Name: Ankit Anil Lade

Roll no.: 18038

Indian Institute of Science Education and Research, Bhopal

Program Name - DSE407

Problem Release date: August 25, 2021
Date of Submission: September 01, 2021

SMS Spam Classification

September 1, 2021

1 Introduction

In this home task I have done classification of spam SMS. We all know that there are a lot of spam messages and mails sent daily. These spam messages or mails can be used for many illegal purposes. To overcome such issues I and many other people try to classify the messages that are spam or not. In this task we try to perform Naive Bayes, Logistic Regression and Support vector machine algorithms to determine if the given message is spam or ham(not spam. For this task we have taken the data set from the UCI repository which provides data set for free for the projects. This corpus data set has been collected from free or free for research sources at the Internet by the people at UCI[1]. It is a labelled data set with 5572 corpus with ‘spam‘ or ‘ham‘ as labels.

2 Methods

First and very important step before using a data set is to perform text preprocessing to make it clean and helps in training the model. I did stemming on the given data set to produce morphological variants of a root/base word. Then I convert a collection of raw documents to a matrix of TF-IDF features using TfidfVectorizer[2]. TF-IDF stands for **Term Frequency — Inverse Document Frequency**. This step was basically a Feature extraction step. After this step we try to classify the class of messages as ’spam’ or ’ham’. The last and most important step is to evaluate the predicted class. This step also tells us about the performance of the algorithms with their specific parameters. Here we compare the precision, recall and f1-scores which is discussed in evaluation section. For the evaluation purpose we split the data in 80% training and 20% testing data. So we got 4457 training corpus and 1115 testing corpus.

Now let us discuss about the classifiers used in this task.

2.1 Multinomial Naive Bayes Classifier

The multinomial Naive Bayes classifier is suitable for classification with discrete features. It is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. Bayes Theorem is given as

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{1}$$

2.2 Logistic Regression Classifier

Logistic regression is one of the basic supervised learning algorithms. It is used to predict the probability of a categorical dependent variable. The output of a Logistic regression function is sigmoid function, but for classification problems we define a threshold and set the values to 0 or 1 depending upon how close the values are from 0 or 1. The sigmoid function [3] used in logistic regression is given as

$$C(X) = \frac{1}{1 + exp(-X\beta)} \tag{2}$$

2.3 Support Vector Machine Classifier

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection[4]. We have used the Linear Support vector classifier it is similar to using a kernel with 'linear' parameter but it is more lenient on penalties.

3 Evaluation Criteria

I have used 5 evaluation criterias. They are listed as follows:

3.1 Confusion Matrix

The aim to use Confusion Matrix or error matrix, is to count the number of times instances of class A are classified as class B.[6] tp is true positives, fp is false positives, tn is true negatives and fn is false negatives.

3.2 Precision

Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances[5]. From the confusion Matrix we can calculate the Precision(p) as

$$p = \frac{tp}{tp + fp} \tag{3}$$

3.3 Recall

Recall (also known as sensitivity) is the fraction of relevant instances that were retrieved[5]. From the confusion Matrix we can calculate the Recall(r) as

$$r = \frac{tp}{tp + fn} \tag{4}$$

3.4 Macro averaged f1-score

Macro F1-averaging is performed by first computing the F1-score per class/label and then averaging them.

3.5 Micro averaged f1-score

Micro averaged f1-score measures the F1-score of the aggregated contributions of all classes. f1 score is given by

$$f1 - score = \frac{2 * p * r}{p + r} \tag{5}$$

4 Analysis

Classifier	Multinomial Naive Bayes	Logistic Regression	SVM
Precision	0.7166	0.9838	0.9892
Recall	0.7166	0.9838	0.9892
Macro F1-score	0.6420	0.9651	0.9764
Micro F1-score	0.7166	0.9838	0.9892

From above table we can clearly say that for classification of SVM classifier is the best. But also it should be noted that the Logistic regression classifier also performed well on the same data. Also it is to be noted that the precision and recall of the Multinomial Naive Bayes is very poor when compared to

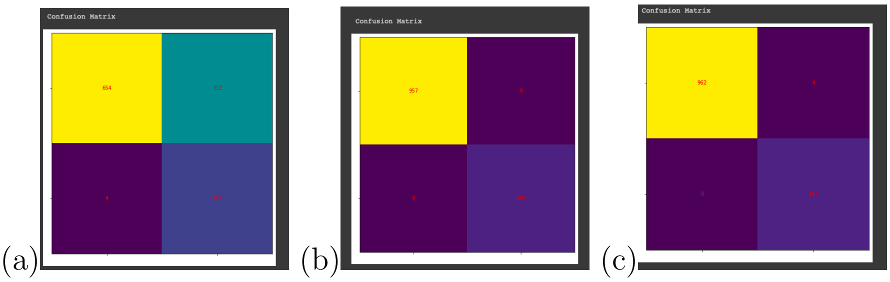

Figure 1: (a) Multinomial Naive Bayes (b) Logistic Regression (c) Support Vector Machine

SVM and Logistic Regression classifier. The main reason for this is because of the more false negatives predicted by this classifier as shown below. We show the confusion Matrix of the above classifiers.

5 Discussion and Conclusion

The scams related to SMS spam are increasing at a very rapid rate. Algorithms need to perform well. It may consider important messages like transaction messages as spam which is dangerous. The powerful algorithms available can be used to perform these tasks to get better performance. A large data set must be used for making the algorithm learn better. We can also conclude that stemming may or may not be helpful always. Stemming can cause some wrong spellings to be correct, which can lead to classifying spam as ham. It is also a time-consuming process and frequently fails to form words from the stem.s

References

[1] https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[3] https://online.stat.psu.edu/stat462/node/207/

[4] https://scikit-learn.org/stable/modules/svm.html

[5] https://en.wikipedia.org/wiki/Precision_and_recall

[6] geeksforgeeks.org/confusion-matrix-machine-learning/