone messages. The data were collected for the purpose of mobile phone spam research and have already been labeled as either spam or ham**.**

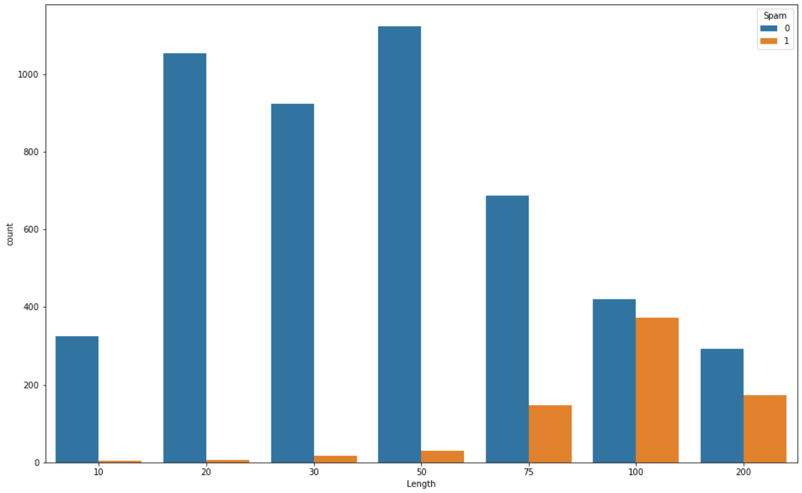
5.Proposed Model

5.1Data Preprocessing

The dataset we are using have 5columns and 5572 rows. In dataset both ham and spam examples are given. In that 5colums three of them are having the null values. By using the isnull() function we find null values and removed those columns. The ham messages are 4872 and spam messages are 747. We have changed the ham and spam with the label of 0 and 1. The below figure shows the count plot of the label.



We removed all special characters and stopwords also. We find the length of each message. And print the bar chart of the message’s length with the count. The below figure shows the bar chart.



We can calculate number of spam words and the number of ham words. The no. of spam words is 219 and ham words are 59. We split the data into train data and test data.

5.2Model

In the model building we used Random Forest Classifier, SVM and Gaussian Naïve Bayes.

Random forest is a Supervised Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

And Support Vector Machine is also a Supervised Learning algorithm. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes.

Gaussian Naive Bayes supports continuous valued features and models each as conforming to a normal distribution. An approach to create a simple model is to assume that the data is described by a Gaussian distribution with no co-variance (independent dimensions) between dimensions. This model can be fit by simply finding the mean and standard deviation of the points within each label.

Metric of Gaussian NB is:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

And to know the performance of the classifiers we used Confusion Matrix and Classification Report belongs to the Supervised leaning methods in our model. The confusion matrix is an N x N table (where N is the number of classes) that contains the number of correct and incorrect predictions of the classification model. To create the confusion matrix, we use sklearn confusion_matrix() function, which takes the real values (y_test) and the predicted values (y_predict). We use seaborn to print a heatmap of the confusion matrix. The values returned by the confusion matrix are divided into the following categories:

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)

By using those categories, we find the accuracy, recall, precision, F1-score and specificity.

Metrics of these are as follows:

Accuracy = (TP + TN)/ (TP + TN + FP + FN)

Precision = TP/ (TP + FP)

Recall = TP/ (TP + FN)

Specificity: TN/ (TN + FP)

F1-score = (2 x Precision x Recall)/ (Precision + Recall)

And the classification report also uses these values and apart from these this will calculate the support, macro average and weighted average.

Support: number of observations for each class.

Macro average: the arithmetic average of a metric between the two classes.

Weighted average: the weighted average is calculated by dividing sum (metric of interest x weight) by sum(weights).

Metrics are as follows for the precision:

Macro average(precision) = (p0 + p1)/ 2 **[p0 and p1 are precision of ham and spam messages]**

Weighted average (precision) = (p0 x s0 + p1 x s1)/ (s0 + s1) **[s0 and s1 are support of ham and spam messages]**

6.Results

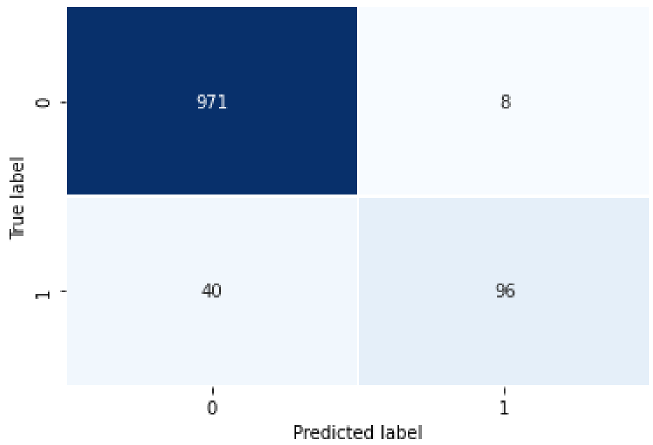
➔After implementing these models, we find out the accuracy for each classifier to check which on will be more fit to our model. And among them Random Forest Classifier has the more accuracy than other classifiers. The accuracy of each classifier are as follows:

Accuracy with Random Forest Classifier: 0.95695067264574

Accuracy with SVC: 0.8780269058295964

Accuracy with Gaussian Naïve Bayesian: 0.9426008968609866

➔Output for the confusion matrix and the values of the catogires as follows:



TP=96 **FP=8**

TN=971 **FN=40**

➔The following are the results of calculation done by using the above values.

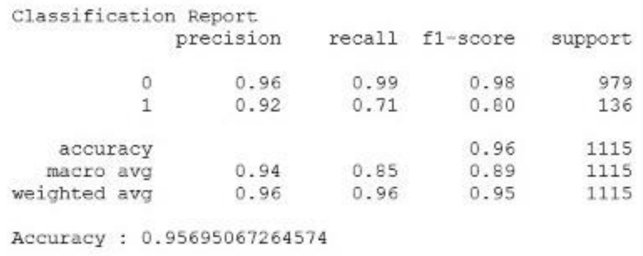
Accuracy = 0.96

Precision = 0.92

Recall = 0.71

F1-Score = 0.80

➔Output of the classification report is given below.



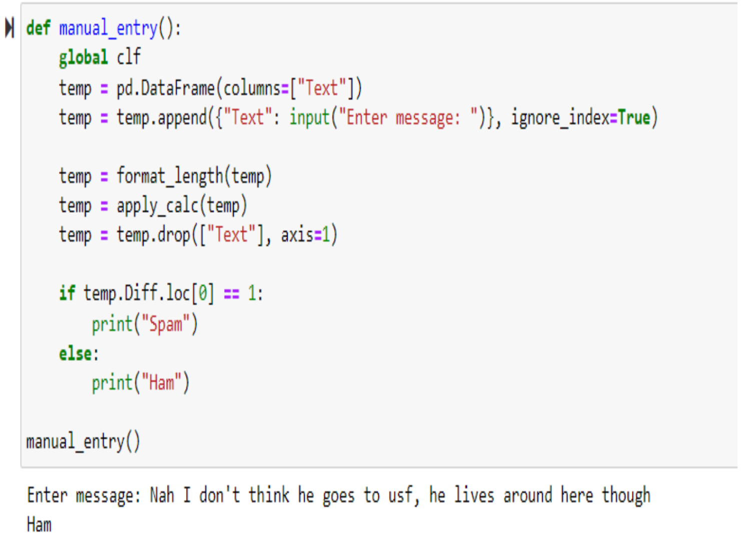
```
Classification Report
              precision    recall  f1-score   support

           0       0.96      0.99      0.98       979
           1       0.92      0.71      0.80       136

    accuracy                           0.96      1115
   macro avg       0.94      0.85      0.89      1115
weighted avg       0.96      0.96      0.95      1115

Accuracy :  0.95695067264574
```

➔These are some predictions made:



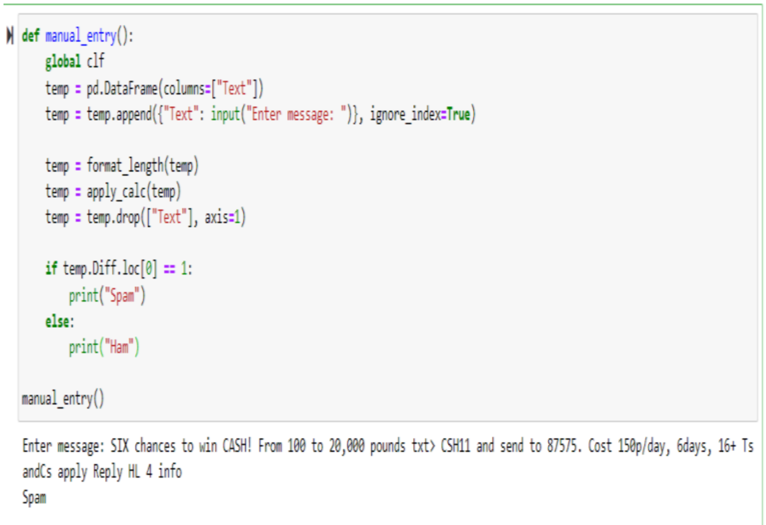
```
def manual_entry():
    global clf
    temp = pd.DataFrame(columns=["Text"])
    temp = temp.append({"Text": input("Enter message: ")}, ignore_index=True)

    temp = format_length(temp)
    temp = apply_calc(temp)
    temp = temp.drop(["Text"], axis=1)

    if temp.Diff.loc[0] == 1:
        print("Spam")
    else:
        print("Ham")

manual_entry()
```

```
Enter message: Nah I don't think he goes to usf, he lives around here though
Ham
```



```
def manual_entry():
    global clf
    temp = pd.DataFrame(columns=["Text"])
    temp = temp.append({"Text": input("Enter message: ")}, ignore_index=True)

    temp = format_length(temp)
    temp = apply_calc(temp)
    temp = temp.drop(["Text"], axis=1)

    if temp.Diff.loc[0] == 1:
        print("Spam")
    else:
        print("Ham")

manual_entry()
```

```
Enter message: SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ Ts
andCs apply Reply HL 4 info
Spam
```

7.Conclusion

Training the data using supervised learning techniques and calculated the accuracy for the each classifier and find out the best one among them. But still we used the confusion matrix and compute the accuracy, precision etc. And produce the classification report and evaluate the code and made predictions. By the produced output the model was working well and producing the correct prediction.

8.References

1. https://www.ijarcce.com/upload/2016/november-16/IJARCCE%2041.pdf
2. https://www.sciencedirect.com/science/article/pii/S1877050919318617
3. https://www.researchgate.net/publication/326391846_SMS_spam_detection_using_association_rule_mining_based_on_sms_structural_features
4. https://medium.com/swlh/confusion-matrix-and-classification-report-88105288d48f#:~:text=The%20confusion%20matrix%20is%20an,the%20predicted%20values%20(y_predict)
5. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
6. https://www.javatpoint.com/machine-learning-naive-bayes-classifier
7. https://www.javatpoint.com/machine-learning-random-forest-algorithm